

# **Making Sense of Virtual Risks**

## **A Quasi-Experimental Investigation into Game-Based Training**

### **Proefschrift**

ter verkrijging van de graad van doctor  
aan de Technische Universiteit Delft,  
op gezag van de Rector Magnificus prof. ir. K.Ch.A.M. Luyben,  
voorzitter van het College voor Promoties,  
in het openbaar te verdedigen op vrijdag 7 december 2012 om 12:30  
door

Casper HARTEVELD

bestuurskundig ingenieur  
geboren te Delft

**Dit proefschrift is goedgekeurd door de promotor:**

Prof. mr. dr. J.A. de Bruijn

Copromotor: Dr. I.S. Mayer

**Samenstelling promotiecommissie:**

Rector Magnificus, voorzitter

Prof. mr. dr. J.A. de Bruijn, Technische Universiteit Delft, promotor

Dr. I.S. Mayer, Technische Universiteit Delft, copromotor

Prof. dr. ir. A. Verbraeck, Technische Universiteit Delft

Prof. dr. ing. S. Schaap, Technische Universiteit Delft

Prof. dr. J.L.A. Geurts, Universiteit van Tilburg

Prof. dr. S. de Freitas FRSA, Coventry University

Dr. ir. A.R. Bidarra, Technische Universiteit Delft

Prof. mr. dr. E.F. ten Heuvelhof, Technische Universiteit Delft, reservelid

©2012 Casper Harteveld and IOS Press

Reuse of the knowledge and information in this publication is welcomed on the understanding that due credit is given to the source. However, neither the publisher nor the author can be held responsible for any consequences resulting from such use.

ISBN 978-1-61499-170-0 (print)

ISBN 978-1-61499-171-7 (online)

*Cover design:* Matthijs Schaap, Deltares

*Publisher*

IOS Press BV

Nieuwe Hemweg 6B

1013 BG Amsterdam

Netherlands

fax: +31 20 687 0019

e-mail: [order@iospress.nl](mailto:order@iospress.nl)

IOS Press, Inc.

4502 Rachael Manor Drive

Fairfax, VA 22032

USA

fax: +1 703 323 3668

e-mail: [iosbooks@iospress.com](mailto:iosbooks@iospress.com)

PRINTED IN THE NETHERLANDS

*To my parents*

# Contents

<b>Preface</b> .....	vii
<b>Acknowledgements</b> .....	x
<b>1   LOADING...</b> .....	1
Rise of a Potential Powerful Tool .....	3
The Case of <i>Levee Patroller</i> .....	11
Toward a Thicker Description .....	17
What to Expect .....	28
<b>2   Sought: A Professional Hans Brinker</b> .....	30
World of Reality: Levee Inspection .....	31
World of Meaning: Sensemaking .....	38
World of Play: 3D Simulation Game .....	46
Lessons Learned .....	52
<b>3   Toward an Innovative Training/Evaluation</b> .....	55
Evaluating a Futuristic Scenario .....	56
Formulating Working Hypotheses .....	65
Setting Up a Futuristic Scenario .....	72
Lessons Learned .....	84
<b>4   Playing a Futuristic Scenario</b> .....	86
The Participating Organizations .....	87
Setting the Facts Straight .....	101
Errors and Improvements .....	104
Lessons Learned .....	112
<b>5   Solving a Crime Is Easier</b> .....	114
Playing the Game at Home .....	115
Rating the Exercise Every Time .....	120
Retrieving the Prevalent Gameplay Responses .....	132

Lessons Learned . . . . .	146
<b>6 Opening the Black-Box . . . . .</b>	<b>147</b>
Performing in the Game . . . . .	149
Dealing with the Virtual Failures . . . . .	155
Retrieving the Prevalent Gameplay Patterns . . . . .	175
Lessons Learned . . . . .	190
<b>7 Knowing the Pen-and-Paper Generation . . . . .</b>	<b>192</b>
Retrieving the Questionnaire Responses . . . . .	193
Reducing the Data . . . . .	208
Investigating the Relationships . . . . .	215
Lessons Learned . . . . .	222
<b>8 Picture This! . . . . .</b>	<b>224</b>
Organizing the Sensemaking Test . . . . .	225
Picturing the Learning Objectives . . . . .	230
Picturing the Overall Results . . . . .	246
Lessons Learned . . . . .	255
<b>9 It Has an Exit Button, Right? . . . . .</b>	<b>256</b>
Setting Up a Discussion . . . . .	257
The Impact of <i>Levee Patroller</i> . . . . .	259
Suggestions for Improvement . . . . .	278
Lessons Learned . . . . .	295
<b>10 Picture That! . . . . .</b>	<b>297</b>
Playing with Digital Literates . . . . .	298
Sensemaking by Super Experts . . . . .	307
Interviewing the Levee Patrollers . . . . .	312
Exercising on a Real Levee . . . . .	329
Lessons Learned . . . . .	341
<b>11 Integrating the Puzzle Pieces . . . . .</b>	<b>343</b>
Accepting the Hypotheses (or Not) . . . . .	344
Providing Additional Perspectives . . . . .	363
Reflecting on the Puzzle . . . . .	369
Lessons Learned . . . . .	371
<b>12 The Future of Game-Based Training . . . . .</b>	<b>373</b>
Toward Sensegaming . . . . .	374
Advice for Future Training/Evaluations . . . . .	375
Recommendations for Future Research and Practice . . . . .	377
The Future of <i>Levee Patroller</i> . . . . .	381
<b>References . . . . .</b>	<b>383</b>

<b>Summary</b> .....	395
<b>Samenvatting</b> .....	402
<b>Curriculum Vitae</b> .....	409

# Preface

*We know the technology works, we have proven it over and over again, and we just want to get on with using it—Don Johnson, the Pentagon, in Prensky (2001, p. 295)*

When we think of the Netherlands with its levees (or dikes) and water, we immediately think of a tale by Mary Elizabeth Mapes Dodge from her novel *Hans Brinker or the Silver Skates* from 1865. What we think of in particular is when “Hans Brinker” becomes the Hero of Haarlem by putting his finger in the levee to prevent a flood. This legend goes like this:<sup>1</sup>

Many years ago, there lived in Haarlem, one of the principal cities of the Netherlands, a sunny-haired boy of gentle disposition. His father was a sluicer, that is, a man whose business it was to open and close the sluices, or large oaken gates, that are placed at regular distances across the entrances of the canals, to regulate the amount of water that shall flow into them.

The sluicer raises the gates more or less according to the quantity of water required, and closes them carefully at night, to avoid all possible danger of an oversupply running into the canal, or the water would soon overflow it and inundate the surrounding country. As a great portion of the Netherlands is lower than the level of the sea, the waters are kept from flooding the land only by means of strong levees, or barriers, and by means of these sluices, which are often strained to the utmost by the pressure of the rising tides. Even the little children in the Netherlands know that constant watchfulness is required to keep the rivers and ocean from flooding the country, and that a moment’s neglect of the sluicer’s duty may bring ruin and death to all.

One lovely autumn afternoon, when the boy was about eight years old, he noticed how the autumn rains had swollen the waters. He thought of his father’s brave old gates and felt glad of their strength, for, thought he, “If they gave way, these pretty fields would all be covered with the angry waters—Father always calls them the angry waters. I suppose he thinks they are mad at him for keeping them out so long.”

While thinking about this, the boy was suddenly startled by the sound of trickling water. Whence did it come? He looked up and saw a small hole in the levee through which a tiny stream was flowing. Any child in the Netherlands will shudder at the thought of a leak in the levee! The boy understood the danger at a glance. That little hole, if the water were allowed to trickle through, would soon be a large one, and a terrible inundation would be the result.

Quick as a flash, he saw his duty. The boy clambered up the heights until he reached the hole. His chubby little finger was thrust in, almost before he knew it. The flowing was

---

<sup>1</sup> The excerpt from Mary Elizabeth Mapes Dodge’s novel *Hans Brinker or the Silver Skates* from 1865 is based on the old English version as analyzed by the Dutch folktale researcher Theo Meder and revised by me to suit its purposes here.

stopped! Ah! he thought, with a chuckle of boyish delight, the angry waters must stay back now! Haarlem shall not be drowned while I am here!

Although it is a nice little story, it is wrong. Putting a finger in a hole is more likely to cause a flooding than prevent one. I do not want to get too (geo-)technical, but such an action increases the pressure onto the levee which will ultimately undermine it. It would be better to manage the flow of water instead of stopping the “angry waters” right away.

This book is about investigating how we can ensure that practitioners recognize risks, like Hans Brinker did when he heard “the sound of trickling water,” and know what to do when they encounter them. If Hans would have been properly trained, he would know not to put his finger in the levee. He would have *made sense* of the situation differently.

Many ways exist to achieve proper training and this book is geared toward exploring one potentially powerful one: the use of *digital games*. Like the tale of Hans Brinker, the value of game-based training is almost legendary. Its application has risen dramatically in the past decades and has been embraced gracefully with little to no foundations for why it works. When it comes to games, it seems as if people are putting their fingers in holes, because “that is how the story goes.” Some people, like Don Johnson, believe such stories so zealously they do not even want to look into the foundations. They “just want to get on with using it.”

The truth is that we are just getting an idea about the value of game-based training. Whereas this book concerns a small step in the larger scheme of things, it provides invaluable insights to anybody interested in using and evaluating games to train practitioners. These insights go beyond stories, fairy tales, and legends. They are based on a rigorous attempt to get to the bottom of it.

This attempt concerns a “small step” because the insights are derived from a single game: the game *Levee Patroller*, used to train practitioners in making sense of flood risks, such as the one encountered by Hans Brinker, by letting them make sense of virtual risks first. Since its initial release in 2006 this game has received widespread attention in the Netherlands and beyond, and users have responded positively to it. We could have decided to “get on with using it,” but we wanted to get to know its actual value and see what this means for the field in general.

Delft, the Netherlands, August 2012

Casper Hartevelt

## Notes

1. This book is based on and continues from my book called *Triadic Game Design*. This book describes the design of *Levee Patroller* in detail and gives an overview of the field of *serious gaming*. Serious gaming refers to the use of game technology for serious purposes, such as training and education.
2. Similar to the previous book, this one is divided into levels instead of chapters too. Playfulness is not the sole domain of games. I even added *progress bars* to keep you engaged.



3. Because I wanted to find a balance between rigor and readability in writing this book, you can find gray boxes like this throughout the book. These boxes give an *in-depth explanation* about what is described in the main text.
4. I tried to be consistent with the *Publication Manual of the American Psychological Association* (6th ed.), but for purposes of readability I deviated from it occasionally. For example, percentages are displayed in whole numbers (and because of this they may not total 100% due to rounding).
5. The statistical analyses in this book are based on the steps and advice by Field (2005) who has the gift of making statistics into something playful.
6. Most analyses were performed with *Microsoft Excel 2010* and *PASW Statistics 18*. The word analyses in Level 8 were done with *Wordsmith Tools 5.0*.
7. The names of the participants and organizations in this book are fictional, but it is not based on fiction.
8. Quotes by participants and from reports in Dutch have been freely translated by me.
9. I used several codes throughout the book for referencing my empirical material:
  - IPpre/post-# = Interview Patroller-pre- or post-interview-Participant number.
  - GQexercise-# = Game Questionnaire-exercise (e.g., ex1 or ex3)-Participant number.
  - GDexercise-# = Game Data-exercise (e.g., ex2 or ex4)-Participant number.
  - Dgroup-# = Discussion-group number (e.g., A1 or C2)-Random number to distinguish contributors.
10. The research presented here was performed as part of fulfilling the requirements of the Ph.D. degree at Delft University of Technology and was funded by Deltares. Deltares is a research institute for delta technology and is the product owner of *Levee Patroller*.

# Acknowledgements

*It does not take long. Time flies away—Participant #97 about playing Levee Patroller*

The journey to my dissertation has been one of collaboration—on a professional and personal level. I owe my gratitude to many people I worked with and I would like to take this opportunity to thank those who supported me throughout these years.

Let me start off with my two supervisors: Hans de Bruijn and Igor Mayer. Hans, your quick, pragmatic, and insightful advice ensured I got the necessary focus to write a book that is valuable in an academic as well as practical way. Your role nicely balanced Igor's role. Igor, you made me challenge myself to get the most out of my research and always pointed me toward a direction I did not think of before.

I also want to thank my colleagues. Although our fields are very different, it was inspiring to notice how much we can learn from each other. Specific mention goes out to Carla Haelermans, Emiel Kerpershoek, Harald Warmelink, Hester Goosensen, and Maartje van den Boogaard with whom I had many discussions about my research. The TU Delft Gaming Street members Daan Groen, Melvin Mukrab, Gert-Jan Stolk, and Linda van Veen deserve a mention too because their help was vital during my field work.

Theory without practice is not useful and the research institute Deltares made it possible for me to get both. They provided the necessary funding to perform my research and field work. I would like to acknowledge and thank Gerben Beetstra, Jos Maccabiani, Mandy Korff, and the Deltares Game team which consists of Rens van den Bergh, Arne Bezuijen, Micheline Hounjet, Almar Joling, and Matthijs Schaap. I further want to thank Jurjen van Deen for supporting me with publishing this work in the Deltares Select Series.

Field work without a field is not possible and so I am grateful that I was able to set up training sessions at the three water authority organizations whose names I will not reveal to preserve anonymity of the research data. I want to thank the coordinators in particular. I am of course also thankful to the 147 levee patrollers that were willing to participate with my research. Same goes for the students and experts whose data helped to validate the results.

Personally, I want to express my appreciation to my friends who have coped with my absence on weekends and trips. I am grateful that my roommate at work, Geertje Bekebrede, as well as my roommate at home, Niels de Vries, accepted to

be my paranymphs. Both have witnessed my journey from the start and gave me invaluable support and advice. I further want to especially thank my sister, Laura, who also witnessed my journey from the start and supported me with my research at several times.

The book is devoted to my parents. They not only made it possible for me to get to this point, they actually participated on various occasions. Like my supervisors they were a balanced team. My mom helped me with the practical issues—from arranging tea and coffee for the levee patrollers to printing this book—and my dad helped and advised me with the research aspects.

Time flies away. I can say this about the journey to my dissertation, but *especially* when I am together with Jordan. I want to thank her for her sustaining love.

This page intentionally left blank

# Level 1

## LOADING...

*I have been saying all my life that games have the power to change the world. We are proving that every day—Doug Whatley, CEO of BreakAway Games*

*Using new workers trained on the Service Rig Trainer is like getting a worker with six months experience—Shawn Primosch, Rig Manager, Concord Well Servicing, about Coole Immersive's Service Rig Trainer*



0%

“I died...AGAIN.”

Needless to say, no human being can die more than once. But these are not the words of a human being but rather of a player playing a game. Players die quite often. Maybe ten times a day. Or maybe even a thousand times if they are really out of luck. The second (or one thousandth) chance is one of the great things about games, which offer unlimited tries until you learn how to do it.

This possibility to die and try again is why players attempt the most dangerous activities imaginable: from crawling, jumping, and slinging over skyscrapers like Spiderman to starting an one-man army campaign against thousands and thousands of enemies—who may not even be human at all. When they happen to die, they take a sip of their drink, followed by a deep breath, and then start again.

For players this eternal life is not the only reason they play. It otherwise hardly explains why players play virtually a game of golf. Or of fishing. However, it does explain why a growing number of organizations have become interested in using games. Think of the military, the police, or the fire department. By letting their personnel get virtual experience in a game, they do not have to gain this in the real world, where they only have one life, and so do the people who depend on their actions. Even if it is not a matter of life or death such virtual experience may save something else—such as time and money.

The big question is: does it work? Is it valuable to use *game-based training* or do all those lost virtual lives not make any difference? At the moment we have true believers and non-believers. The true believers draw upon the rich history in which

games were used for serious purposes (Harteveld, 2011). They will refer to how the game of chess was used to develop war strategies in the Middle Ages (Smith, 2010; Vale, 2001) and will stress how the military has ever since embraced gaming wholeheartedly (Prensky, 2001, pp. 295–317). They will most likely cite Huizinga (1938/1955), author of the seminal work “Homo Ludens,” and argue that most of civilization came into existence by playing, or cite other thinkers who highlight for example the importance of leisure (Pieper, 1948/1998), flow (Csikszentmihalyi, 1991), and of learning by experience (Dewey, 1938; Kolb, 1984).

They will further point out how flight simulators have been used successfully for decades (Lee, 2005) and that in the sixties over 200 business games were available for use (Klasson, 1964). Or like Doug Whatley and Shawn Primosch they base it on their own experiences and simply quote participants who played their games and said “It was so much fun” and “Today I learned more than reading from a book.”

Some true believers go as far to suggest that “Long before today’s teenagers have grandchildren, Digital Game-Based Learning...will be totally taken for granted as the way people learn” (Prensky, 2001, p. 3) or that “The development of and adoption of simulations will change the nature of work, change the skill sets of our culture, and create an international industry that will eventually account for billions in revenue” (Aldrich, 2004, p. 229).

Some, such as McGonigal (2011), even suggest “that reality is broken, and we need to start making games to fix it” (p. 9) and that “if we commit to harnessing the power of games for real happiness and real change, then a better reality is more than possible—it is likely” (p. 354). Others are less pronounced (and provocative), but do make an argument for why games are actually good for you (Johnson, 2005), provide good learning environments (Gee, 2003; Koster, 2005; Shaffer, 2006; Squire, 2011), will change business as we know it (Beck & Wade, 2004; Edery & Mollick, 2009; Reeves & Read, 2009), and are much more than entertainment (Bogost, 2011; Jones, 2008; Sawyer, 2002).

The non-believers (or naysayers, see Prensky, 2001, pp. 372–377), on the other hand, may hold on to the strict difference between work and play. For them work cannot be play and play cannot be work (for this “commonsense tendency” and other misconceptions about play, see Rieber, 1996). They will cynically but rightly so point out that no “hard evidence” proves its utility and that existing evidence actually shows that other methods are more efficient.

Critics will also argue that games are far more expensive and take far more time to develop compared to other forms of technology-delivered training despite the uncertainty about return on investment (Sitzmann, 2011). They will think that this is “another hype” and that soon this will all be over, much similar as to how edutainment (Egenfeldt-Nielsen, 2007), virtual reality<sup>1</sup> (Castronova, 2005; Stone, 2005, 2009, pp. 286–294), and virtual worlds such as *Second Life* (Dibbell, 2011) emerged and declined. Today it is about gamification (Kapp, 2012; Zichermann & Cunningham, 2011), gamefulness (Deterding, Dixon, Khaled, & Nacke, 2011), and serious

<sup>1</sup> Castronova (2005) describes that virtual reality is “now re-emerging with considerable force” (p. 6), see also Blascovich and Bailenson (2011).

games (Bergeron, 2006; Michael & Chen, 2006). Tomorrow it will be about something else.

## Rise of a Potential Powerful Tool

Whether true believer, non-believer, or somewhere in between, everyone has witnessed the rise of *digital games* in the past decades. Since *Pong*, a minimalist digital version of tennis, was released in 1972, digital games have matured and pervaded our society. It is one of the fastest growing industries, with an impressive annual growth rate of around 7% (PwC Entertainment & Media, 2012). Worldwide revenue is currently about \$60 billion and total global spending is expected to expand to \$83 billion in 2016. To compare, filmed entertainment (includes box office revenues, DVDs and Blue Rays purchases and rentals, TV subscriptions, and pay-per-view revenues) is growing at a rate of around 3% and is projected at a worldwide revenue of \$99.7 billion in 2016.

With the emergence of digital games as a mainstream medium came a growing (academic and professional) interest in investigating what these games are, how they can be used and improved, and how they affect individuals, organizations, and society at large.<sup>2</sup> One such interest concerned a renewed interest in applying games for education (Egenfeldt-Nielsen, 2007). But people started to realize that games are a potentially powerful tool for many other serious purposes too (Harteveld, 2011, pp. 55–69). Games have been used to *a*) change *attitudes*, e.g. to persuade people to buy a product or change their diet (Bogost, 2007); *b*) *assess* organizational structures, processes, tools, instruments, or even people, e.g., to evaluate a new financial system before it is implemented (van Bueren, Mayer, Harteveld, & Scalzo, 2009); *c*) *collect data* useful for other purposes, e.g., to improve search engines (Von Ahn & Dabbish, 2004); *d*) *explore* the possibilities without having a clear idea up front, e.g., observing and understanding strategic behavior of different parties in a decision making process (Kuit, 2002); and *e*) *test theories* if users do have a clear idea up front, e.g., to determine the strategic behavior of buyers and suppliers in a supply chain (Meijer, 2009).

In addition, people have started to realize that games are a potentially powerful tool in many different domains (Harteveld, 2011, pp. 39–54). The use of games, including for teaching *knowledge* and *skills*, has found applications in business and management, health, the military, politics and society, public policy, safety and crisis response, and science and education. “There can be little doubt,” write Michael and Chen (2006, p. 232), “that serious games represent one of the most significant trends in video game development since the move into the third dimension.”

<sup>2</sup> It also resulted in the establishment of (digital) game research as a field (for an overview, see Mäyrä, 2008; Egenfeldt-Nielsen, Smith, & Tosca, 2008). According to Aarseth (2001), 2001 can be seen as the Year One of “Computer Game Studies” as an emerging, viable, international, and academic field.

Various names have appeared to coin this movement (or parts of it), but “serious games” or “serious gaming” is the most frequently referred to (for a short discussion on the “babel problem” in the field, see Hartevelde, 2011, pp. 6–7). The name first appeared as the title of a book by (Abt, 1970), highlighting that this movement already started around the same time *Pong* was released. Several well-known successes are to be noted in this emerging field, such as *America’s Army*, a “First-Person-Shooter” (FPS) game to make young civilians familiar with the US Army (Zyda et al., 2003); *Re-Mission*, a “Third-Person” action game for teenagers and young adolescents with cancer (Beale, Kato, Marin-Bowling, Guthrie, & Cole, 2007; Kato, Cole, Bradlyn, & Pollock, 2008; Tate, Haratatos, & Cole, 2009); *Foldit*, an online puzzle game about folding proteins (Cooper et al., 2010); the *ESP Game*, an online multiplayer game for collecting picture labels (Von Ahn, 2006); the *World Without Oil*, an Alternate Reality Game (ARG) for thinking how the world would be without oil (McGonigal, 2011, pp. 302–316). The increased interest and these initial successes seem to proof the true believers right: games are a potential powerful tool that will become ever more important.

However, in the past decade, much research—in particular about educational games—seems to suggest little evidence for games’ advantages.

- Leemkuil, de Jong, and Ootes (2000, p. ii) say that “Much of the work on the evaluation of games has been anecdotal, descriptive or judgmental, but there are some indications that they are effective.” They add, however, that “there is general consensus that learning with interactive environments such as games, simulations, and adventures is not effective when no instructional measures or support is added.” Support involves feedback, additional information, and assignments for example.
- Kirriemuir and McFarlane (2004, p. 28) say that “Though a rapidly growing and maturing body of research is helping to develop a clearer understanding of the educational potential of games, there are as yet a small number of games that have a clear contribution to make to the educational agenda.”
- “The evidence of potential is striking, but the empirical evidence of games as learning environments is scant” (p. 168) is what O’Neil, Wainess, and Baker (2005) conclude and according to them “games themselves are not sufficient for learning, but there are elements in games that can be activated within an instructional context that may enhance the learning process” (p. 465).
- According to Hays (2005), “empirical research...is fragmented” because it “includes research on different tasks, age groups, and types of games” and is “plagued with methodological flaws” (p. 53). He further emphasizes that “There is no evidence to indicate that games are the preferred instructional method in all situations” (p. 53) and like Leemkuil et al. (2000) he stresses that support is needed: “games should be used as adjuncts and aids, not as stand-alone instruction” (p. i).
- The Federation of American Scientists (2006, p. 6) observe that “Effective use of games and other new technologies is likely to be limited unless educational institutions are willing to consider significant changes in pedagogy and content,



and rethink the role of teachers” and “Outcome data from large-scale evaluations of educational games are needed to demonstrate that these technologies are equal to or offer comparative advantage vs. conventional instruction methods.”

- Egenfeldt-Nielsen (2006, pp. 188–190) wants to cure researchers’ amnesia about prior research and says that the “findings on learning outcome are positive and promising” but that “skepticism is warranted, however, because the lack of control groups, researcher bias, weak assessment tests, and short exposure time is not addressed sufficiently.” He concludes “that video games facilitate learning, but the evidence for saying any more than this is weak.”
- “One of the main obstructions to uptake games in learning contexts is a lack of empirical data to support the fact that they work, as well as a lack of understanding about how these games might be used most effectively in practice” and therefore a need exists for “more rigorous baseline studies that can quantify how much and in which ways games and simulations are currently being used most effectively to support learning” and for “guidelines, case studies, and exemplars from current practice to inform and improve the quality of delivery of games-based learning across the sector and to support better future planning and resource allocation” is what de Freitas (2006) suggests.
- Van Eck (2006, p. 30) asserts that game-based learning has been advocated for twenty-five years and “much of that time without any evidence of success,” but according to him “this has much less to do with attitude and learner preferences than it does with a technology that supports some of the most effective learning principles identified during the last hundred years.” He recommends to “focus on the strengths of the medium and provide the support and infrastructure needed to implement” games successfully.
- Vogel et al. (2006, p. 229) performed a meta-analytic analysis of 32 studies and found that “across people and situations, games and interactive simulations are more dominant for cognitive gain outcomes” but that when “teachers controlled the programs, no significant advantage was found” and “when the computer dictated the sequence of the program, results favored those in the traditional teaching method.”
- Ma, Williams, Prejean, and Richard (2007, p. 517) say that “the field has limited experience designing or implementing effective educational computer games...Empirical research should be conducted to develop a knowledge base that provides guidance for educators.” They further argue that “design-based research may inform the methodology for research on educational computer games.”
- Pivec and Pivec (2008, p. 1) assert that research “has been carried out over the past 20 years, but with very mixed results” and conclude that “Video games can supplement traditional learning but not replace it” and “the knowledge and skill level required to implement this technology successfully is lacking.”
- Ke (2009, p. 24) also finds research fragmented based on examining 89 studies and proposes “that instead of one-shot, incoherent experiments, future gaming research should take a systematic, comprehensive approach to examine dynam-

- ics governing the relations among multiple influential variables.” She further notes that “the empirical research on instructional gaming tends to focus on traditional learner groups while ignoring adult learners, especially the elderly.”
- Wouters, van der Spek, and van Oostendorp (2009, p. 246) say that much more research is required and one of their recommendations is to develop new ways of assessing game effectiveness. In particular,, they recommend a “visually-oriented assessment” because “video games are highly visual” and this may “reveal learning of knowledge that would probably not have been found with a text-based assessment method.”
  - Sitzmann (2011, p. 489) performed a meta-analytic examination of 65 studies who used a comparison group and concludes that “Trainees learned more, relative to a comparison group, when simulation games conveyed course material actively rather than passively, trainees could access the simulation game as many times as desired, and the simulation game was a supplement to other instructional methods rather than stand-alone instruction,” but learned less “when the instruction the comparison group received as a substitute for the simulation game actively engaged them in the learning experience.”
  - Young et al. (2012) identified 300+ articles and conclude that “The inconclusive nature of game-based learning seems to only hint at the value of games as educational tools...evidence for their impact on student achievement is slim...we can report finding evidence only for language learning and, to a lesser degree, physical education” (p. 80). According to them the slim evidence is a result of a “disconnect between the possible instructional affordances of games and how they are integrated into classrooms” (p. 80) and they recommend researchers to “utilize log files to establish complex connections between players and the virtual environment” and use “techniques...to understand how gaming unfolds across time and (virtual) space” (p. 83).
  - Girard, Ecalte, and Magnan (2012) focused on Randomized Controlled Trials (RCTs) with games, because that is “the gold standard for the evaluation of both medical treatment and educational interventions” (p. 2) and found that of the nine studies considered “only a few of the games resulted in improved learning, with the others having no positive effect on knowledge and skills acquisition when compared with more traditional methods of teaching...or to a control group which received no training” (p. 8). They further argue that because of the “lack of empirical studies,” more “experimental studies comparing the effect on learning” and “longitudinal studies to assess the long-term effect” need to be conducted (p. 10). We should also “avoid becoming overenthusiastic about the SGs [serious games] that are currently flooding the market until their effectiveness for learning has been scientifically demonstrated” (p. 10).
  - Connolly, Boyle, MacArthur, Hainey, and Boyle (2012) considered 129 papers with empirical evidence and note above all “the diversity of research on positive impacts and outcomes associated with playing digital games.” They developed a framework to categorize the research and highlight “the persistent difficulties associated with classifying learning outcomes.” Although they found “empirical evidence concerning the effectiveness of games-based learning,” they too

call “for more RCTs to provide more rigorous evidence,” but say that “more qualitative studies would also help to extend our understanding of the nature of engagement in games.”

Such evidence makes clear that we need to speak of “the rise of a *potential* powerful tool.” Gaming has potential, theoretically and based on some of the “hints” from literature, but we need to figure out how to utilize and proof that potential. The reports and articles provide us some directions.

### *Need for a Common Taxonomy*

One of the first struggles in pushing the field forward is the question of what is exactly being studied. Almost every report or article starts by defining what they mean by a game, simulation, serious game, and so forth, and each one decides differently:

...a closer inspection reveals that the “simulation games” selected for this meta-analysis [by Sitzmann (2011)] were not equivalent and do not fit with our definition of “simulation games” or “SGs” [serious games]. The author included in her analysis games which are too old to be simulation games, games which do not meet the criteria necessary in order (according to us) to be categorized as simulation games and games which have no ludic content whatsoever (Girard et al., 2012, p. 3).

This may explain why little overlap exists between the studies included in the reports and articles (see also Tobias & Fletcher, 2012). Some decide to include simulations (e.g., Sitzmann, 2011), whereas others explicitly separate games from simulations (e.g., Young et al., 2012). In short, no agreed-upon definitions exist and “this lack of organisation is regarded as an obstacle to progress in understanding the effects of games, developing more effective games and proposing guidance about how best to use games in learning” (Connolly et al., 2012, p. 662).

The problem is first of all due to the blur between the terms play, game, simulation, and simulator and the different associations scholars from different disciplines attach to each. I will not attempt to resolve this problem here, but I would like to point out that significant attempts have been made in establishing a common taxonomy what constitutes play (Huizinga, 1938/1955; Caillois, 1958/1961), what games are (Juul, 2005; Salen & Zimmerman, 2004), what the differences are between some of the terms (Deterding et al., 2011; Narayanasamy, Wong, Fung, & Rai, 2006), and how to categorize or classify serious games (Ratan & Ritterfeld, 2009; Sawyer & Smith, 2008). Also, I would like to provide my working definitions here to avoid confusion on this matter.

*Play* In a broad sense this refers to all voluntary activities that are deemed pleasurable, from gambling to playing games, and that could be coined as playful, such as denoting chapters in books as levels. In a strict sense it refers to a voluntary and unstructured activity with few to no rules and with no clear goal. Think of a make-believe activity by children, such as playing *Doctor & Nurse*.

*Game* A voluntary activity which is governed by rules and that includes a clear goal and feedback about the progression toward this goal. By “governed by rules” I mean that the course of the activity is determined by what has been agreed upon up front (or programmed). This cannot be changed.

*Digital game* A synonym for computer or videogame. I prefer the term “digital games” over “computer games” and “videogames,” as these terms refer in a strict sense to either PC-based games or console games (i.e., games played on Playstation, Xbox, or Wii), respectively. The term “digital” includes all games with a computerized backbone. In addition, it is the perfect antithesis of analog. Analog games are, among others, board and card games.

*Hardcore simulation* The term simulation could refer to advanced calculators used in operations research and management science which have a visual output and that require no involvement of the user, except for manipulating input variables. I consider these “hardcore simulations.”

*Simulation exercise* The term simulation could also refer to imaginative activities with a close correspondence to reality and that do require involvement from participants. Think of a fire drill exercise—real or virtual. I consider this a “simulation exercise.” It is more play-like, because it is very unstructured.

*Simulation game* The term simulation refers finally to a game genre. We speak of such a “simulation game” if the activity is close to reality, but also has game-like characteristics, such as rules and feedback. *Sim City* is a well-known example of a simulation game. Similar to the serious game definition, a simulation game does not need to be digital.

*Simulator* With the advances in computer graphics, simulators are barely distinguishable from simulations anymore (Narayanasamy et al., 2006). The differences remain especially in a much higher need to accurately model reality and the use of custom input and visualization devices, such as playing in an actual cockpit.

*Serious game* A game intentionally designed with a purpose other than entertainment in mind. Repurposed entertainment games used in education, such as *Civilization* or *Sim City*, which were not intentionally designed for these occasions excludes them from being a serious game. Unlike what some suggest (e.g., Bergeron, 2006), a serious game does not need to be digital.

*Educational game* A game intentionally designed with an educational purpose in mind. This concerns a subset of serious games.

## ***Need for Specialization***

The majority of the reports and articles put the reviewed games on one big pile, but it is known that we should not generalize research of “one game in one learning area for one group of learners to all games in all learning areas for all learners” (Hays, 2005, p. 53). Encountering the large diversity, Young et al. (2012) as well as Connolly et al. (2012) decided to classify the findings. Young et al. decided to split

their findings based on the content areas of mathematics, science, language learning, physical education, and history. Connolly et al. categorized games based on their primary purpose, genre, subject discipline, platform/delivery, and then categorized the effects too. The latter corresponds closely to what I suggested earlier (Harteveld, 2011, 31–88), to categorize a game by the *domain* it relates to (health or military?), the *value* it attempts to bring forth (knowledge or data collection?), and the type of *genre* it represents (puzzle or strategy game?).

How games are classified is one concern, but it is clear that further specialization is desirable to better understand the potential of games. Specialized reviews of games have appeared within certain domains such as health (Baranowski, Buday, Thompson, & Baranowski, 2008; Kato, 2010), public policy (Mayer, 2009), and the military (Smith, 2010). An even further specialization has taken place in examining games with a certain value within a certain domain. Exergaming, the use of games to stimulate people to exercise, is one example (Peng, Lin, & Crouse, 2011). This relates to the domain of health and to the value of attitude.

### *Need for Effective Design and Use*

We are still looking for answers on how to design and use games. Much recently, Tobias and Fletcher (2012, p. 234) even asked the community of researchers to answer “How do we produce games that reliably yield prespecified instructional objectives?”

However, extant research has taught us that a game is only effective if it *a*) is combined with other instructional methods or else learners “will learn to play the game rather than learn domain-specific knowledge embedded in the game” (Ke, 2009, p. 21); *b*) is integrated within a curriculum in a way that the strengths of the medium are harnessed; *c*) is employed by people who are knowledgeable about the technology (see also Mishra & Koehler, 2006); *d*) is actively played by players and more than once; and *e*) “is designed to meet specific instructional objectives and used as it was intended” (Hays, 2005, p. 23).

About the design in particular, a consensus exists that the game needs to align, integrate, or balance content with game characteristics (Harteveld, Guimarães, Mayer, & Bidarra, 2010; Harteveld, 2011). The field has learned from the mistakes from the edutainment movement, which concerned the first educational games, and knows that it should avoid “chocolate-covered broccoli” (Laurel, 2001) or “sugar-coated learning” (Egenfeldt-Nielsen, 2007) and that it should involve instructional and subject-matter experts throughout the design process in addition to programmers and artists. From reflections on design experiences this has become clear too (Frank, 2007; Hussain et al., 2010; Winn & Heeter, 2006; Marsh et al., 2011).

We also know that the field can learn a lot from commercial games (Becker, 2008). The better games—the “good” ones—apply good principles of learning and this has evolved according to a Darwinian process:

If a game, for whatever reason, has good principles of learning built into its design—that is, if it facilitates learning in good ways—then it gets played and can sell a lot of copies, if it is otherwise good as well. Other games can build on these principles and, perhaps, do them one step better. If a game has poor learning principles built into its design, then it will not get learned or played and will not sell well...In the end, then, video games represent a process...that leads to better and better designs for good learning and, indeed, good learning of hard and challenging things (Gee, 2003, p. 6).

According to Gee (2003) these principles have currently evolved in such a way as to align with what he believes are the best theories of learning in cognitive science and Van Eck (2006) seems to agree with him on this. But it makes a difference if a game is about collecting coins and shooting aliens or about learning quantum physics and Bayesian statistics. For educational games what needs to be taught is much more complex and has to be transferable to the real world. In addition, no huge competitive market exists that will decide in a Darwinian manner what will work. For each subject few educational games will be developed. Research is therefore needed to look into the effective design and use of serious gaming.

### ***Need for Rigorous and Innovative Research***

We learn especially that despite a decade of strong interest (if not longer) and an “explosion of publications and research studies dealing with the value and effects of games” (Tobias & Fletcher, 2011, p. 4), the field is still in dire need of comprehensive, rigorous studies into the *effectiveness* of games—that is, studies that go beyond anecdotal, descriptive, or judgmental evidence that Leemkuil et al. (2000) speak of and without any of the methodological flaws that Hays (2005) and others refer to. Some say these studies need to be RCTs, longitudinal, or at least compared to other instructional methods. Others are less directive but use words as systematic, comprehensive, rigor, and other denotations to indicate that we need to become more serious about serious game research.

Speaking of being systematic, strangely enough none of them clarify “effectiveness,” which could be interpreted in several ways. Effectiveness refers strictly to “doing the right thing.” In other words, does a game do what it is intended for? The answer could be “yes” or “no.” It does not say anything about the extent to which results are achieved. Strictly speaking, “how things are getting done” is encompassed by “efficacy.” It answers the question to what extent does a game do what it is intended for? The use of effectiveness in a strict sense is rather uninformative in an educational context and in this book I consider effectiveness and efficacy as synonyms.

A less trivial question is whether they include—again strictly speaking—what is considered *efficiency*. Efficiency is about “doing things in the most economical way.” This is about the amount of resources that are needed to achieve results. With games we can think of facilitator and player effort, the costs for developing and using games, and many other variables that allow us to compare the input and output of using a game. Although efficiency is important to consider, because in the end

one needs to also wonder if it was all worth it, I will consider this as something separate from what I just defined as effectiveness.

What becomes clear as well is that rigor is not enough. It requires *innovation* too in how games are used (e.g., Federation of American Scientists, 2006) and how results are measured (e.g., Wouters et al., 2009; Young et al., 2012). Otherwise we may not be able to make use of games effectively and are not able to capture what impact games have.

## **The Case of *Levee Patroller***

Amidst the rise of this potentially powerful tool in the past decade, I found myself involved with the development of a very unique game. The idea for this game sprung when in the summer of 2005—when Hurricane Katrina flooded a large swatch of the southeastern United States—a symposium was organized to exhibit innovative products for flood and water management in the Netherlands, a country with a long history of flooding. Two vivid “gamers,” a young manager who used to work in a game store and a student who worked part-time on creating three dimensional (3D) models, decided to create an interactive digital environment for this event. It did not need to serve any other purpose than entertain the visitors with the latest technology.

They made use of one of their favorite games to create this environment: *Unreal Tournament 2004*. This “First-Person Shooter” (FPS) was released the year before and is based on its successful predecessor from 1999. In both games, players control soldiers that have to fight each other with a massive arsenal. New to the 2004 edition is the ability for users to easily create their own content or add new content to the game. It is even possible to make a completely different game. This ability to modify an existing game is called *modding*, its derivative a “mod,” and if it results into a completely different game we speak of a “total conversion mod” (Postigo, 2007).

In the end, the two created in less than a month a total conversion mod of *Unreal Tournament 2004* in which players wander around an authentic Dutch landscape (without any guns of course). Players can hit some question marks scattered throughout the environment to receive information about flood management. They have to hurry if they want to find all of them, because at some point the water level rises and floods the virtual region. Although it was a simple demonstration, it sparked a discussion among the visitors about how digital game technology could serve a purpose for flood and water management. Such a tool, visitors reasoned, would allow calamity response organizations to safely learn about risks by exposure to virtual ones.

Inspired by the symposium, a team of people, including myself, started designing the game in February 2006. I was one of team members. My role was that of the “lead game designer.” This means I had to think of what the game had to be. I had to think of how the subject matter would be conveyed in an effective manner by playing this game. This was quite challenging because

I was packed with *a*) no knowledge or understanding of the subject matter at hand, *b*) a basic understanding of the workings of the human brain and how people learn, and *c*) little knowledge of games beyond playing them...I did not give up. Instead, I consulted experts with various backgrounds, read many books and articles, rooted in psychology to game design, and critically analyzed (and played) several games, from entertainment to serious ones. Looking back I can conclude that the experience was sometimes frustrating and sometimes a bit boring. On occasions, it took many hours, days, or even weeks before I figured out how to deal with a design dilemma. At other times, I was busy translating design documents, writing help files, and doing other activities that are not the most fun imaginable (for me at least). Nevertheless, the project kept me going: I was in a “flow” (Harteveld, 2011, p. 1).

We succeeded despite the challenge. Nine months later the initial version of *Levee Patroller* was released. The name refers to the game’s target group. Levee patrollers are considered the “eyes and ears” of the calamity response organizations. They inspect levees (also known as dikes/dykes), the artificial and natural barriers that protect a region from flooding, and report any risks they encounter.<sup>3</sup>

Much similar to the actual practice, in the game players have to find all virtual failures in a region and report these. If they do not find the failures in time or report them incorrectly, it could result in a levee breach that floods the whole virtual region. The game looks realistic and that is why people would consider it a simulation game at first sight. They are right, but in designing the game we used elements from a variety of game genres. It is first and foremost a simulation game, but it has characteristics of action, adventure, and puzzle games too.

The game received much media attention after its release—especially in the Netherlands but also abroad, in newspapers, magazines, and books. It has furthermore been exhibited at the Science Center NEMO in Amsterdam and at the Science Centre Delft. The first reactions by the actual levee patrollers and the calamity response organizations were positive too.

My role as the lead game designer was more or less finished. I could have gone on to focus primarily on other projects, but something was nagging. Would it be truly used by the patrollers? And if so, how is it used and does this happen effectively? Does the game work? If not, what needs to be changed? This is what I wondered after the initial release and in the past couple of years. They are the same types of questions the whole field continues to address.

## ***Unique But Not Alone***

*Levee Patroller* is a unique game because of the practice at which it is directed. Levee patrollers are not the most well-known types of practitioners. The calamity response organizations patrollers belong to are called the *water authorities* and these are the oldest form of democratic government in the Netherlands (Rijkswaterstaat,

---

<sup>3</sup> Formally the terms dikes or dykes, which are derived from the Dutch word *dijken*, are used to refer to the large barriers, those that protect the land from the rivers and the sea. The smaller barriers are seen as levees.





**Fig. 1.1** Monument “The Levee Patrollers” by artist Frans Ram (1991) on one of the levee segments of the Dutch island Ameland. It is a symbol of the safety of Ameland. The inscription reads “Though the storm rushes the waves sometimes fearfully high, the levee patroller is prepared with a vigilant eye”

2011, p. 18). The first ones were established in 13th century and they were established to start dealing with the “angry waters” in an organized manner—with practitioners called levee patrollers. Especially in the Middle Ages the Netherlands had suffered from many floods.

Although *Levee Patroller* is a unique game, many similar game-like (or game-based) digital technologies have also been developed in the past decade. To name but a few: *a) Hazmat: Hotzone*, an instructor-based simulation that used videogame technology to train first responder response to hazardous materials emergencies; *b) Triage Trainer*, a game to train the process of determining the priority of patients’ treatments based on the severity of their condition; *c) Hazard Recognition Game*, a game that enables to safely acquire and demonstrate the competencies necessary to supervise critical tasks in the oil industry; *d) Pulse!!*, a virtual reality learning platform where players get into a virtual intensive care unit and need to assess, diagnose, and treat the injuries of patients during catastrophic incidents, such as combat or bioterrorism.

What these technologies have in common is that they are situated in the same domain and attempt to bring forth the same sort of value. They even use a similar type of game genre to accomplish this.

### **Same domain: safety and crisis response**

The domain of a game is the subject matter or discipline it relates to. *Levee Patroller* and its affiliated technologies relate, of course, to the domain of safety and crisis response.

To be safe we try to prevent or mitigate the consequences of incidents and accidents and invest in security to deal with malicious acts, such as sabotage and terrorism. We especially want to prevent or mitigate a *crisis*. A crisis is an urgent threat that “marks a phase of disorder in the seemingly normal development of a system” (Boin, ’t Hart, Stern, & Sundelius, 2005, p. 2). A (natural) disaster, such as a flood, could lead to a crisis when it severely disrupts basic infrastructures and no accurate response is given or possible.

Many disciplines are concerned with this domain: safety science, crisis management, disaster management, and security management among others. Differences exist between these related disciplines, but they all involve dealing with *risks*. When I talk about risks, I refer to the Oxford Dictionaries definition that speaks of “the possibility that something unpleasant or unwelcome will happen.” In some cases, such as with a triage, something unpleasant or unwelcome has already occurred and then we have to deal with risks that make it possibly even more unpleasant or unwelcome. By taking proper action, which involves accurately and timely diagnosing victims, the consequences can be minimized.

Gaming represents new potential for addressing 21st century risks, such as pollution, diseases, and Internet security threats. In addition, the world of today is much more complex, interdependent, populated, and connected and this increases the consequences of risks. Ale (2009, p. 1) says that “The risks of modern technological society can be managed by using the means society has developed.” One of these means are games. That is why a number of organizations that are preoccupied with risks are increasingly interested in this medium. The Dutch water authorities were most certainly not the only ones who thought that games might be a potential powerful tool.

### **Same value: knowledge by means of sensemaking**

The value of games in the domain of safety and crisis response is manifold. They could be used for assessment, such as the testing of equipment and new procedures, or for exploration, such as finding the best evacuation route. The most common association in this area is with training and this is what *Levee Patroller* shares with the other technologies that have been developed in this area. The reasons why the technologies are valuable in this domain relate to the reasons why flight simulators are used. According to Rolfe and Staples (1988, p. 2) these are the five major advantages for using flight simulators:

1. Increased efficiency, as training will not be interfered by factors such as adverse weather conditions or aircraft availability. In addition, situations can easily be repeated or changed on-the-fly;
2. Increased safety and the ability to control the level of task demand;
3. Lower overall training costs;
4. The reduction in operational and environmental disturbance;
5. The facility to practice situations which for reasons of expense, safety, and practicability cannot be rehearsed in the real world.

It is striking that they do not mention “increased effectiveness.” It turns out that as with games, few rigorous evaluation studies have been done “due to the high cost and substantial logistical problems involved in conducting transfer-of-training studies” (Lee, 2005, p. 75). It seems that flight simulators have become commonplace, because they are less expensive, faster, safer, and more flexible than real-life training.

Games in the domain of safety and crisis response are especially developed because of the fifth advantage. We want practitioners to get experience in dealing with incidents, accidents, security issues, and threats without real-life damage to person or property.

Furthermore, it becomes further clear that the first generation of these game-like technologies are especially oriented at events that rarely occur. Dealing with levee failures (*Levee Patroller*), hazardous materials emergencies (*Hazmat: Hot-zone*), triaging patients (*Triage Trainer*), and treating patients during catastrophic incidents (*Pulse!!*) are rare events. Essentially what they are used for is *sensemaking*: a process by which people give meaning to experience (Dervin, Foreman-Wernet, & Lauterbach, 2003; Weick, 1995). The outcome of this process is *knowledge* about what, when, and where risks occur and what needs to be done once they are encountered.

This ability to make sense is at the heart of what these technologies are about. That this possibly happens with lower cost, faster, safer, and more flexibility than a real-life training is an additional bonus.

### Same genre: 3D simulation

There are game-like technologies in the domain of safety and crisis response that predate *Levee Patroller*, such as a simulation game called *Firestorm: The Forest Fire Simulation Program*. As the title suggests, the goal is to extinguish fires. This was published in 1995 and uses unlike the first generation games two dimensional (2D) graphics.

At the time this game was published true three dimensional (3D) graphics was about to become possible.<sup>4</sup> This happened with the appearance of *Quake* in 1996 and its accompanied *Quake Engine*, the game engine behind the game. At this point the use of game engines and modding started to take an enormous flight, as it became easy to develop a completely different game in a short time. The *Quake Engine* was not much later, in 1998, eclipsed by the *Unreal Engine*, the technology behind *Unreal Tournament* and *Levee Patroller*. This engine and its successors are still one of the most popular engines to date, also for developers with a serious purpose in mind.

Although in the 21st century true 3D games became the standard in the entertainment industry, and the technology became more accessible in terms of costs and use,

<sup>4</sup> The first 3D games date back from the seventies and eighties. These are not true 3D, because objects in these games are actually 2D and it is often not possible to look up and down. That is why the graphics in these games are referred to as “pseudo-3D” or “2.5D.”

the use of 2D remains prevalent. Because the use of 3D is prevalent among the first generation of game-like technologies in the domain of safety and crisis response this makes it a defining feature that sets them apart from for example games used for marketing and advertising (e.g., advergaming), the news (e.g., newsgaming), and societal critique or involvement (i.e., social impact games or games for change).

Other early attempts concern the use of computer or game technology as (interactive) visualizations within a training. The visualizations are basically a theater or stage that provides the setting for a simulation exercise. Participants make decisions, as individuals or as team, and one or more instructors decide how the exercise unfolds. This is for example a description of *XVR*, a virtual reality training for safety and security professionals (E-Semble, 2010):

Using a joystick *XVR* allows one or more incident response professionals to walk, drive or fly around in the simulated reality of an incident. This gives them the opportunity to train in observing and assessing the environment. Furthermore they have to assess risks and dangers, decide which measures to take and what procedures to apply, and report to the other rescue crew members.

An essential feature of *XVR* is that the instructor can easily build an incident scenario and has full control over the course of events in the scenario during the exercise. After starting the exercise, the instructor presents the student with questions and asks the student to motivate his decisions. He can also give feedback, for instance by role-playing the control room or other rescue staff. The instructor can respond to the student's decisions by activating events in the virtual scenario. The instructor may also decide to condense time and jump to a next phase in the incident.

This idea of using a real-time, interactive virtual environment goes back to at least 1992 when the Environment Tectonics Corporation (ETC) started to build their *Advanced Disaster Management Simulator (ADMS)* which they released in 1994 to train incident commanders, first responders, and incident command teams. This means that around the same time one of the first simulation games was released and 3D technology became available, training with virtual environments already started too.

What the first generation of game-like technologies share with these interactive virtual environments is a focus on realism. The close similarity to the real world is another defining feature that sets these game-like technologies apart from others. The difference with these virtual environments and also the difference among the technologies is how game-like they are. Some such as *Levee Patroller* and *Triage Trainer* are completely governed by rules. Because of their realism I consider them simulation games. Others such as *Hazmat: Hotzone* are led by instructors. This makes them similar to the interactive virtual environments, such as *ADMS* and *XVR* who are unstructured and therefore more play-like. They are what I consider simulation exercises.

With this we have come to an end in defining to what types of technologies *Levee Patroller* relates to. These are games and simulation exercises that belong to the domain of safety and crisis response and who aim at letting players make sense of risks in a realistic 3D virtual environment.

## ***Unique But Unique Opportunity***

Although *Levee Patroller* might be unique, it provides for a unique opportunity to contribute to maturing the field. First of all, the game belongs to a certain specialization in the field and not much is known about this specialization. Whereas at the moment a wealth of research is pursued in health (Arnab, Dunwell, & Debattista, 2012) and classroom education (Tobias & Fletcher, 2011), little is known about the use of games in the domain of safety and crisis response.

Previous research includes a published effectiveness study of *Triage Trainer* (Knight et al., 2010); a qualitative study about the design of *Hazmat: Hotzone* (Harz & Stern, 2008); a discussion on the experiences of developing the *Hazard Recognition Game* (Warmelink, Meijer, Mayer, & Verbraeck, 2009); a description of the research model and development of *Pulse!!* (Dunne & McDonald, 2010) and of how medical curricula could be provided in virtual space by using *Pulse!!* as an example (McDonald, 2010); an empirical investigation of design guidelines of another triage game called *Code Red: Triage* (van der Spek, 2011; van der Spek, Wouters, & van Oostendorp, 2011); a very short description of yet another triage game called *Burn Center* (Kurenov, Cance, Noel, & Mozingo, 2009); and a technical account of again a triage game, this time called *UnrealTriage* (McGrath & Hill, 2004). I am sure much more exist, but it would not be that much more. And if so they would most likely involve (technical) descriptions of the games (and other triage games).

Moreover, of the known closely related technologies many remained prototypes and never found a real application (so far). Of the examples mentioned, some have been discontinued (*Hazmat: Hotzone* and *UnrealTriage*) and others are still in development (*Pulse!!* and *Hazard Recognition Game*). Despite the promising results that were accomplished with *Triage Trainer*, this game is even still a prototype. It only allows for an hour of gameplay and has not been fully embedded within a program. *Burn Center* did. It is now part of a certified eight hours training program. However, this program involves 12 video lectures and an assessment besides the game and, therefore, I do not expect that the game is played for too long.

Unlike these similar technologies, *Levee Patroller* has been fully developed to facilitate many hours of training. And unlike most it also found an application, as five water authorities participated in its development and wanted to build a curriculum around it. Not less importantly, unlike some of its affiliated technologies, it is a game by every definition. In sum, *Levee Patroller* is unique but it provides for a unique possibility to contribute to the maturity of the field—for safety and crisis response in particular but also for serious games in general.

## **Toward a Thicker Description**

The objectives behind the investigation of *Levee Patroller* was two-fold. The first objective relates to the dire need for evidence about the effectiveness of games. This objective was to design and implement an innovative game-based training interven-

tion and evaluate its effectiveness in a comprehensive and rigorous manner. The following questions are associated with this objective:

1. What is the effectiveness of the training with *Levee Patroller*?
2. What factors contribute to its effectiveness?

Because so little is known about game-based training and in particular regarding the domain of safety and crisis response, the second objective was to develop a substantiated understanding of what makes a game successful in training practitioners to make sense of risks. Such understanding would be developed by considering the following questions:

1. How do participants experience the game-based training?
2. How do participants play the game?

To answer the questions I applied several methodologies and methods—quantitative and qualitative. However, this quote by Geertz (1973) explains what the research presented in this book is ultimately about:

This, it must be immediately said, is not a matter of methods. From one point of view, that of the textbook, doing ethnography is establishing rapport, selecting informants, transcribing texts, taking genealogies, mapping fields, keeping a diary, and so on. But it is not these things, techniques and received procedures, that define the enterprise. What defines it is the kind of intellectual effort it is: an elaborate venture in..."thick description" (p. 6).

What was really aimed for in this investigation was to establish a *thick description* of a game-based training. The study did not only aim for measuring the results, but also for providing a context and a better understanding on how these results were established. In fact, what I was really looking for with the investigation into the game-based training with *Levee Patroller* was this "elaborate venture" to describe and understand what happened and to make this meaningful to others. Because a mix of methods and methodologies were use to get this description (and not just ethnographic ones), we could possibly speak of establishing a "thicker description."

Ten evaluation principles were kept in mind for implementing and evaluating the game-based training with *Levee Patroller*. These principles are based on my understanding of the field and how it can proceed to "the next level." They also declare the focus, scope, and assumptions behind the study.

## ***1. Rome Was Not Built in a Day***

Although gaming has a long and rich history, it has not been studied comprehensively until recently. We cannot expect that the first games work perfect right away. Also, the serious games market does not have the mass market evolutionary mechanism that its entertainment counterpart has and where Gee (2003) speaks of. For understanding and improving games we need design, use, and evaluation theories, frameworks, and methodologies and precisely these are missing—at least ones that

are tested, accepted, and widely used, because various attempts have been made (e.g., de Freitas & Oliver, 2006; Kriz & Hense, 2006; Winn, 2009). We further seem to need a universally agreed upon taxonomy. Here attempts have been made too (e.g., Sawyer & Smith, 2008).

As we are building the knowledge base on games I considered that an understanding of what works and what does not is more fruitful than the evidence itself. It is nice to know that the *Triage Trainer* works better in some ways than a card-sorting exercise (Knight et al., 2010), but we do not know why. By knowing and understanding the successes and failures we are able to build the Rome of games.

It is this principle that made me decide to focus on effectiveness and leave out efficiency. We first of all need to make sure games work and understand how they might work better. After that we can worry about efficiency.

This principle also led me decide to pursue a *mixed methods* study. The unique challenges of the fundamental research questions require a combination of quantitative and qualitative methods.

This principle highlights above all my starting assumption: games have enormous potential and we need to give them a chance. To allow games to come to fruition a constructive research attitude is needed and according to Squire (2007, pp. 53–54) this is one that avoids using “cookie cutter applications of textbook research methodologies” but one that seeks theories

to explain how particular game-based approaches...work within particular contexts. Of course, you would want to collaborate with practitioners to implement such programs (allowing them to adapt materials as necessary), but the idea is that researchers iteratively design and research these pedagogical models as “proof” of what games can do, and then systematically design the necessary and sufficient conditions for them to work.

Only then we are able to build the Rome of games. If this happens to be an unsustainable and unproductive city, the Germanic tribes of traditional instructional methods will ambush and destroy it at some point. If we let them ambush it now already, we may risk of having to use another phrase, that of “throwing the baby out with the bath water.”

## ***2. No Comparison of Apples and Oranges***

Girard et al. (2012, p. 2) called Randomized Controlled Trials (RCTs) the “gold standard” for evaluation and urged game researchers to pursue this. Egenfeldt-Nielsen (2006, p. 190) called the comparison of games with other teaching styles “the ultimate test.” The idea according to them and others (e.g., Clark, 2007; Connolly et al., 2012) is that we need to proof that games work by comparing them to alternative treatments. Although RCTs and other comparative designs are in some situations suitable and provide valuable information (e.g., Jennett et al., 2008; Beale et al., 2007; Brown et al., 1997; Kato et al., 2008; Knight et al., 2010; Ke & Grabowski, 2007), in many cases this results in comparing apples and oranges and I will explain why:

- A game is often developed because of a niche it could fulfill in a curriculum or particular context. This means that nothing comparable exists and researchers have to find a way to make a comparison. They could do this by developing alternatives in addition to the game, but this has the risk of being “weak, ‘straw man’ alternatives” (Clark, 2007, p. 56).
- “To simply ‘use games to meet the same old demands’ of education may miss the point” (Squire, 2007, p. 53–54). Gaming is an instructional method in its own right, with its own strengths and weaknesses that we need to find out, and should be used to teach in an entirely different way. It will not replace traditional education, but rather augment, reshape, and change it.
- Previous research convincingly highlights over and over again that a game works best if it has been integrated into a curriculum with other instructional methods. This confirms that a game fulfills a niche and does not replace traditional education but rather augments, reshapes, and changes it. Isolating the game from other instructional methods will lead to failure and integrating it with others will confound the research.
- To teach in an entirely new way requires innovation, by for example not sticking to a workshop format of an evening or a classroom schedule of 30 or 45 minutes. Sticking to these traditional formats puts games at a disadvantage and by innovating the alternative treatments are put at a disadvantage.

I get that policy makers and investors want to see a value over existing practices. With levee inspection this is no different. But this gold standard or ultimate test often results in forced comparisons that are not sensible and do not lead to the innovations that are necessary to bring the field forward. In line with the principle of “Rome was not built in a day” I am also of the opinion that it is more valuable to focus on the games themselves—to give them a chance and see what they (could) do by devoting time and energy to investigating how they (can) work. Policy makers and investors can then decide if what is achieved is worth pursuing further.

With *Levee Patroller* in particular no doubt other solutions are imaginable, but this game was developed precisely because of the impossibility to practice with levee failures, fulfilling a niche in the education of patrollers. Instigated by the demands of applying the gold standard and ultimate test of research I thought of many possibilities to compare the game to alternatives. Every time I came to the same conclusion: it feels like comparing apples and oranges and it will hamper the innovation that is needed. Therefore, I decided to focus completely on *Levee Patroller* instead.

### 3. See the Big Picture

An area in which RCTs and other comparative designs become more sensible is in varying in different design options or game attributes by for example modifying task difficulty (Orvis, Horn, & Belanich, 2008), instructional support (Cameron & Dwyer, 2005; Leemkuil, 2006; Yaman, Nerdel, & Bayrhuber, 2008), and design



principles (van der Spek, 2011). Wilson et al. (2009) recommended this type of research, because

...in all of the studies reviewed, multiple game attributes were embedded in the games. It is not clear whether one attribute had a greater impact on learning than another, or whether it was the combination of attributes that led to success. Therefore, future research must seek to understand which specific game attribute(s) have the greatest impact on learning (p. 259).

The manipulation of design variables (of the game and its use) seems to be a trend and acknowledges that other researchers implicitly also decided to give games a chance and see how they could be improved by empirically and systematically investigating their implementation. Squire (2007, p. 53), however, considers this as belonging to a traditional educational technology paradigm which

...involves looking for blanket statements about whether games “work,” or even isolating variables (like removing teachers from the equation and seeing what happens) in an effort to come up with variables that can be universally applied. Imagine the problems with making blanket statements about “books” as an effective instructional medium, or the instructional effectiveness of “color” in educational film...this body of work [referring to an emerging paradigm of game-based learning predicated on theories of situated cognition] seeks to avoid the “no-significant-differences phenomena”...and seeks to use iterative research, theory building, and design to generate useful theory (p. 53).

Although Squire (2007) makes some good points, the manipulation of design variables, as suggested by Wilson et al. (2009), has proven its usefulness already. Otherwise we would not have known as much about the importance of instructional support as we do right now.

With *Levee Patroller* I decided to not focus on the details by isolating or manipulating variables, opting instead to focus on the big picture. This is because of the exploratory and innovative nature of the research I had in mind and because I simply had no idea if it would work at all. An approach as described by Squire (2007) fits such an uncertain environment much better. But I focused on this above all because my objective was to see if the game as a whole works and this precludes looking into any of the details.

I mention this principle for another reason too. Up front I was aware that many variables play a role in how a game is experienced (Kriz & Hense, 2006): the quality of the game, the facilitators, the participants' computer skills, and their game experience are one out of many variables that may determine the outcomes. It is impossible to focus on each of these variables in a detailed manner and here again I made a choice of breadth over depth.

#### ***4. More Than the Tip of the Iceberg***

As with most training evaluations (Haskell, 1998), most game evaluations have been based on a smile sheet at worst and a questionnaire at best (e.g., O'Neil et al., 2005). This is an issue, because a meta-analysis on knowledge self-assessments in education and workplace training by Sitzmann, Ely, Brown, and Bauer (2010) indicates

that self-assessments are strongly correlated with motivation and satisfaction and only moderately with cognitive learning. Therefore, researchers need to be “more prudent in their use of self-assessments of knowledge” (p. 30) as a proxy measure for cognitive learning. Sitzmann et al. recommend to limit its role in evaluation research and practice and rely instead on objective tests to assess learners’ knowledge level and gain.

In another article the researchers recommend that “In cases where we seek a measure of cognitive learning, this will often mean taking the less easy path, measuring multiple outcomes, and striving to create new measures and use new analytical techniques that are free from the biases associated with self assessment” Brown, Sitzmann, and Bauer (2010, p. 352). Moskal (2010, p. 314) captured this advice already succinctly by saying that “Self-assessment is one piece of the puzzle, but used in isolation, the puzzle remains incomplete.”

Beyond this primary reason to measure “more than the tip of the iceberg” provided by self-assessments, the need to “create new measures” and “new analytical techniques” for measuring cognitive learning corresponds to other needs called for in game research. Traditional measures may not be able to capture the impact of games who are especially visually oriented (Wouters et al., 2009) or how they unfold over time and space (Young et al., 2012).

To get more than the tip of the iceberg, this requires to innovate in this space as well. Not only do we need to rethink how we use and implement games, we also need to rethink how we measure the results.

## ***5. The Proof of the Pudding is in the Eating***

Squire (2007, pp. 54) stresses that we should avoid using “cookie cutter applications of textbook research methodologies.” He does not elaborate what he means by this, but this is my interpretation: If games are evaluated, even if that happens by what we consider “rigorous” according to the scientific standards, by for example applying a RCT or another strong research design, playing the game is considered as an “independent variable” or “intervention” and is not further considered. It is treated as a black-box (Chen, 1990).

What I mean by this black-box is that researchers are only concerned about the measures that they retrieve from their questionnaires and tests. They are not concerned about what happens throughout the game. For example, researchers may randomly assign participants to two groups, one is assigned to play a game and the other gets an alternative treatment. Then both have to take a test and from this it turns out that the game group performs better. What do we learn from this other than that in this particular context with this particular game the game performs better on this particular test?

The hypothetical example I just described is, in fact, the research design of the evaluation of *Triage Trainer* (Knight et al., 2010). The researchers say their study “contributes towards an understanding of the issues surrounding the use of serious

games in healthcare education, and the factors influencing their efficacy” (p. 1178), but I really cannot see how. Their empirical study does not give us any understanding of what mechanisms account for the game’s success. The lack of contextual understanding, of the game’s design and the possible factors contributing to its success, hampers any generalization beyond this game and the context of the workshop with which they evaluated it. What the study does provide is empirical evidence about the potential of games.

The neglect of what happens throughout the game contrasts with the game research frameworks that have been developed (Garris, Ahlers, & Driskell, 2002; Kriz & Hense, 2006; Winn, 2009). Each is a logic model (Bickman, 1987), representing how particular results are produced. They are called logic models, because they are supposed to provide a “logical” way of showing how this happens. Each stresses that the design and other input variables influence how a game is played and that this, in turn, determines the outcomes. Consequently, it is important to consider the gameplay systematically and not consider this a black-box. As they say, “the proof of the pudding is in the eating.”

## 6. *The Icing and the Cake*

In many studies the game and the gameplay are not completely neglected. For example, Beale et al. (2007) recorded how many hours participants played *Re-Mission* and the number of unique missions they completed to see if this influenced the outcomes. Furthermore, measuring both—game and gameplay—is often addressed by researchers with questionnaires. Participants have to rate in those instances the quality of the game and how they experienced playing it. This makes it possible to see if participants who enjoyed the game also learned more compared to those who did not.

Game scores or performance are occasionally considered too, especially in the evaluation of typical business games and have been investigated early on for purposes such as employee selection and appraisal (Vance & Gray, 1967). Many are however critical about relating these indicators to learning, because performance does not necessarily entail learning (Washbush & Gosen, 2001).

My point here is that if game aspects are considered in educational game evaluations, they concern some questionnaire items and possibly some data from the game, be it game scores or amount of gameplay time. However, they are rarely of primary importance. These game-related measures are used to explain for and support the findings of the primary measuring instruments. In other words, evaluators consider the icing and not the cake of what happens throughout the game.

With this minimal consideration of game aspects, scholars investigating educational games lag far behind compared to the investigation of games for other serious purposes. There they consider the icing and the cake by making fully use of the game data. The studies in which games are used for data collection come to mind

(Von Ahn, 2006; Cooper et al., 2010), but also studies in which games are used for theory testing (Meijer, 2009) and exploration (Kuit, 2002).

Of course, in these studies researchers depend completely on the game as their primary research method. But the fact that these scholars find gaming environments valuable to retrieve data should signal that with educational games, the cake could be considered as well as the icing. One particular recommendation is to “utilize log files” (Young et al., 2012, p. 83).

## ***7. Ain't Nothing Like the Real Thing***

Researching games requires participants and here scholars are confronted with the “college-students-as-research-subjects issue.” It is a generally accepted research practice to use college students in research studies, though there is a question as to whether (undergraduate) college students are representative surrogates for a broader population. According to Peterson (2001) the issue has been formally recognized, empirically examined, and heatedly debated in various disciplines for over decades and he performed a second-order meta-analysis to assess the implications of using college students subjects in social science research. He concludes that

The primary implication of the present research is that social science researchers should be cautious when using college student subjects and be cognizant of the implications of doing so if the purpose of the investigation is to produce universal principles. More specifically, the present research suggests that, by relying on college student subjects, researchers may be constrained regarding what might be learned about consumer behavior and in certain instances may even be misinformed (p. 458).

Sitzmann (2011, p. 507) reports in her meta-analysis that among 65 samples and 6,476 trainees “Learners were undergraduate students in 77% of samples, graduate students in 12% of samples, employees in 5% of samples, and military personnel in 6% of samples.” Sitzmann adds that “The effect of simulation games, relative to a comparison group, on learning did not significantly differ across undergraduate, graduate, employee, or military populations.” This seems to suggest that it is not much of an issue to use students for investigating learning with games.

It is, however, a *sine qua non* to play with the target group because it is generally assumed that

Play is greatly influenced by not only the design, but also the player, including his or her cognitive, social, cultural, and experiential background that he or she brings to the given play experience. Therefore, the experience of one player may be profoundly different than the experience of another player. The target audience for the game must be strongly taken into account throughout the design process (Winn, 2009, p. 1014).

I would argue that the target audience needs to be strongly taken into account during the evaluation for these reasons too. It seems likely that a 47 year old farmer will play and experience a game such as *Levee Patroller* quite differently than a student or teenager. Especially with games using students may seriously constrain what

might be learned or misinform researchers. Like Peterson (2001) I am however not against the use of ‘college-students-as-research-subjects,’ but ideally, researchers should attempt to play their games with the actual target group. There “ain’t nothing like the real thing.”

## 8. *Practice Makes Perfect*

Earlier I mentioned that the Federation of American Scientists (2006, p. 6) has observed that “Effective use of games and other new technologies is likely to be limited unless educational institutions are willing to consider significant changes in pedagogy and content, and rethink the role of teachers.” My interpretation is that we need to innovate, by for example not sticking to a workshop format of an evening or a classroom schedule of 30 or 45 minutes. Young et al. (2012) seem to agree:

Many educational games have assimilated game features into the constraints of the school day, becoming 20-minute activities with associated work sheets that lack a multiplayer continuity and the extended engagement characteristic of games played for purely entertainment value. Such adaptation may mask the learning benefits of video games...there appears to be a disconnect between the possible instructional affordances of games and how they are integrated into classrooms...most schools trade off extended immersion for curriculum coverage, individual play, and short exposures, goals that are not well aligned with engaging video game play (p. 80).

Among training practitioners, the need for extended play is even more evident. It is generally known that for becoming an expert in something practice is required and this is a slow process that takes up a great deal of hard work (Anderson, 1995). We cannot expect that playing a game for just a little bit transforms players into experts—even if the game happens to have the best theories of learning in cognitive science integrated with its gameplay. The “law of practice” was not for nothing formulated by Thorndike (Knowles, Elwood F. Holton, & Swanson, 1998; Merriam & Caffarella, 1999) and, as they say, “practice makes perfect.”

Fortunately, unlike simulation exercises, which still require much preparation and facilitation, games allow for this extended practice, but similar to the limited use of games in education (with the notable exceptions of Egenfeldt-Nielsen, 2007; Ke & Grabowski, 2007; Squire, 2004), few studies have used full-fledged digital games to train practitioners for more than just a single workshop (with the exception of a couple of military reports, see Beal & Christ, 2004; Surface, Dierdorff, & Watson, 2007).

Of course, some games do not need extended practice. This would be true for especially small games that are easily understood, quickly picked up, have a simple and clear message, and take a short amount of time to play. *Levee Patroller* is much the contrary and to use this game effectively, I knew I needed to use the principle of “practice makes perfect” and implement it in a way that participants play it more than once.

## 9. *Big Fish in a Big Pond*

About one particular issue a major consensus and a wealth of empirical evidence exists: games work best if they are integrated into a curriculum with other instructional methods and facilitation and instructional support are needed to enable learning with games (Pivec & Pivec, 2008). Using a game is not magic. It requires hard work and this might be overwhelming to some. The teachers/facilitators of some studies using games can attest to this (Egenfeldt-Nielsen, 2007; Squire, 2004).

In the military the need to integrate games into a curriculum and use facilitators is also clear. According to Chatham (2007), who describes his experiences with *Ambush!* among others, one of the “ugly” games-for-training lessons is that

There is no golden disc and no “trainerless trainer” that compels trainees to use it by themselves. Humans must be available to ensure effective training (p. 39)...The mere acquisition of a disc of training software seldom results in effective training. *Ambush!* itself reduces quickly to a free-for-all unless it is used in a setting with an instructor, training goals, and enforced AARs [After Action Reviews] (p. 43).

These AARs or debriefing sessions are, according to many, the key to using games effectively (Crookall, 1995). This is where “spontaneous concepts” generated from playing the game are developed with the assistance from the facilitator and other players into “scientific concepts” (Egenfeldt-Nielsen, 2007). It is that crucial that the checklist for manuscripts to be submitted to the journal of Simulation & Gaming includes the item “adequate discussion on debriefing (if none, adequate explanation of why none)—do not ignore this item.”

Does this mean that games are a “small fish in a big pond” or the “cherry on the cake”? Some games may serve this purpose, to make trainees curious or allow them to finally put to practice what they have learned before. But for others treating them as a small fish may limit their potential. Much like forcing games into traditional curricula, the learned benefits will be masked.

*Levee Patroller* was designed to be a “big fish” and it needed to be treated like that for making it work. With this I mean that the game needed to play a pivotal role in the training. In addition, from previous studies we know that a big fish cannot swim by itself. It needs to have a pond and preferably a big one, because this would increase the benefits from using the game. This explains the relevance of the “big fish in a big pond”-principle.

## 10. *It Takes Two to Tango*

I have done previous research with *Levee Patroller* (Harteveld, 2011) with focus on the design of the game. While designing this game I noticed—similar to others (Frank, 2007; Hussain et al., 2010; Winn & Heeter, 2006; Marsh et al., 2011)—that trade-offs have to be made in the design. For example, should players know up front how many failures they have to find or should we keep it realistic and not tell

them? Should we use a joystick or use the conventional but more tedious keyboard and mouse? Do players need to measure cracks or is this just silly and should it be ignored?

I realized that these design trade-offs, dilemmas, or tensions revolve around three “worlds”: the worlds of Reality, Meaning, and Play. I called them worlds because each one provides a different *Weltanschauung* (German for worldview) on how to accomplish a game and is inhabited by different people, disciplines, aspects, and criteria. Let me elaborate on each world:

- *Reality*: Each game has a relationship with the real world. It is situated in a domain and it involves a certain topic. Disciplines and subject-matter experts are related to the topic. With *Levee Patroller* the disciplines are geo-engineering and crisis management. This world is especially concerned with making sure that the game corresponds correctly to the real world, something that is referred to as *validity* (Feinstein & Cannon, 2001; Peters, Vissers, & Heijne, 1998).
- *Meaning*: Each game aims to attain a value beyond the game itself, be it knowledge, an attitude change, or collecting data. What is attained for determines what type of people and disciplines are involved. With knowledge this concerns teachers and the learning sciences, whereas with data collection this could involve data mining experts and database management. What this world is striving for is to make sure that what happens in the game also finds an application—a value—outside the game. This is called *transfer* (Barnett & Ceci, 2002; Thorndike & Woodworth, 1901).
- *Play*: Each game is first and foremost a voluntary activity which is governed by rules and that includes a clear goal and feedback about the progression toward this goal. A wide variety of game types and genres exist to which a game relates to, from a simple board game with four players to grand-scale virtual worlds with thousands of players. Various actors are affiliated with this world, such as game designers, programmers, and artists, and disciplines, such as game studies and game design. This world is concerned about *engagement* (Oblinger, 2004), the involvement and commitment players have in playing a game.

Various tensions may arise within and between two or three of these worlds. For example, regarding the issue of measuring cracks clients and experts demanded that measuring becomes part of the gameplay. From a game designer’s perspective measuring seems however rather silly, tedious, and pointless. An instructional designer would agree that it might be necessary for raising awareness about the importance of measuring, but would stress that any solution should be meaningful. Showing measurements without any user effort would not create for the needed impact.

These tensions force designers to make trade-offs and in making these I hypothesized that it is fundamental to keep these three worlds in balance. Each world is important in designing game, because it needs to be valid, transfer a value, and be engaging. If any of these criteria are not met, the game collapses. It becomes meaningless and/or unplayable. That is why I wrote “It takes two to tango, but it takes three to design a ‘meaningful game’” (p. 1).

Now if these worlds play an important role in the design of a game, which is something I argued for, then they play—or actually should play—an important role in the evaluation of a game too. That is what this principle is about. Game evaluation needs to occur interdisciplinary, by looking at it from the *Weltanschauung* of Reality, Meaning, and Play.

## What to Expect

In the next level background information is provided of the game about which this book is all about. This level discusses the world of levee inspection, the game's learning objectives, and its game design. From this level it will also become further clear why the game was designed and why a game-based training needed to be designed.

In Level 3 the design of the training and evaluation is discussed. The various methods and methodologies are explained that were used to measure the game's effectiveness and how the game was experienced and played. It also explains how I designed a structured three-week training with a special research version of the game.

The subsequent level, Level 4, is about the setup and implementation of the training. Three water authorities agreed to participate and initially 160 patrollers were recruited. The level details how the training proceeded at each water authority and gives training facts concerning participation and what went right and wrong.

Levels 5 and 6 are about how participants experienced and played the game, respectively. These levels reveal the empirical results that were derived from the game questionnaires and the game data. These levels open the black-box and give a better understanding of what happened throughout the training.

Levels 7 and 8 discuss two methods, the pre- and post-questionnaire and the pre- and post-test, that were used to determine the effectiveness of the training. The questionnaires revealed the characteristics of the participants and their attitudes and perceptions regarding games in general, *Levee Patroller* and the training in particular, and levee inspection above all. The test concerns a new method to measure the phenomenon of sensemaking. Level 8 explains how it works and what its results are.

Levels 9 and 10 provide for a triangulation of all the previous levels and their results. Level 9 details a discussion with the participants at the end of the training and Level 10 explains a number of alternative methods and studies that were done to validate findings.

All the separate puzzle pieces of the empirical levels are integrated in Level 11. This level answers the questions put forward in this level and reflects on the complete investigation. Recommendations and a perspective of the future of game-based training are provided in Level 12.

Before continuing to the next level, I will clarify that in this investigation I did not set out to proof the true believers right and the non-believers wrong. I was out to get a



realistic and substantiated proof of the potential power of games. Although possible researcher bias may have occurred, because I was involved as the lead designer of the game I evaluated, this designer's role was behind me. After completing the design, I wanted to attain the "truth" about its actual use and so I remained critical throughout. I was out to see the positives and the negatives, because from failure we learn, maybe even more so than from success (Schank, 1997).

## Level 2

### Sought: A Professional Hans Brinker

*If history repeats itself, and the unexpected always happens, how incapable must Man be of learning from experience—George Bernard Shaw*

*Climate change is expected to cause more severe and more frequent natural hazards. As our cities and coasts grow more vulnerable, these hazards can lead to disasters that are far worse than those we have seen to date—Ban Ki Moon, Secretary-General of the United Nation*



8%

It is the year 2053. Due to global warming the ice of the North Pole has melted to such a degree that the sea level has risen significantly. Countries that are below sea level, such as the Netherlands, are potentially at risk. Why they are at risk, and the Netherlands in particular, becomes clear on what seemed a beautiful Sunday autumn afternoon. At that day the weather suddenly changed dramatically. In just one night 175,000 hectare of land transformed into a sea as a result of a storm surge. In total 1,836 people died and 72,000 were evacuated.

This is a hypothetical example but it is not unimaginable. And in fact, it already happened. In 1953, the Netherlands experienced its largest natural disaster when it was hit by a storm surge (Slager, 2003). The previously mentioned numbers are the actual facts of this disaster. How can we learn from the experience of 1953 to make sure history does not repeat itself in 2053 or any other year?

The Dutch government immediately responded with the “Delta Works”—a series of gigantic levee constructions. The largest one is about nine kilometers long. These technical solutions, costing billions, has protected the country so far from the “angry waters” of the sea. Inland the danger by the “angry waters” of the rivers and the canals had not been taken care off. This became clear when a flooding almost occurred in 1993 and in 1995 (ten Brinke & Bannink, 2004). With the latter near-flooding about 250,000 people were evacuated, which is most likely the largest evacuation ever in Dutch history (van Duin & Hendriks, 1995).

This time “softer” solutions were thought of, such as giving more “room for the rivers” to enable waterways to expand in times of high water (Ministry of Trans-

port, Public Works and Water Management, 2006) and by installing calamity response organizations dedicated to flooding (Trimension, 2001). But how can these organizations be prepared for something that rarely happens?

The need for preparedness became more evident in the summer of 2005 when the levees broke in New Orleans, Louisiana (Cooper & Block, 2006). That summer Hurricane Katrina hit the southern American coast and led to the costliest and one of the deadliest natural disasters in the American history. At least 1,836 people lost their lives, 1.2 million were ordered to evacuate, and the total property damage was estimated at \$81 billion due to the hurricane and the subsequent floods.

The disaster of Hurricane Katrina posed another wake-up call to the Netherlands. With global warming and other 21st century developments, the Dutch realized that they might be facing their Katrina one day as well (Dykstra, 2009). The impact might very likely be worse than in 1953 and will extend beyond its borders, as the Netherlands has become more densely populated and has acquired many critical infrastructures and industries of international importance (Ale, 2009).

Amidst this background the game *Levee Patroller* was initiated and developed. The game provides for another solution and—to speak with the words of George Bernard Shaw—one that allows Man (and Woman) to become capable of learning from experience.

Because this book is about the evaluation of this game, by using the triadic game design framework the goals of this level are to describe

- The practice of levee inspection, consisting of the water authorities, levee patrollers, and levee failures (Reality);
- The objectives of *Levee Patroller* and how these should be achieved (Meaning); and
- How this 3D simulation game works and what its implications were before the training as well as how it was used (Play).

## World of Reality: Levee Inspection

The disaster of Hurricane Katrina illustrates that the Netherlands is not the only country with flooding issues. Similar inspection practices—with water authorities and patrollers—exist elsewhere too. However, the world of levee inspection is and remains quite unique to this small country. This world has deep historical roots and undoubtedly originated because flood risks have been greater than all other risks combined in this country, of which more than half of it lies below sea level (ten Brinke & Bannink, 2004, p. 11). For example, between 1750 and 1800 alone, there were 152 floods (Rijkswaterstaat, 2011).

As a practice levee inspection is a responsibility of the Dutch “water authorities” and is performed by practitioners called “levee patrollers.” The risks that they are concerned about are the possibility of the occurrence of a “levee failure.” These water authorities, patrollers, and failures make up the world of levee inspection. I will discuss each one of them in more detail to give you the context behind the game.

### **In-depth explanation: gathering data about the context**

The description of the world of levee inspection is not only based on articles, reports, and books. Since I started with the design of the game I have been immersed into the world of levee inspection and made detailed observations on various occasions and events that I attended. In addition, I acquired data by means of a questionnaire and interviews, to confirm some of my thoughts and elaborate on others. This data collection proceeded as follows:

- Throughout the design of the game I had several interviews and informal talks with “inspection coordinators,” patrollers, and subject-matter experts about levee inspection. This was with the intent to translate their world into that of the game. The inspection coordinators are in charge of the calamity response organization.
- Those interviews and informal talks were with five out of 27 water authorities and I wanted to get an overview of how inspection works at all of the water authorities. Because it was impossible to find the data about the other water authorities in articles and reports, I decided to submit a questionnaire to the inspection coordinators of each water authority with basic questions about levee inspection, such as what types of levees they inspect, what types of failures they could encounter, and what responsibilities their patrollers have. I received 22 responses.
- The answers on the questionnaire led to more questions and so I made a follow-up with telephone interviews of about 30 minutes with 11 inspection coordinators.

I realize that most readers will not be interested in the exact details about levee inspection and so I decided to provide only the necessary information in this level.

## ***The Water Authorities from Old to New***

For a long time the water authorities were the only accepted authorities in the Netherlands (Dyckmeester, 1940). They were established in several local areas by farmers to manage the flood defenses. Only by collaborating people could keep their feet dry. The oldest water authorities were established in the 13th century. Their influence extended much beyond the fight against flooding (Kienhuis, Westerwoudt, van der Wal, & Berge, 1993). The water authorities had a social function as well as judicial power. Damaging a levee back in the days could end in your hands getting chopped off. Even the death penalty was sometimes applied to those who did harm to the levees.

The water authorities functioned independently from each other and the government, which explains why each developed its own vocabulary and customs, until the Dutch constitution was implemented in 1848. After that the role and position of the water authorities were questioned and this has continued until this very day. Emperor Napoleon, who occupied the Netherlands from 1795 to 1813, made the first attempt to abolish them, but he met his Waterloo here too (Dyckmeester, 1940, p. 4). But he did manage to take their judicial power away.

Over time, two other remarkable and structural changes have occurred. The first concerns their number. In 1850 the number of known water authorities were 3500

(Rijkswaterstaat, 2011). This number was reduced to 2500 in 1950, to 50 after the major flooding disaster in 1953, and to the number of 27 that we have today.<sup>1</sup> It is expected that this reduction will continue to take place in the nearby future.

The second change is that the water authorities became semi-governmental institutions (de Graeff, van der Heide, Mouwen, & van der Wal, 1987). Although they still have their own democratically elected board and collect their own taxes, they fall under the supervision of the regional authorities (i.e., the provinces). In practice this leads to some difficulties, as the regions of the regions authorities and the water authorities are not one and the same. Certain water authority regions overlap with two or more regional authorities. In addition, the regional authorities have hardly any knowledge about water management. This makes it difficult for them to judge the policies of the water authorities.

By law the water authorities have one specific dedicated task (Kienhuis et al., 1993): to manage the “water system” in their region. A water system is the complete constellation of waters in a region, from the sea, lakes, rivers, canals, to even ditches. Managing this system involves three subtasks. It involves taking care of *a*) water quantity, *b*) water quality, and *c*) water defenses. The latter subtask is why *Levee Patroller* was designed. This subtask is about ensuring that the levees are in good condition.<sup>2</sup> Levee inspections are done to ensure this state—regularly and during emergencies.

These inspections have been done from the beginning of the establishment of the water authorities in the Middle Ages, when farmers and other people already walked over the levees and used the equipment they had at their disposal to keep the levees in good condition. However, it was not until 2001 that a proposal was made to change the ad hoc nature of how the water authorities dealt with emergency situations by installing a calamity response organization which prepares, plans, and coordinates the prevention and after care of calamities (Trimension, 2001). This was done as a response to the floods in 1993 and 1995, from which it was concluded that a “better” organization was needed. With better they meant an organized organization with practitioners who are prepared to deal with the unexpected. Before, it was basically a bunch of individuals put together when needed. In 2004 such a calamity response organization was installed at each of the water authorities. That year a critical (and somewhat cynical) report stated:

With regards to the preparation of calamities an alarming picture is drawn. The scale and fidelity of the exercises is too limited, calamity plans are lacking, and the levee administrators and the crisis management organizations do not work as a team. It is doubtful whether the Netherlands is able to handle a possible large flooding. The realistic “exercise” of 1993 did lead to a considerable improved calamity organization in 1995 (ten Brinke & Bannink, 2004, p. 16).

<sup>1</sup> A peculiar and funny detail is that one of these 27 water authorities, Blija Buitendijks, is run by only one person and has the supervision of an area of 100 hectare.

<sup>2</sup> The water authorities are not responsible for all of the water defenses in the Netherlands. Most of the large infrastructural water works, such as the Delta Works, and large parts of the sea defenses are controlled by the Directorate General of Public Works and Water Management. This is a national agency.

The game *Levee Patroller* was developed as part of this transformation from an informal network of practitioners who come together when needed to a professional organization which is prepared to deal with the worst kind of situations. The game was specifically aimed at transforming the “bunch of individuals” into knowledgeable practitioners—into “professional Hans Brinkers”—and to make it unnecessary to have a near-flooding to “exercise.” As the game is devoted to this specific target group, let us have closer look at what these practitioners are and do.

### ***The Real Hans Brinkers***

Hans Brinker became known as the mythical hero who put his finger in the levee to prevent a flood from occurring. The real “Hans Brinkers” are the levee patrollers who are affiliated with the water authorities and who—like Hans Brinker—walk over and around the levees to prevent floods. When asked, each water authority described patrollers as: “They are the eyes and ears of the organization.” The patrollers are used as an extension of the organization to “sense” what is out there and communicate this to others.

This does not mean that the inspection is similar at each and every water authority and that the patrollers are one and the same. Patrollers differ, for example, in type and responsibilities. Almost all water authorities have *experts*. These are practitioners who inspect levees regularly. Certain water authorities have in addition a number of *employees* who get involved when needed, but perform other tasks within the water authority. Inspection is not part of their job description. Then a number of water authorities make use of *volunteers*. These people have a job somewhere else but they can be called to assist when necessary as well. What types are used differs per water authority.

The number of patrollers also varies widely from 0 to up to 750 practitioners. The numbers differ greatly, because it depends on the region to what extent inspection is needed.

How it is organized is not only dependent on the region. The organizational culture, including past routines, plays a role too. This became apparent to me when two water authorities merged, one that used volunteers and the other that used employees. The managers told me they were investigating how to continue: adopt one or the other for the new organization, use a mix, or keep it like this? They did not have a clue and are still struggling with reorganizing their calamity response organization. In total, the number of patrollers in the Netherlands is said to be about 3,500.<sup>3</sup>

To address the difference in responsibilities, it is first necessary to decompose the main task of levee inspection into subtasks. Patrollers essentially follow these five steps:

---

<sup>3</sup> This is what a subject-matter expert told me. Based on my questionnaire with a response from 22 out of 27 water authorities, I came to the number of 2,564. This is an underestimate, because not every respondent gave a number.

1. Finding failures.
2. Reporting signals that make up a failure.
3. Communicating reports to others.
4. Diagnosing the situation.
5. Taking measures when necessary.

Although how it is done differs per water authority, the first three steps are taken by every type of levee patroller. These steps make up the core responsibilities of patrollers. The first step involves finding the risks that patrollers have to deal with: the levee failures. To find these failures patrollers have to recognize the signals that indicate the possible occurrence of a failure.

Upon finding a signal, a patroller needs to report this signal and look for other signals to get a complete picture of the situation. The patroller may also need to make measurements to accurately describe the size of the signals found. When this is completed, the patroller can initiate the third step, to communicate the findings. At some water authorities patrollers communicate their findings to the field commanders. At others they communicate them straight to a central coordination office called the *Action Center*.

The two other steps require more substantial knowledge in the process of putting the puzzle pieces together, to judge what is occurring and to subsequently act. That is why these steps are at most water authorities privileged to the experts only. The actual execution of measures is largely delegated to construction companies. The water authorities made contracts with these companies that in times of emergency they will provide the personnel and equipment to deal with detected failures.

At most water authorities, the levees receive a regular inspection once or twice per year, before and/or after the storm season, which runs from October to April. This regular inspection is done by the experts, sometimes with the help of others. Inspection further happens when extreme situations are expected. Such situations, which could involve high water, a storm, or drought, happen rarely. To illustrate, one voluntary patroller indicated that in his 48 years of involvement he did a real inspection but once (IPpre-#14).

Patrollers hardly trained before the establishment of the calamity response organization. Knowledge was passed from older workers to newer ones and from father to son. After the establishment of the calamity response organization, water authorities started to organize more activities. Other than social gatherings, which are organized to get to know one another, the following activities are most prevalent:

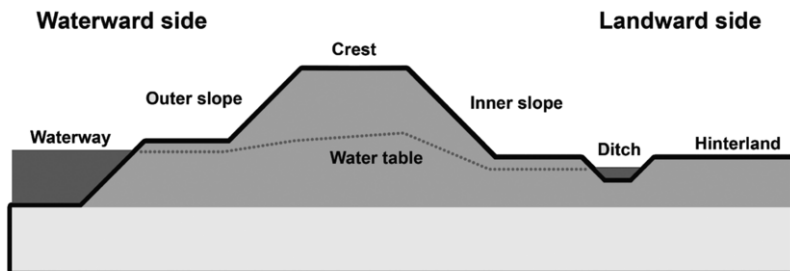
- *Instructions*: In an afternoon or evening patrollers receive information about procedures and failures from an expert.
- *Field trip*: On a Saturday afternoon patrollers go into a bus and get a tour of the region and learn about its specifics, such as what levee segments are vulnerable.
- *Field exercise*: Here patrollers pretend on an evening that it is an emergency situation. They have to inspect particular levee segments in teams and find signs. The signs contain a picture and information of a failure, which the patrollers have to communicate to an Action Center.

The frequency of these activities differs among the water authorities. The more active ones organize field trips once every five years, instructions once every two or three years, and field exercises once every year.

With the creation of *Levee Patroller*, another possible activity was added to the list. This game was not made with the intention of replacing any of the above-mentioned activities but rather offers an added value beyond these activities. Despite (or in fact because of) its “virtuality,” the game provides the only practical means to get experience in finding and reporting “levee failures.”

### ***Rare But Disastrous Failures***

Levees themselves exist in all kinds of shapes and colors. Most of the levees came into existence in a natural way. By excavating soil from the land the land itself lowered in comparison to the waterways running next to it over the centuries. At some point in time the Dutch began to fortify these natural levees with left over materials. Major parts of the Netherlands would otherwise be under water. In addition, over time the inhabitants created artificial levees, out of protection but also to make sure the reclaimed land from the sea would remain dry. Some of these artificial levees are made with natural material, such as sand, clay, and grass. The natural and artificial levees made with natural material are referred to as *green levees* (Fig. 2.1).



**Fig. 2.1** A crosscut of a generic green levee

Besides the green levees, technical constructions exist, such as sluices and the major constructions known as the “Delta Works.” These major constructions and the green levees around the larger waterways, such as the sea, lakes, and rivers, equal a length of 17,500 kilometers (Rijkswaterstaat, 2011). These major levees are referred to as *primary levees*. Tens of thousands of kilometers of smaller barriers exist on top of that, such as around the canals. These smaller ones are referred to as *regional levees*.



All levees share that they have one dedicated function: to protect the land from flooding. If a levee is not in a good condition and, therefore, not likely to maintain its function, we speak of a levee failure. If eventually a levee cannot protect the land from flooding anymore, when it fails completely, a “levee breach” has occurred.

Which failure occurs is dependent on the type of levee, its material composition, the type of waterway, and its cause. How a failure fails is what is called a *failure mechanism*. Step four during an inspection, diagnosing the situation, involves analyzing what failure mechanism is happening. This is important, because what failure mechanism occurs determines what measures need to be taken. Five failure mechanisms are to be distinguished: macro-instability, micro-instability, erosion outer slope, erosion inner slope, and sand boils.

#### **In-depth explanation: delving into the failure mechanisms**

The failure mechanisms relate to two basic ways of failing. To explain these, you have to understand that a levee does not prevent all the water from reaching the hinterland (Fig. 2.1). Waterways remain connected to the hinterland by means of groundwater. The height of this groundwater is called the *water table* (or phreatic surface). The soil below the water table is saturated and the soil above the table is unsaturated. In other words, everything below the table is wet, everything above it is dry. The two basic ways of failing are:

- *Stability*: When the water level changes, the water table changes as well. This change affects the composition of saturated and unsaturated soil and could make the levee unbalanced. If the water level is too low, it could lead to settlements toward the waterward side. If it is too high, settlements could appear toward the landward side. It could also happen that due to the water pressure cracks appear on the crest or slopes, from which the levee further deteriorates.
- *Erosion*: The revetment of a levee, the cover or the most outer layer of a levee on the inner or outer slope, plays an important role in this regard. This revetment protects the levee itself by making sure that the soil underneath does not get washed away when water runs over or splashes against it or when objects, such as floating waste or ships, hit the levee. If the soil gets washed away, the levee slowly degrades and loses its function.

As suggested by the terminology, macro- and micro-instability relate to stability and erosion outer and inner slope to erosion. Sand boils is a special type of erosion, as with this mechanism the levee erodes from the inside out.

What patrollers need to be concerned about above all is to observe the *signals* of a failure. Examples of signals are cracks, water outflow, damaged revetment, and settlement. Every failure consists of one or more of such signals. Based on the description of the signals, a failure mechanism can be retrieved and action can be taken if needed.

Such action is rarely needed. Except for one levee breach in 2003, no serious failure has occurred in the past decades. The failures that did occur were “routine jobs.” Aside from damaged or poorly maintained revetment, this involved damage by human activities, such as dredging, excavation, or construction. These risks were noticed in time and easily repaired.

Although failures rarely occur, it is not accepted that they lead to an eventual levee breach. The consequences are simply too large. To prevent such a “rare but disastrous failure,” it not only requires to have “healthy levees,” it also requires personnel that knows what the risks are and how to deal with them. We are therefore looking for professional Hans Brinkers.

## World of Meaning: Sensemaking

How do the water authorities get their professional Hans Brinkers? In the summer of 2005, visitors of a symposium saw a technical demonstration of the use of gaming technology to visualize levees and came to the conclusion that gaming might provide a potential powerful solution. A game would give the patrollers the needed experience with levee failures. To realize this idea, we defined in consultation with all stakeholders the following learning objectives (in order of importance):

- *Observing*: To recognize signals of a failure.
- *Reporting*: To report in the correct way the observed signals associated with a failure.
- *Assessing*: To recognize the different phases and the severity of a failure.
- *Diagnosing*: To recognize a failure mechanism behind a failure.
- *Taking measures*: To know how a further progression of a failure can be prevented.

These specified objectives indicate that the value of the game seems to involve *knowledge* and *skills*: knowledge and skills about recognizing failures and how to deal with them. This is achieved by means of virtual *sensemaking*. In the game, players make sense of virtual failures and use this to make sense of real ones. But what all of this ultimately should improve is the *communication* between patrollers, between patrollers and the Action Center, and between water authorities. I will now elaborate on this by considering possible learning outcomes in training evaluations, the concept of sensemaking, and how all this can impact the communication.

## The Learning Outcomes

Regarding training evaluations the learning outcomes model by Kraiger, Ford, and Salas (1993) is frequently referred to. They make a distinction into cognitive, skill-based, and affective learning outcomes and note that they “are often interrelated...changes in one learning outcome may imply changes in another” (p. 322). I will use this classification scheme to further explain what *Levee Patroller* tries to achieve. I will start with the one that we focused the least on during the design.

### Affective learning outcomes

Affective learning outcomes relate to attitude and motivation. *Attitude* is a person's "learned predisposition to respond in a consistently favorable or unfavorable manner with respect to a given object" (Fishbein & Ajzen, 1975, p. 11). In training programs this is often measured by means of "trainee reactions," that is how well trainees liked or judged the training. If a trainee is unfavorable against playing games for example, we may expect him or her to dislike a game-based training.

This is however more a measure of the quality of the training than a direct measure of learning. Those attitudinal outcomes that might be impacted by a training are organizational commitment, self-awareness, and changing values (e.g., importance of safety). For example, playing *Levee Patroller* may encourage players to have a different "predisposition" to their environment, levee inspection, and the water authorities. However, in developing the game we did not aim such attitudinal outcomes.

Kraiger et al. (1993) share *motivation* among affective learning outcomes, because it "is also an internal state that affects behavior" (p. 318). Although motivation may be an intended outcome of a training, it most certainly plays an important role at the start and throughout the training (Harteveld, 2011). Without the commitment and willingness of players to invest in learning the subject-matter of the game, it is unlikely that the value will be achieved. This importance of motivation is one of the reasons why games are considered potential powerful tools. It is suggested that games are effective, because they increase player's motivation (Garris et al., 2002; Malone, 1981; Malone & Lepper, 1987b, 1987a) and *Levee Patroller* may, therefore, encourage trainees to become committed and willing to learn the material.

As an outcome motivation could engender a concern for increasing one's competence pertaining to the subject-matter (mastery orientation); an intention to do well or better (performance orientation); confidence in having learned the information taught and being able to perform well (self-efficacy, see Bandura, 1977); and of setting goals of exerting further effort into learning the material (goal setting). Similar to attitudes, in developing the game we did not specifically aim at accomplishing these effects. That does not mean they are not valuable or influential. As Kraiger et al. (1993) stress, the outcomes are often interrelated and, therefore, attitudes and motivation will impact the cognitive outcomes.

### Cognitive learning outcomes

The cognitive learning outcomes refer to "the quantity and type of knowledge and the relationships among knowledge elements" (Kraiger et al., 1993, p. 313). A distinction is traditionally made into *declarative knowledge* and *procedural knowledge* (Anderson, 1983, 1993). Declarative knowledge is about "knowing that." It concerns "knowledge elements" such as facts, events, and sequences that we can *verbalize*, that is make known to others. The relationships among facts and information, how knowledge is organized, also forms part of this. Some such as Kraiger et al. like to

point out this difference by for example referring to knowledge elements as “verbal knowledge” and to the relationships among knowledge elements as “knowledge organization” (or mental model). In *Levee Patroller*, we can identify the following declarative knowledge items:

- *Inspection concepts and vocabulary*: Players learn what concepts are relevant to levee inspection and how to label them, from failure signals to failure mechanisms.
- *Mental model formation of failures*: Players learn what failures exist, where they can occur, how to recognize them, and what failure mechanisms and measures relate to them.
- *Mental simulation of failures*: Players learn how a failure develops over time. Based on this they can develop expectations.

Procedural knowledge is about “knowing how.” It refers to knowledge about how to perform a task or action. This knowledge is often implicit, therefore more difficult to articulate, and makes use of declarative knowledge. It consists of IF-THEN constructions, such as IF an English verb needs to be written in the past tense THEN add -ed unless it is an irregular verb.

With *Levee Patroller*, players should learn the task or action of how to inspect failures. They should learn what they need to pay attention to when encountering a certain failure (IF I see failure X THEN I need to look at this and that) and what steps are necessary in reporting a failure (mention the location, notice the failure characteristics, and then call the Action Center). To facilitate the construction of this knowledge, the game teaches players an *inspection protocol*, which is a set of rules and procedures that must be followed. This protocol is based on a prototype checklist form developed in 2005 by the Foundation for Applied Water Research<sup>4</sup> and on expert heuristics that I elicited in my conversations with subject-matter experts (e.g., IF soil starts to flush THEN the failure becomes riskier).

Patrollers need to articulate what they see. How well their procedural knowledge may be developed, they should above all be judged on their ability to describe what they see. This leads into the third outcome: skills.

### Skill-based learning outcomes

Skill-based learning outcomes concern the development of motor or technical skills. Although playing *Levee Patroller* may result in improving players’ skills in playing games or their hand-eye coordination, this is not its intent. Its intent is to develop technical skills. These skills are specific to a particular occupation or group of occupations in the performance of a particular task. I further defined skills as the ability to apply what is learned (Harteveld, 2011). This means that players do not simply recall facts or images from the game, but actually use that information/knowledge to perform on a particular task.

<sup>4</sup> In Dutch the Foundation for Applied Water Research is known as the “Stichting Toepast Onderzoek Waterbeheer” (STOWA).

With levee inspection, this particular task concerns dealing with levee failures, something with which patrollers have minimal experience. What the game offers is the development of *sensemaking skills*, the ability to make sense of (virtual) failures—which includes distinguishing failures from non-failures and the ability to articulate what is made sense of. Based on the learning objectives we can decompose these sensemaking skills into skills for observing, reporting, assessing, diagnosing, and deciding what measures to take.

If skills become sufficiently developed, “Individuals also learn to apply newly learned behaviors to unique settings (generalization) and to modify existing skills depending on the situation (discrimination)” (Kraiger et al., 1993, p. 317). Relating this to levee inspection and the game, it means that if skills become enough developed, players may be able to make sense of failures that are slightly different from the ones they practiced with or even make sense of completely new failures.

### *The Process of Sensemaking*

The concept of sensemaking has not been related to games so far.<sup>5</sup> It has, however, been related to disasters (Weick, 1993), communication (Dervin et al., 2003), decision making (Klein, Moon, & Hoffman, 2006a, 2006b), performance of military command teams (Jensen, 2009), use of information technology in organizations (Orlikowski & Gash, 1994), human-computer interaction (Russell, Stefik, Pirolli, & Card, 1993), and modeling and simulation (Tolone, 2009)—all areas that closely relate to *Levee Patroller*. It is broadly defined as a process by which people give meaning to experience or as how people make sense out of their experience in the world. More specifically it is about “such things as placement of items into frameworks, comprehending, redressing surprise, constructing meaning, interacting in pursuit of mutual understanding, and patterning” (Weick, 1995, p. 6) or about “a motivated, continuous effort to understand connections (which can be among people, places, and events) in order to anticipate their trajectories and act effectively” (Klein et al., 2006a, p. 71).

Consistent with Dervin and Naumer (2009), I noticed four major contributions in sensemaking, each in a different context (Harteveld, 2009). The contributions concern the work by Russell et al. (1993) in human-computer interaction, Klein et al. (2006a, 2006b) in cognitive systems engineering, Weick (1995) in organizational communication, and Dervin et al. (2003) in library and information science. Although these major contributions differ in perspective and focus among others, I further noticed that they share many similarities. Each speak of a process, challenges, the importance of context, creating order from chaos, and about the (re-)construction of knowledge.

---

<sup>5</sup> I found one dissertation in Dutch by van der Meer (1983) who used “social simulation” to investigate how sensemaking takes place in organizations.

### **Sensemaking is a process and not a product**

Weick (1995) warns us to not confuse sensemaking with “interpretation,” which is often used as a synonym for sensemaking. Whereas interpretation can be a process as well as a product, sensemaking is about an activity or a process:

It is common to hear that someone made ‘an interpretation’. But we seldom hear that someone made ‘a sensemaking’. We hear, instead, that people make sense of something (p. 13).

In addition, interpreting is one of the steps that make up sensemaking. The (iterative) process of sensemaking consists of deciding (mostly unconsciously) of what to focus on and how to look at the world, extracting the important cues from the environment, constructing meaning based on these cues, making a (re-)interpretation, and implementing a certain action.

Like Weick, each contribution focuses on the process. It is not about what happened, but rather “how” what happened. How people build bridges over information gaps, organize themselves, make decisions in complex, real world tasks, or retrieve information is what is important. It is about what occurs “in between.”

And like sensemaking, play is not a product. This is a process too and one in which meaning is constructed as well, by deciding on what to focus and look at in the *gameworld* and extracting the important cues from the game environment, and so on. A game could for this reason be redefined as a product that allows for (virtual) sensemaking to take place. In case of *Levee Patroller* players experience the process of finding and reporting failures.

### **Sensemaking does not occur in a vacuum**

Sensemaking occurs within a particular context. The history, culture, identity, existing mental models, the task at hand, the information systems used, all of this and more determines how people make sense of a situation. This story about the merchant Marco Polo makes this clear (Eco, 1998).

When Marco Polo traveled to China, he was obviously looking for unicorns... On his way home, in Java, he saw some animals that resembled unicorns, because they had a single horn on their muzzles, and because an entire tradition had prepared him to see unicorns, he identified these animals as unicorns. But because he was naive and honest, he could not refrain from telling the truth. And the truth was that the unicorns he saw were very different from those represented by a millennial tradition. They were not white but black. They had pelts like buffalo, and their hooves were as big as elephants’. Their horns, too, were not white but black, their tongues were spiky, and their heads looked like wild boars’. In fact, what Marco Polo saw was the rhinoceros (p. 55).

If you were to travel to Java, you are unlikely to make this mistake. That is because of the history, culture, and knowledge of this time. However, if you were to play *Levee Patroller* you might not make the same decisions as someone else. What is a dangerous situation according to one might be completely safe to another. Personal characteristics, experiences, and other contextual influences determine how sense is made—also in games.

## Sensemaking happens when challenged

Klein et al. (2006a, p. 72) say that “sensemaking does not always have clear beginning and ending points.” What is clear is that it is often triggered when regular routines are interrupted, when we are confronted with information gaps, when we are in ambiguous, equivocal, or uncertain situations, when we have to make decisions, or when we have to look up information and process this. In each activity, a challenge is defining what is being done. If no challenge exists, no need exists to make sense.

This is also one of the reasons why gaming is an interesting tool with regards to sensemaking. Games are about challenges (Rollings & Adams, 2003). For this reason, Crandall, Klein, and Hoffman (2006) suggested the following:

Computers are being increasingly used to create gaming environments and to present humans with varieties of experience in simulated settings and artificial worlds. All of these settings offer the potential for putting humans into cognitively complex and challenging circumstances in order to understand how we perform tasks, make sense of what is going on, act, and react (p. 19).

Instead of studying humans we could help them in dealing with “cognitively complex and challenging circumstances.” That is the intent of *Levee Patroller*. The game further challenges players with failures that they have never seen before and puts them therefore in a situation where they have to make sense of what they see.

## Sensemaking is creating order from chaos

Every contribution reconceptualizes human beings as passive receivers toward ones that actively influence their environment. Human beings are “authors” of their sensemakings and “make and unmake, develop, maintain, resist, destroy, and change order, structure, culture, organisation, relationships, self (Dervin, 1999, p. 45). Making the world more orderly—structuring it—is what players do too and what makes games different from for example music videos according to Johnson (2005):

To non-players, games bear a superficial resemblance to music videos: flashy graphics; layered mix of image, music, and text; the occasional burst of speed, particularly during the pre-rendered opening sequences. But what you actually *do* in playing a game—the way your mind has to work—is radically different. It is not about tolerating or aestheticizing chaos; it is about finding order and meaning in the world, and making decisions that help create that order (p. 62).

Nevertheless, structures, such as routines, categories, and cultural values, are inscribed into people and they are of influence as well. This becomes clear with the story of Marco Polo. Entire traditions steered him in seeing the rhinoceros as a unicorn. So people are active agents exerting influence, but are influenced by structures too.

This steering of how sense is made is called *sensegiving* by Gioia and Chittipeddi (1991), who define it as “the process of attempting to influence the sensemaking

and meaning construction of others toward a preferred redefinition of organizational reality” (p. 442). With this in mind, we could conceive of games as “sensegivers” if they attempt to steer players into a preferred direction. With *Levee Patroller* this is very much so. It provides a structure of what failures exist and how they need to be recognized and dealt with.

### **Sensemaking leads to the (re-)construction of knowledge**

Sensemaking is a process and not a product, but it ultimately produces something: knowledge. Dervin (1998) states this explicitly and sees no distinction between knowledge and information (much like others in the behavioral and brain sciences; see for example Dienes & Perner, 1999). She considers sensemaking as “information/knowledge as product of and fodder for sense making and sense unmaking” (p. 36)

Others refer to the (re-)construction of mental models, frames, cognitive maps, representations, schemata, knowledge structures or any other similar concept. What these concepts have in common is that they are about how information/knowledge is organized, structured, stored, and represented mentally (Kitchin, 1994; Spicer, 1998). Experts have more elaborate models than laypeople and that is why they recognize patterns faster and more accurately, conceive more detail, and above all, know more than others about a certain subject (Anderson, 1995; Chi, Feltovich, & Glaser, 1981).

Playing a game such as *Levee Patroller* should lead to the (re-)construction of knowledge based on the experience with virtual risks and make players into experts. This (re-)constructed knowledge can subsequently be used as “fodder” in making sense of real risks.

### ***Impact on Communication***

The game was not developed with the purpose of improving communication. As a matter of fact, in the development of the game we explicitly decided to leave out, demarcate, and limit communication processes (Harteveld, 2011). However, I increasingly started to realize that although we limited communication in the game, the game might still indirectly make an impact on communication. The first explanation is relatively simple. Sensemaking is “an issue of language, talk, and communication” according to Weick, Sutcliffe, and Obstfeld (2005, p. 409), who quote Taylor and Van Every (2000):

We see communication as an ongoing process of making sense of the circumstances in which people collectively find ourselves and of the events that affect them. The sensemaking, to the extent that it involves communication, takes place in interactive talk and draws on the resources of language in order to formulate and exchange through talk (or in other media such as graphics) symbolically encoded representations of these circumstances. As



this occurs, a situation is talked into existence and the basis is laid for action to deal with it (p. 58).

Levee patrollers talk a failure “into existence” and create the basis “for action to deal with it.” They talk with each other, because they almost always patrol in teams, and then communicate their findings to someone else. During these conversations the patrollers need to make clear what they are seeing via (until now) language only and not by other media. The language they use provides for a “symbolically encoded representation” of the failure situation.

For example, one patroller told me that once he found a failure and he called up the water authority.<sup>6</sup> They did not believe him and said this did not need their immediate attention. He disagreed. A couple of days later he called them again and this time they decided to look at it. Not much later the area surrounding the failure situation was fenced off and the water authority started repairing the levee.

The game will obviously not prevent situations like this, but it does provide for a *common vocabulary* and *shared experience* that will provide for a foundation for future interactions among patrollers and between patrollers and field offices. Speaking the same language and having a similar experience to draw upon will more likely lead to sharing the same meaning, as to whether or not we are dealing with a failure and action is needed. To improve organizational structure and behavior Weick (1995) encourages the creation of a shared experience:

If people want to share meaning, then they need to talk about their shared experience in close proximity to its occurrence and hammer out a common way to encode it and talk about it. They need to see their joint saying about the experience to learn what they jointly think happened. This may be why outdoor adventure retreats seem to be a successful means to build teams. There is novel, joint experience for which no one has a ready label, and which tends to be made meaningful, on the spot, with a common vocabulary, while the joint experience is still fresh in everyone’s mind. People construct shared meaning for a shared experience (pp. 188–189).

If we replace “outdoor adventure retreats” with a game such as *Levee Patroller* then we see may be why games could be a successful means to improve communication. Although players may not share the same meaning from the game, because each player may make sense of the experience differently, they do share the exact same experience.

The second explanation is based on the communication between water authorities. Water authorities have to increasingly work together and share information and experiences about levee inspection. Such collaboration and sharing is hampered because every water authority has its own organizational culture and its own vocabulary. This is evident in how they call their “levee patrollers”: some say *dijkwacht* or *dijkbewaker* and others *wachtloper* or *dijkinspecteur*.

With this in mind, we could conceive of every water authority as a *community of practice*, because they represent “a set of relations among persons, activity, and world, over time and in relation with other tangential and overlapping communities of practice” (Lave & Wenger, 1991, p. 98) and have developed “their own practices,

<sup>6</sup> Participant #31 told me this anecdote when I visited him to bring a loan laptop.

routines, rituals, artifacts, symbols, conventions, stories, and histories” (Wenger, 1998, p. 6). *Levee Patroller* is then an artifact developed to be used by many of these communities and as such it functions as a *boundary object* (Star & Griesemer, 1989). Boundary objects mediate interactions between different communities of practice by providing a common basis for conversations and enabling knowledge exchange across organizational and professional borders (Carlile, 2002). They may also become a vehicle for innovation despite of possible professional and organizational barriers (Dodgson, Gann, & Salter, 2007).

## World of Play: 3D Simulation Game

The previous made clear that *Levee Patroller* is a game grounded in the world of levee inspection, which is affiliated with the domain of safety and crisis response and which is a world consisting of water authorities, levee patrollers, and levee failures. It also made clear that the game is related to the world of sensemaking, a world that tries to achieve the value of knowledge by means of making sense of challenges in a game environment and which may impact the communication between constituents of an organization and even between organizations.

This section will examine the game from the perspective of the world of Play, because what did not become clear is the design of *Levee Patroller*. I will further discuss what implications it made in the field of levee inspection and how the game has been used so far by the water authorities.

## The Game Design

Although *Levee Patroller* draws upon many genres, it is first and foremost a 3D simulation game. The player is situated in a 3D realistic game environment in a “first-person” perspective. In this perspective players control an *avatar*, the user’s representation in a game, but do not see the avatar’s body. Players look at the game from the avatar’s own eyes. This viewpoint is often used in flight and racing simulators and in First-Person-Shooters (FPS), such as *Unreal Tournament*. The game also uses the technology behind the latter FPS game, the “Unreal Engine,” a *game engine* that helps in the creation and development of games by providing a software framework with many core functionalities, such as sound, animation, and artificial intelligence.

Players assume the role of a levee patroller and have to navigate a game environment (or *gameworld*) that mimics the Dutch landscapes (including windmills and greenhouses). The current version has four different regions that players can access and ten failure types. Each region has particular characteristics regarding the types of waterways, levees, and failures (which relate to each other, because the type of waterway determines the type of levee and this determines the types of failures that

could occur). The game includes a training exercise option, where players learn how to play the game; a full game option, which includes a sequence of exercises in increasing difficulty; and a scenario generator, which allows users to create their own exercise by choosing a region, their responsibilities, the weather, and what failures to include.

Like most simulation games, the game stays close to reality (Rollings & Adams, 2003). However, many trade-offs have been made such as that players patrol on their own and not in teams. The weather and other contextual variables that may influence failures are not considered too. This has only a visual impact. For example, rain results in reduced visibility. Like most simulation games, it also lacks a narrative within the game. No princesses are to be saved or monsters to be slaughtered.

Unlike typical simulation games it does have a clear goal (Juul, 2005). The goal of every exercise is to find all failures, report and diagnose these, and—if necessary—take measures to prevent them from flooding the region. An exercise ends when all failures are dealt with, a levee breaches, or time runs out. Time runs out after approximately 24 minutes. Even levee patrollers need to take a break some time.

To find the failures players need to navigate the gameworld. Once they find one, they need to report it. When the exercise ends, players receive a score and feedback about their performance.

### **Navigating the gameworld**

To navigate the gameworld players have to use the keyboard and mouse in a similar way as most FPS games. To explain this navigation to patrollers I always say that they need to think of the arrow keys as their feet and the mouse as their eyes, because the arrow keys allow players to move around and the mouse to look around.

With a click on the mouse button players can access their inventory. This inventory consists of a map, handbook, notebook, statistics tool, and reporting and measuring marker. The *map* serves several purposes. First, it gives an overview of the region. Players are able to see where the waterways and levees are. Second, it shows where the player is located in the region. In this way, players can orient themselves and navigate themselves to where they want to go. Third, if failures are found these are shown on the map. By clicking on a failure on the map (represented as a mini reporting marker) players go immediately to that failure. This was implemented to save players considerable time (and frustration) in going back to a failure.

The handbook, notebook, and statistics tool provide various information to players. The *handbook* is based on the original handbook of levee inspection and gives information about what it means to be a levee patroller. It further provides a short documentation about failure signals and mechanisms. The *notebook*, on the other hand, allows players to make notes and look back at these at a later moment. It also keeps automatically track of all measurements that have been made. The *statistic tool* gives information about the score players have at any time during the exercise.

The reporting and measuring marker are tools to report failures. The *measuring marker* (or yellow marker), as its name suggests, allows to measure failure signals. Players need to place one marker at one end of what they want to measure and then place a second marker at the other end. After placing the second marker, the length and height is automatically calculated.

Players need to use the *reporting marker* (or red marker) to indicate that they found a failure. Placing this marker results in another menu, the report menu, and with this they can report the failure.

## Reporting failures

Similar to the inventory, the report menu features several items. Through this menu players can report the location, signals, failure mechanism. They can also contact the computerized Action Center and take measures after approval by the Action Center. Reporting the *location* involves marking where in the region the failure is located. This happens on a map similar to the one from the inventory, except that here players do not see where they are. This was implemented to increase the situation awareness of players. In practice, problems often occur with the reporting of failure locations.

A failure consists of one or more signals and players need to make a report for each signal they observe and for each change in a signal that they reported earlier. Making a report involves filling out a check list with questions pertaining to a particular signal. For example, to make a report about a crack players need to report its type (parallel or perpendicular?), length, width, and on what surface it occurs (asphalt or grass?). With water outflow or a settlement other questions need to be answered.

Once players report the location and filled out at least one report, they can contact the *Action Center* to communicate their findings. Otherwise they will receive a critical remark from the Action Center stating they cannot assist with matters they have no information about. During the conversation players need to decide about the severity of the situation. They do this by selecting from a drop-down menu whether they find the situation

- *Reportable*: No direct danger. The failure does need to be reported and checked to see if it remains stable or if it becomes worse.
- *Severe*: Danger exists. The failure needs to be checked regularly to see if it remains stable or if it becomes worse.
- *Critical*: Immediate danger! The levee is about to breach. Measures need to be taken right away.

After assessing the situation, players receive an explanation from the Action Center from how they perceive the situation severity based on the information received from the player. This is where the expert heuristics are used. If the assessment by players is different from the Action Center, players are urged by the Action Center to look at the failure again and call them back to tell if they changed their reports or their assessment—or if they simply disagree with them.

Failures can change over time and must be checked if they remain stable or become worse. If a failure becomes worse, it means that existing signals change or that new signals appear. Changes as well as new signals have to be reported and communicated to the Action Center. The Action Center will remind players about checking a failure if they have not provided a status update for a while.

Players can make a *diagnosis* at all time. This involves determining what failure mechanism is occurring. They make this decision by selecting between the five failure mechanisms discussed earlier: macro- and micro-instability, erosion inner and outer slope, and sand boils. Like with the diagnosis, with *taking a measure* players need to select from a list of possibilities, but this time they are only allowed to do this after approval by the Action Center. We implemented this to highlight to players that they cannot decide on taking measures on their own. The water authorities were worried that the game might provide this message otherwise.

A levee expert will appear after deciding on a measure, and will explain why or why not the measure was successful.

### Getting a score

At the end of the exercise the final score is shown. The scoring system is based on the following seven criteria:

1. *Found failures*: This relates to the number of failures that are found ( $W = 10$ );
2. *Location accuracy*: Patrollers need to be aware of the location of a failure, so other units can easily find the failure ( $W = 1$ );
3. *Observed signals*: Besides finding the failure, patrollers have to recognize the signals of a failure ( $W = 5$ );
4. *Reporting accuracy*: This criterion assesses to what extent the reports are correct. Patrollers receive points if they report 70% or more of the reporting items correctly ( $W = 1$ );
5. *Assessment accuracy*: This relates to the situation assessment which the Action Center requires patrollers to make during each conversation about a failure ( $W = 2$ );
6. *Diagnose accuracy*: This relates to the identification of the failure mechanism of a failure ( $W = 5$ );
7. *Measure effectiveness*: This relates to choosing the correct measure ( $W = 5$ ).

The total number of points is calculated by multiplying each criterion with a weight ( $W$ ) and then summing the products of these multiplications. This calculation is done because not every criterion is equally important. Finding a failure is, for example, weighted ten times more important than accurately reporting the location. By dividing the total score by its maximum, a percentage, representing the final score, is presented.

Besides the final score, another factor of importance plays a role in judging players' performance: whether or not a levee breach occurred. The levee patroller's task is above all to prevent this from happening and so when one occurs we cannot say

the performance was up to the mark. Therefore, if a breach occurs the performance is always judged as “insufficient.” In all other cases the percentage determines the judgment and as follows:

- < 55% = insufficient;
- 55–70% = sufficient;
- 71–85% = good;
- 86–100% = perfect.

### ***Designing: Recreating Reality***

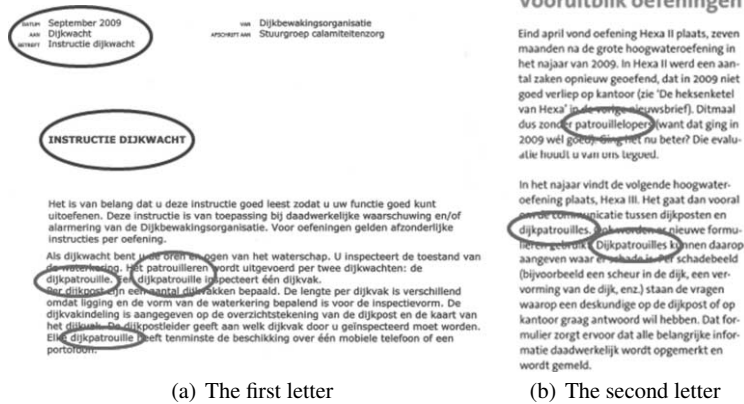
Game design is a creative effort. It requires to integrate, synthesize, and balance various perspectives into one, that of a game. Reality is not simply modeled. It is recreated and this recreation could result in new and fresh insights about that same reality. Before the start of the evaluation I encountered a number of instances where this becomes clear. Such instances also demonstrate the game’s impact on communication, its role as a boundary object, and its use as a vehicle of innovation.

One instance relates to the name of the game. One of the hardest parts and most discussed design issues during the design was actually how to name the game. The Dutch name of the game, *Dijk Patrouille*, was invented because we wanted a name that fit well with the associations most people have with games: active, engaging, and exciting. The existing Dutch words that refer to patrollers are much the contrary. Most of them have *wacht* (in English, “guard”) as part of their name, as in *wacht-loper* or *dijkwacht*. Furthermore, the word *wacht* is closely related to the Dutch *verwachten* which means “waiting.” In one of the brainstorm sessions we thought of *patrouille* which has a much more active connotation. Patrollers do not guard and wait, but are actively on patrol in search of failures.

Over the years I received numerous newsletters from the water authorities and to my own surprise I started seeing the term *dijk patrouille* or *dijkpatrouille* appear in those letters from two of the water authorities I worked with. They mentioned these words without any relationship to the game. It was used to refer to their levee patrollers. Figure 2.2 show two of those letters. In the upper left of the first letter the term *dijkwacht* is still used, but the rest of the letter uses the game’s term.

Another instance relates to the inspection protocol and then in particular the reporting procedure that is used. One water authority decided to adopt this reporting procedure and use it in their actual practice. They are of the opinion that the game introduces a great structure as to how failures need to be reported and decided to translate to their own practice.

What is interesting is that in translating the reporting procedure to their own practice the game inspired them to “score” failures. Each option on their adopted reporting form has a number of points attached to it. For example, regarding the question “What is the water quantity?” the option “little” gives 0 points and the option “much” gives 5 points. By adding all these point a total score is achieved and



**Fig. 2.2** Two letters that use the name of the game from two different water authorities

this gives an indication of the severity of the situation—with the more points, the more severe the failure.

The reasoning behind this *gamifying* of failures is that they want to objectify observations. Otherwise, the Action Center needs to rely at first at the patroller's assessment and then we get into the issue of language. I consider this gamifying because this is an example of applying game principles to a non-game context (Deterding et al., 2011).

## The Actual Use

Although these two specific instances demonstrate the early impact of *Levee Patroller*, in the end we want to know if the game accomplishes its purpose: to turn practitioners into professional Hans Brinkers. Here, I noticed that much work remains, highlighting the often challenging gap in game production between concept and execution.

Immediately after its release one of the water authorities enthusiastically started to make use of it, on their own initiative. They created a special room in their office building and called this the “game room,” a small room with about ten PCs, each with *Levee Patroller* installed.

I attended a number of their sessions. There I noticed that participants enjoyed it, but were not able to play the game enough to get much out of it. Time was even too short to finish one exercise. In addition, I noticed that the facilitators were not knowledgeable enough about the game and levee inspection. With much confidence and conviction they were explaining the wrong things to the participants.

Based on this, I suggested to organize “train-the-trainer” sessions, whereby we would educate the facilitators about the use of the game and its content. Although we received some positive responses about this idea, it was never organized. The feedback we received is that the water authorities were still struggling in how to use the game. The other water authorities did not get much farther than giving the game to some of their key personnel and operating bases to start a discussion on how to use it.

Another use of the game is as part of the levee inspection instruction course. The research institute Deltares provides these instruction courses at the request of some of the water authorities. The instructors have tried to integrate the game within the course. However, they quickly realized that the allotted time for the course was too short to make fully use of the game. Currently, the game’s visualizations are used as part of the course sheets and occasionally participants can get acquainted with the game at the far end of the course.

In response to the infrequent and insufficient use of the game, I revisited the design, this time with the intention to create not a game per se, but an innovative training with a game.

## Lessons Learned

One of the recommendations from the literature reviews is that a common taxonomy is needed, a taxonomy that helps us to clearly define what is exactly being studied (Level 1). In my opinion, it helps to define games according to the three perspectives of triadic game design: Reality, Meaning, and Play (Fig. 2.3). This is what I have done in this level.

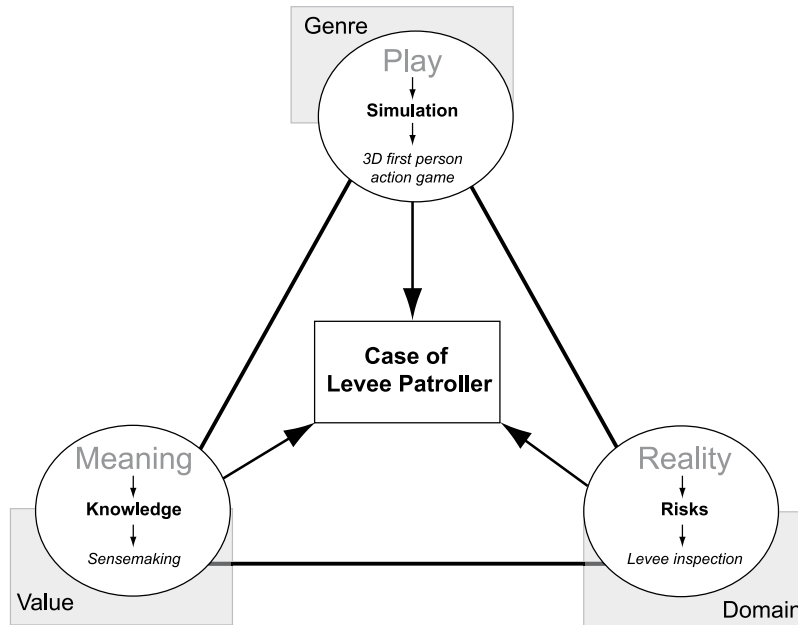
From the perspective of Reality we delved into levee inspection and showed important aspects that relate to it, such as the water authorities, the failures, and—of course—the patrollers. It became clear that levee inspection has a long history in the Netherlands and that much variety exists among the water authority organizations. Another important fact is that failures occur rarely, which is one of the most important reasons the game was developed. Furthermore, we can identify different types of patrollers: volunteers, employees, and expert employees.

Based on the world of Meaning, we have been able to define what it needs to bring forth, how the game does this, and what it could possibly impact. To explain this I introduced various concepts and I will represent here how all these concepts relate to each other.

*Meaning* This is one of the worlds in triadic game design. It refers to the world that is preoccupied with bringing forth a certain “value.” Games could be used for different values, such as data collection, theory testing, but also for transferring knowledge.

*Transfer* Games, at least serious ones, are played to make an impact on the real world. To make an impact, something from playing the game needs to shift





**Fig. 2.3** The case of *Levee Patroller* according to triadic game design.

from the virtual to the real world. What is supposed to be transferred depends on the value the game aims to bring forth. For *Levee Patroller* this concerns especially knowledge and skills.

**Knowledge** It concerns the outcome of playing a game, something that needs to be transferred to the real world. Knowledge is equated with information and with “cognitive learning outcomes” as defined by Kraiger et al. (1993). The knowledge to be transferred from *Levee Patroller* are inspection concepts and vocabulary, mental model formation of failures, mental simulation of failures, and an inspection protocol.

**Skills** Are the ability to apply what is learned. It involves the use of knowledge to perform on a particular task. In *Levee Patroller* technical skills are taught, which are referred to as “sensemaking skills.”

**Affect** Relates to moods, feelings and attitudes, and the affective learning outcomes therefore focus on motivation, attitude change, and self-efficacy. Although *Levee Patroller* may accomplish affective outcomes, it was not designed for this.

**Sensemaking** Is a process that leads to the (re-)construction of knowledge and happens especially when people are challenged and want to create order out of chaos. This is the process that leads to the learning outcomes of *Levee Patroller*.

**Communication** Is about sharing and conveying information among people. Acquiring knowledge and skills from a similar sensemaking process may have an

impact on this. The common vocabulary and shared experience derived from this will be baseline in how practitioners converse with each other and may even mediate interactions between different communities of practice.

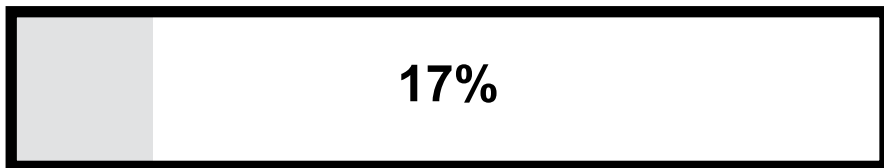
Finally, using the perspective of Play we can define *Levee Patroller* as a simulation game because its virtual world and gameplay follow reality closely. In the game players take up the role of a patroller and have to look for failures, report them, and possibly deal with them before it is too late. Although the game seemed to have an impact on the field by just being created, its actual use is lacking behind, which is one of the reasons I decided to design and implement my own training.

## Level 3

### Toward an Innovative Training/Evaluation

*My apologies for finishing week 1 too late—Participant #103 on the game questionnaire*

*This is again such a question that does not make any sense. At my own levee I know everything. If you put me somewhere else, then I do not of course—Participant complains when filling out the questionnaire*



*Levee Patroller* is an example of a game which was purposefully designed for sense-making. This game was created so practitioners are able to make sense of levee failures. Back in 2010 we only did not know if it worked—which is due to the mere fact that it was barely used and most certainly not as it should have been used.

In this level I explain what *training/evaluation* I came up with to fulfill this gap based on the ten evaluation principles as mentioned in Level 1 and the characteristics of the game in Level 2. I speak deliberately of a training/evaluation, because what I am about to describe is a design of the training with *Levee Patroller* as well as a design of its evaluation. Both designs are tied together, but have slightly different purposes. With the training the purpose was to improve the trainees; with the evaluation the purpose was to get “objective” results about how the game improved the trainees. These different purposes led to some inherent tensions in the design and execution of the training/evaluation.

The term training/evaluation also stresses my double-role. On the one hand, I was the facilitator of the training. As part of this role I had to ensure the learning objectives were achieved. On the other hand, I was the evaluator. This role prescribed that I would systematically retrieve data according to established scientific standards in order to judge the game’s effectiveness.

The goals of this level are to describe

- The motivations and strategy behind the evaluation;
- The working hypotheses developed to guide the evaluation process; and

- The setup of the training with *Levee Patroller*.

## Evaluating a Futuristic Scenario

From the previous level we can retrieve that for *Levee Patroller* to be considered effective, it needs to transfer knowledge and skills about levee inspection. Possibly it may impact communication too. In a nutshell, the major hypothesis behind the game-based training is then that after the training participants will have increased their knowledge and skills and are able to communicate better with each other.

To assess this rigorously and comprehensively, the evaluation needed to be more “more than the tip of the iceberg” and consider “the icing and the cake.” To consider how such an evaluation takes place, I decided to turn to the idea of triadic game design (TGD) based on the evaluation principle “it takes two to tango.” From there I developed an evaluation strategy grounded in *mixed methods research* and *quasi-experimental design* to proof whether the game indeed transfers knowledge and skills about levee inspection.

## Inspiration from Triadic Game Design

Again, the design of *Levee Patroller* demonstrates the triadic interrelationship of Reality, Meaning, and Play:

- The world of Reality relates to the people, disciplines, aspects, and criteria of the domain and subject-matter of the game. With *Levee Patroller* the domain is safety and crisis response (or risks) and its subject matter “levee inspection.”
- As we also understood from Level 2, the game aims to make players more knowledgeable about the topic of levee inspection by training them in “inspection knowledge” and “sensemaking skills.” These knowledge and skills will help them in dealing with actual failures. In the far end it may even have an impact on the “communication” in the organizations. The world that is preoccupied with this creation of value is Meaning and various people and disciplines relate to this too.
- Then this 3D first person simulation game belongs to a world that is concerned with the creation and study of these types of (digital) phenomena. This makes up the third world, that of Play.

It seemed reasonable to assume that if these worlds played an important role in the design, they would play an important role in the evaluation too. What I wondered is how I could use this design framework to evaluate “players” (*player-centered approach*) and determine the “outcomes” of a training (*outcomes-based approach*). In developing the evaluation TGD was not the only source of inspiration. I used other models and frameworks for developing a *logic-based approach*.

### Toward a logic-based approach

A first decision was to make the evaluation logic-based, because to explain for causal relationships it helps to theorize up front about the sequence or flow of events in an evaluation program and the relationships between them. This is exactly what a logic-based approach does. It attempts to highlight how the desired outcomes are achieved. TGD is static and does not reflect a sequence or flow of events. It only represents aspects to be considered. However, many other existing frameworks have theorized about such a flow or even explicitly referred to their framework as a “logic model” and I used them for inspiration (Garris et al., 2002; Kriz & Hense, 2006; Winn, 2009). For evaluating a game they speak of a sequence of input (or design), process (or play), and output (or experience).

Regarding input, it is important to know who the player is. His or her characteristics will likely make a difference on how the game is played and what results are derived from it. That is why I decided to use TGD as inspiration for determining how to look at players. Another input is of course the game. The design of the game will ultimately influence the outcomes and so understanding how a game is designed is helpful in appreciating what happens during a game-based training.

According to the evaluation principle “The proof of the pudding is in the eating” I had to consider the process of how the game is used. This requires to not treat the game as a “black-box.” Instead, it would require evaluating how players played and experienced the game. I call this stage “use,” to stress my focus on the game usage. With process one may consider many other aspects surrounding the training.

The game is used as an intervention and ultimately it needs to impact something. This something is what I call the “outcomes” of the training. Like with the player characteristics I used TGD for defining the outcomes of the game-based training. But the foundations of the evaluation approach was to consider the logic of input–use–outcomes.

### Toward a player-centered approach

In TGD, the player is one of the “context” aspects that designers need to take into account. In other approaches the “player” or “target group” takes a much more prominent place (e.g., de Freitas & Oliver, 2006). With evaluation this is also necessary and especially with educational games, such as *Levee Patroller*. The player is here the *unit of analysis*, the major entity being analyzed. Instead of the game which is the focal point of attention in design, in evaluation this becomes the player. It is possible to relate the player to TGD and as follows:

- *Reality–Player as person*: The player is a person in the real world. He or she has demographics, a personality, attitudes, and so on, that could affect how the game is experienced.
- *Meaning–Player as interpreter (or learner)*: People interpret information differently and so do players. This depends, for example, on the existing knowledge, education, learning styles, and expectations.

- *Play–Player as person*: Players differ among each other. This means that among other things previous experience with games as well as game preferences can make a difference in the results.

From each of these perspectives I identified the possible *contextual variables* that could play a role during the training with *Levee Patroller* (also based on Kriz & Hense, 2006). From Reality this concerns the type of job the player has, the organization he or she belongs to, and the commitment that person has toward the organization; from Meaning this is the education the person received, and the motivation the person has to learn about the subject and participate in the training; and from Play this is about the computer skills, game skills, and game attitude a person has. These were the contextual variables (among some others) I deemed most relevant for this particular game.

Regarding the player, Garris et al. (2002) argue that players experience a *game cycle*. Gameplay can lead to “certain user judgments or reactions such as increased interest, enjoyment, involvement, or confidence; these reactions lead to behaviors such as greater persistence or intensity of effort; and these behaviors result in system feedback on performance in the game context” (p. 445). So by becoming engaged with the game, players become intrinsically motivated to learn more, and (positive) feedback from the game will continue the cycle.

This part of their model shows that playing a game is a *dynamic process*. Players may adapt their judgments and behavior based on how the game progresses. Therefore, we should not regard the “player” as a fixed entity that we analyze based on contextual variables only. In addition, these player dynamics are another reason to consider the evaluation principle “the proof of the pudding is in the eating.” A game fails or succeeds because of what happens in between. If players’ game performance is poor and does not increase over time, it might be a reason for a player to discontinue the game.

For evaluating games this means that it is necessary to capture the player characteristics as well as track how player involvement and experience changes over time.

### **Toward an outcomes-based approach**

In the end, evaluating the effectiveness of a game is about measuring the outcomes (also referred to as output or effects) of its use. Similar to the player, we can discuss outcomes using the three worlds. Based on Play we can define the first outcome: *judgments*. These judgments refer to the participants’ satisfaction with regards to the game *and* training, because these two are hard to separate from each other. It is sometimes also referred to as *trainee reactions* (Sitzmann, Brown, Casper, Ely, & Zimmerman, 2008). From this world, the idea was to create a tool that is satisfactory and so if judgments are positive, this is an outcome that satisfies this world.

In the previous level I defined what outcomes we expect based on the world of Meaning. First, we expect the following cognitive learning outcomes: inspection concepts and vocabulary, mental model formation of failures, mental simulation of

failures, and the acquisition of an inspection protocol. For the sake of simplicity sake I will refer to these outcomes as knowledge.

Second, we expect skill-based outcomes, which are dependent on the cognitive ones, and I defined these as “sensemaking skills”: the ability to make sense of failures. To consider whether participants acquired these skills, we want them to improve their performance in making sense of failures.

Third, another impact the game may have is on the communication between patrollers, between patrollers and the Action Center, and between water authorities. I did not intend to measure this directly, because of its lesser importance in determining the effectiveness of the training. It could, for example, be measured by considering the knowledge increase in vocabulary. If this happens, we can assume communication will improve.

Then we have Reality. This world has not a specific outcome attached to it, but it is concerned about whether the outcomes will make an impact on practice. To consider whether the player experience is of value in reality, the concept of *transfer* is of importance. According to Sitzmann et al. (2010, p. 496) who cites Baldwin and Ford (1988) “transfer refers to the successful application of the skills gained in a training context to the job.” With levee inspection it is impossible to look into actual job performance, but a proxy—a variable that is used to measure an unobservable variable of interest—can be thought of. And this leads into a discussion of my concrete strategy to measure the effectiveness of *Levee Patroller*.

## ***The Evaluation Strategy***

In determining my evaluation strategy I first decided to consider the design as a given. I assumed that the game was designed to the best of its possibilities considering the available resources. It has furthermore been validated with experts and improved over the course of several years before it was finally used for the training. Just to be sure, I verified the game content and the exercises before the start of the training. This led to some small improvements, because I discovered inconsistencies that were not observed before—for example inconsistencies in the text, in the approved answers, and in the representations of the failures. This verification process took about one month and involved significant time and attention.

I considered the game-based training, as described in this level, as a given too. Although the training could have been organized differently, I designed this to the best of its possibilities as well. But I had no way of knowing whether it would work. I did not run a pilot study first, because it required significant effort to even be able to run a training session. I only ran a small test session to see if everything worked technically.

Due to the uncertainty surrounding the implementation of the training, I strove for pragmatism, opting for flexibility in the normal rigidity prescribed by (analytical) scientific standards. If something worked poorly, this would mean the training objectives would not be realized and, therefore, it would be better to improve this

the next time around. Although such changes threaten the internal validity of an evaluation, they give us a better understanding of what works and what does not in implementing a game-based training. Otherwise we need to set up a complete new study to achieve such insights and risk implementing (and much consciously so) a rather unsuccessful training. In addition, it was a training/evaluation. As a facilitator I felt obliged to improve the training.

This tension between rigor prescribed by scientific standards and the *relevance* of implementing a designed artifact successfully in its application domain relates to the reaction of Squire (2007) on the criticism by (Clark, 2007) on game-based evaluations (Level 1). The latter demands rigorous research, based on “reliable and valid tests” and by comparing the game to a “viable, robust non-game alternative” among others (p. 58), whereas the first says we need to deploy “iterative research, theory building, and design to generate useful theory” (p. 53).

This tension also relates to the distinction made by Klabbers (2006) between *analytical sciences* and the *design sciences*. The main aim of the analytical sciences is to develop generalized scientific concepts and context-independent knowledge. The design sciences, on the other hand, is issue-driven and aims to support and evaluate the development and use of a solution, such as a game, in its practical context. Localized, context-dependent knowledge plays a role here.

And as a matter of fact, the design scientific approaches are primarily concerned with making an impact in the real world by creating new and innovative artifacts and making these relevant to their application domain (Hevner, March, Park, & Ram, 2004; Hevner, 2007; Simon, 1969). Then the analytical sciences are characterized by rigor in order to explain or predict human or organizational behavior. According to this paradigm research needs to be theory-based and make use of reliable, robust, and valid methods and instruments.

A synthesis between the two paradigms has been illustrated by various gaming scholars (Bekebrede, 2010; Meijer, 2009; Klabbers, 2006; Kriz & Hense, 2006). It even *must* be found, because “practical utility alone does not define good design research...It is the synergy between relevance and rigor...that define good design science research” (Hevner, 2007, p. 91).

My evaluation strategy was to find such synergy. The design of *Levee Patroller* and the training with it were created to be of “practical utility”—to make an impact in the application domain of levee inspection. However, how I deployed this training and evaluated it are in accordance with the standards of scientific rigor from the analytical sciences as much as possible. My primary method of investigation concerns even a *quasi-experiment*, one of the hallmarks of this paradigm. I did not deploy this just because of approximating “good design science research,” it was also and especially an attempt at “producing findings that can be transformed and accumulated into generalized knowledge” (Kriz & Hense, 2006, p. 280)—much despite the fact that *Levee Patroller* is a very unique case.

Because of its uniqueness I realized it was important to describe its context carefully, to provide for a *thick description* of it (Geertz, 1973). This was necessary, because “Case-to-case transfer is enhanced by thick description that allows assessment of the applicability of study conclusions to one’s own situation” (Firestone,



1993, p. 18). I further realized such a thick description calls for a need to combine qualitative methods of inquiry with quantitative ones, because to describe the context carefully, qualitative methods of inquiry are necessary, whereas the likes of Clark (2007) remain critical unless quantitative evidence is provided.

The research with *Levee Patroller* is further marked by two needs that each belong to a different research methodology: the need to explain for the factors that contribute to the effectiveness of a game-based training and the need to explore what happens if we implement an innovative game-based training. The first need is instilled by the request of evidence and requires deductive thinking. This is known as *explanatory research* (Babbie, 1989). The second need is especially used when the topic or issue is new and requires flexibility, necessitating inductive thinking. This is known as *exploratory research*.

To close the gap between *a*) rigor and relevance, *b*) the analytical and design sciences, *c*) context-independent claims and a unique case, and *d*) explanatory and exploratory research purposes why I decided to ground my evaluation strategy within that of *mixed methods research* (Creswell & Clark, 2007).

## ***Research Design for a Thicker Description***

Mixed methods research is defined as “the class of research where the researcher mixes or combines quantitative and qualitative research techniques, methods, approaches, concepts or language into a single study” (Johnson & Onwuegbuzie, 2004, p. 17). According to some it could be considered the “third methodological movement” (Tashakkori & Teddlie, 2003, p. ix) after the “quantitative” (QUAN) and “qualitative” (QUAL) approaches.<sup>1</sup> It can further be characterized as inclusive, expansive, creative, pluralistic, rejecting dogmatism, eclectic, and as pragmatic (Creswell & Clark, 2007):

Mixed methods research is “practical” in the sense that the researcher is free to use all methods possible to address a research problem. It is also “practical” because individuals tend to solve problems using both numbers and words, they combine inductive and deductive thinking, and they...employ skills in observing people as well as recording behavior. It is natural, then, for individuals to employ mixed methods research as the preferred mode of understanding the world. When people talk about the Katrina devastation in the southern United States, both words and numbers come to mind. This type of talk is not only more natural, it is also more persuasive than either words or numbers by themselves in presenting a complete picture of the devastation (p. 10).

The purpose with *Levee Patroller* was to gain a “persuasive” and “complete picture” of game-based training and mixed methods research seemed the right choice to go ahead, not only to close the gaps between the apparent contradictions this

---

<sup>1</sup> The terms “quantitative” and “qualitative” are often used to denote a different research approach, but in either one of them researchers might make use of qualitative and/or quantitative methods. Also, I use “approach” as a synonym for *methodology*.

research was facing, but also because the premise is that the combination of a quantitative and qualitative approach provides for a better understanding than either approach alone. This is what is called the “fundamental principle of mixed research” (Johnson & Turner, 2003) and what I refer to as getting a “thicker description.”

### **An embedded concurrent mixed model design**

The evaluation takes the “design” part for granted and focuses on the “use” and “experience” parts of the triadic game evaluation framework. The experience part relates to the explanatory purpose of the research. This is to find evidence for the effectiveness of the game-based training and it belongs to the “quantitative” approach of this research. It belongs to the following two questions that were mentioned in Level 1:

1. What is the effectiveness of the training with *Levee Patroller*?
2. What factors contribute to its effectiveness?

The answer involves considering the three main outcomes—judgments, knowledge, and sensemaking performance—and the possible factors that are the cause of these effects, such as the player’s motivation and computer skills. This is about causes and effects and requires deductive thinking.

Although this quantitative approach will lead to possible new evidence and insights, it will not develop the understanding that is looked for. For example, why did it work with this particular game and this particular training? To make the results meaningful to others we need to open the black-box and look into its use in addition to a “thick description” of the training context. This requires another approach, because we have no idea up front what could possibly account for this or what would happen. A more exploratory, “qualitative” approach seems better suited here. This approach belongs to the other two questions as mentioned in Level 1:

1. How do participants experience the game-based training?
2. How do participants play the game?

Answering these questions requires making observations, looking into the game cycle and game data and then constructing grounded concepts that explain the data. This is about generating theory and requires inductive thinking.

Because the evaluation mixes QUAN and QUAL approaches and not just QUAN and QUAL methods, we should speak of a *mixed model design* and not of a *mixed methods design* (Tashakkori & Teddlie, 2003; Creswell & Clark, 2007). Because the two approaches are used simultaneously and not one after another, the design is *concurrent* rather than sequential. And because I collected and analyzed both QUAN and QUAL data (before, during, or after) within a traditional design, we speak of an *embedded design*. Therefore, we could consider this as an “an embedded concurrent

mixed model design.”<sup>2</sup> The traditional design that I used to embed everything in concerns the traditional quantitative design of the “quasi-experiment.”

### **Embedding the design within a quasi-experiment**

To explain the specific research design with its methods is to explain my logic of thinking about constructing it, because it is not based on a standard design found in a textbook. Rather, how it is constructed is based on the two different approaches as just discussed, the QUAN approach to find empirical evidence about the effectiveness and the QUAL approach to understand how participants experienced the training and played the game. Let us focus on the QUAN approach first.

From an explanatory research perspective (QUAN) I sought to make causal claims about the game-based training on the three main outcomes. In this regard the training concerns the independent or treatment variable and the three main outcomes are the dependent variables. To make valid claims about the treatment variable, (educational) scholars would recommend to compare one group that receives this treatment with at least one other group (Cook & Campbell, 1979). After an initial consideration of possible alternatives, it became clear that none exists, because the game fulfills a unique gap in the education of the patrollers. Creating a reasonable, new alternative seemed difficult too and posed the risk of becoming a “straw man” (see the principle of “no comparison of apples and oranges”). The only possible comparison was to make a *within-subjects* comparison and this becomes possible with a pre- and post-test. From this perspective the basic research design is a *one-group pre-test–post-test design*. This is considered a “pre-experimental design” (Cohen, Manion, Morrison, & Morrison, 2007), because no comparison is made. For this reason some consider this a bad example (Campbell & Stanley, 1963) or not even a quasi-experiment at all (Asher, 1976).

However, unlike most one-group pre-test–post-test design, the extent to which participants are exposed to the treatment will differ, because from the start it seemed that not every participant would play all exercises. Because of the differences in exposure, it becomes possible to make stronger claims about the impact of playing the game. The outcomes were determined with the following two main data collection methods:

*Pre- and post-questionnaire* Before and after the training participants made a self-assessment (using Likert items) of their knowledge and attitudes toward levee inspection. Based on these self-assessments the learning outcomes could be determined. The pre-questionnaire was further used to gather contextual variables, such as age and game attitude, and the post-questionnaire to determine how participants judged the training.

*Pre- and post-sensemaking test* To determine participants’ sensemaking skills, they needed to assess failure pictures before and after the training. To refrain from

---

<sup>2</sup> Various terms are mentioned in the literature to denote mixed methods research designs. I chose those terms that best reflect what research design I developed.

making any sense for them and to see an impact on communication (i.e., vocabulary, word count, and dispersion), participants needed to answer open questions. Content analysis was used to quantify the qualitative data. Just as self-assessment is a *proxy* of the actual cognitive learning outcomes, these sense-making tests are a proxy of sensemaking performance (and communication).

The explorative research perspective (QUAL) considered the treatment itself—to determine what happens in between input and outcomes. To achieve this I used the following two main methods repeatedly after each exercise:

*Game questionnaire* After every exercise participants had to answer a small questionnaire based on a number of closed items and open questions. This was used to understand how participants experienced a particular exercise and see how their experience with the game might change over time. The items were designed according to the three worlds of triadic game design and, therefore, I refer to each set of items as Reality, Meaning, and Play ratings. The answers on the open questions I consider “gameplay responses.”

*Game data* Each played exercise resulted in game data. This game data consists of quantitative data, the scores, and qualitative data of how the participant played an exercise (i.e., gameplay time and gameplay observations). With this data I was able to reconstruct how participants made sense of virtual failures. I made a distinction between the game scores as they were calculated by the game’s scoring system and the “failure correctness score,” which is a score of how a player dealt with a particular failure.

To triangulate (TRIANG) and explore some issues further, two additional methods were added to the research design and three additional small studies:

*Pre- and post-interviews* This is a clear example of an extension of the quantitative design. Before and after the training I selected a number of participants with the purpose to get to know who these patrollers really are and the organization they are affiliated with, test patrollers’ knowledge in alternative ways, and validate the sensemaking test. It was therefore a qualitative procedure to triangulate the outcomes of the quantitative approach.

*Discussion* At the end of the training I discussed with the participants what they thought of the game and the training. Unlike the interviews, this was particularly used to triangulate the outcomes from the qualitative approach. We discussed the game’s effectiveness, the game’s suitability for this target group, a possible increased awareness because of playing, and any suggestions for a future design and use.

*Students* Although I did not use a comparison group to determine the impact of the game-based training, I implemented a part of the training with students. I did this to *a*) assess how knowledgeable patrollers are compared to students at the start of the training (i.e., QUAN triangulation); and *b*) see how patrollers play the game compared to computer-skilled people (i.e., QUAL triangulation). I further used the students as a benchmark for a number of characteristics, such as game skills and attitude.

*Super experts* The game was used to turn patrollers into “professionals Hans Brinkers” and, consequently, it seemed advantageous to compare them to what I term the *super experts*, specialists in the field who I asked to complete the sensemaking test. This enables one to see how patrollers perform compared to these super experts.

*Field exercise* About a half year after the training I made a comparison, instigated by one of the participating organizations (Level 4). I observed how a group who received the game-based training perceived and acted on a field exercise compared to a group who did not. This study provided an opportunity to investigate the communication outcome further. Other opportunities were to consider an affective learning outcome, that of confidence, and the usefulness of the game-based training for performing a field exercise. I consider this a separate study, because it was not embedded in the quasi-experiment. Its insights are however useful as a validation of the training.

Using the logic of input–use–outcomes the above-mentioned methods and their measured variables are illustrated in Figure 3.1. This figure also highlights what I will discuss in the subsequent levels of this book. How the mentioned methods were embedded within the traditional format of the one-group pre-test-post-test design is more clearly visible in Figure 3.5, which is presented at the end of this level.

## Formulating Working Hypotheses

At the start of this level I formulated the major hypothesis of this research, which is basically stating that playing the game improved players on their knowledge and sensemaking performance. I used this initial hypothesis to design the evaluation. Having identified possible influential factors, such as game skills and age, and the methods to measure the results, I will now turn to formulating the working hypotheses. I speak of “working” hypotheses instead of hypotheses for several reasons. First, they “are suggested or supported in some measure by features or observed facts” (Whitney & Smith, 1901, p. 616). In addition to TGD, my expectations are based on my earlier observations with the game, assumptions mentioned in the literature, or facts mentioned in areas outside that of games. They are not based on strong, earlier evidence in game research.

Second, I was not intending to test my theory of how game-based training works, but rather to explore it, to achieve a better understanding and, as a matter of fact, engage in some theory building. I had some intuitive hunches and used these to guide my investigation. If these working hypotheses turn out to be incorrect, this would not be a problem at all.

Third, the working hypotheses give something to “work” with—they help in keeping a focus when delving into an empirical investigation and especially one with many variables. For answering the research questions more answers were needed than if players’ knowledge and sensemaking performance improved. The working hypotheses served as a guide in finding these answers. I need to add that they served

INPUT	USE	OUTCOMES
<b>Pre-questionnaire (Level 7)</b> <i>Play characteristics</i> Computer skills Game skills Game attitude <i>Meaning characteristics</i> Education Motivation Expectations Knowledge perceptions <i>Reality characteristics</i> Age Type Affiliation Commitment Inspection perceptions	<b>Facilitation (Level 4)</b> Facilitator notes (QUAL) Training facts Training errors (QUAL) Training improvements (QUAL)  <b>Game questionnaire (Level 5)</b> Play ratings Reality ratings Meaning ratings Gameplay responses (QUAL)  <b>Game data (Level 6)</b> <i>Game score</i> Total Per learning objective <i>Failure correctness score</i> Per failure Per learning objective <i>Other</i> Number of exercises Gameplay time Game observations (QUAL)	<b>Post-questionnaire (Level 7)</b> Game judgments Knowledge perceptions Inspection perceptions  <b>Post-sensemaking test (Level 8)</b> <i>Sensemaking performance</i> Total Per failure Per learning objective <i>Communication</i> Word count Vocabulary Dispersion
<b>Pre-sensemaking test (Level 8)</b> <i>Sensemaking performance</i> Total Per failure Per learning objective <i>Communication</i> Word count Vocabulary Dispersion	<b>Student game data (Level 10)</b> Facilitator notes (QUAL) Game score Game observations (QUAL)	<b>Discussion (Level 9)</b> Effectiveness (QUAL) Target group (QUAL) Awareness (QUAL) Suggestions (QUAL)  <b>Post-interviews (Various)</b> Personal (QUAL) Organization (QUAL) Knowledge & skills (Level 10)
<b>Pre-interviews (Various)</b> Personal (QUAL) Organization (QUAL) Knowledge & skills (Level 10)	<b>TRIANG</b>	<b>Super experts (Level 10)</b> Sensemaking performance
<b>Students (Level 10)</b> Characteristics Sensemaking performance		<b>Field exercise (Level 10)</b> Confidence Communication Usefulness training Observations (QUAL)

**Fig. 3.1** Research design of the training/evaluation, including the methods used and the variables measured. The methods in the grey area were used for triangulation

especially as a guide for the quantitative approach. With the qualitative approach I had just the questions as an initial guide to work with.

In formulating my working hypotheses I made a distinction into those that relate to the outcomes, those that are about variables that influence the outcomes, and those that involve a difference in outcomes among participants. I want to stress that many more hypotheses could have been thought of and that the ones I am about to describe do not cover all the variables I measured. These are simply the ones I deemed most important at the onset of my training/evaluation.

## ***The Main and Secondary Outcomes***

I defined two main learning outcomes of the training/evaluation: knowledge and sensemaking skills. The first, knowledge, will be measured via the questionnaires on the meetings and concerns a self-assessment. That is why I speak from here on of *knowledge perception*. It is the knowledge that participants perceive to have about levee inspection.

The second, the sensemaking skills, is what participants need to demonstrate on the sensemaking test. They received real and virtual pictures and need to make sense of these. The outcome of this is what I consider their *sensemaking performance*. This is the overall score of how *accurate* participants made sense of the pictures. This accuracy has been defined by the content of the game and input by experts, but especially emerged from an iterative coding process when the data was analyzed.

The game-based training is an intervention—an independent variable—that is supposed to especially affect these two dependent variables. Knowledge perception and sensemaking performance are therefore the two main outcomes and I expected that

**Hypothesis 1** *Post-training knowledge perception (Hypothesis 1.1) and sense-making performance (Hypothesis 1.2) will be higher compared to the pre-training knowledge perception and sensemaking performance, respectively.*

This first hypothesis is the “major hypothesis” I have been referring to. Then because of the evaluation principle “more than the tip of the iceberg” and the results by Sitzmann et al. (2010) which suggest that self-assessment is only moderately correlated to learning (Level 1) I further expected that

**Hypothesis 2** *Knowledge perception does not correlate strongly with sense-making performance; the correlation coefficient will be lower than .50.*

The cut-off of .50 is based on the conventional criteria by Cohen (1988) who implies that correlation coefficient values of .10, .30, and .50 represent small, medium, and large effect sizes respectively. I did not have such specific ideas about how well patrollers would perform on making sense of the pictures, and especially not how they would do before the training, but I did think it would be a fair assumption that they would perform better on the virtual pictures than the real ones. The virtual pictures I used came straight from the game-based training (except for one, the “new failure”; see Level 8) and the real ones they may have never seen. They most certainly did not practice with them. Although transfer would hopefully occur to the real pictures, it seemed that

**Hypothesis 3** *Post-training sensemaking performance on virtual pictures is higher compared to real pictures.*

A third and another outcome of the training/evaluation is what participants thought of it, something I referred to as *judgments*. Based on the idea of triadic

game design and the popular notion that learning is fun (Koster, 2005), I reasoned that this outcome relates to the two main outcomes. If players have fun and are engaged with the game, this would start a feedback loop as described by Garriss et al. (2002) that motivates players to learn more and so

**Hypothesis 4** *Game judgments correlate with knowledge perception (Hypothesis 4.1) and sensemaking performance (Hypothesis 4.2); participants who evaluate the game higher will have a higher knowledge perception and sensemaking performance.*

In Level 2 I explained that in the far end the game may have an impact on communication. This is one of the secondary outcomes and it is something I examined when studying the field exercise (Level 10). However, I also examined this with the content analysis based on the sensemaking test. I used three indicators to measure communication with this test. The first is *vocabulary*. With this I refer to what types of terms are used by the patrollers. One can hypothesize that if patrollers have a shared vocabulary, they are able to communicate better (Level 2). By playing the game patrollers will acquire or adapt to the game's vocabulary and so I assumed that

**Hypothesis 5** *Post-training vocabulary use will resemble the game's vocabulary closer compared to pre-training vocabulary use.*

The other two indicators are *word count* and *dispersion*. Word count concerns the number of words used to make sense of a phenomenon such as a risk. One could hypothesize that better trained practitioners not only are more accurate in what they describe (that is their sensemaking performance) but that they also need fewer words to do so. They will say, "This is a macro-instability" instead of "I see a levee that is severely damaged, because it seems like a part of it is settling toward the hinterland." If the other person understands this message, it will make the communication more efficient.

Dispersion is about the diversity of possible sensemakings. The game has a single and comprehensive classification system for reporting failures and if players adopt this we could expect seeing less variety in the responses they provide and thereby increase communication because less confusion will arise as a result. Therefore,

**Hypothesis 6** *Post-training word count (Hypothesis 6.1) and dispersion (Hypothesis 6.2) will be lower compared to the pre-training word count and dispersion, respectively.*

Throughout the training/evaluation I measured other secondary outcomes as well, such as the perceptions participants have about levee inspection (on the questionnaires) and a possible increased awareness they achieved by it (on the discussions). These are affective learning outcomes (Kraiger et al., 1993) and I did not make a rigorous attempt at measuring these, nor did I have any specific ideas about them.



## ***The Likely Moderators***

I hypothesized that a number of variables could affect the outcomes. In discussing these possible “intermediaries” between the independent and dependent variables it is useful to distinguish between *moderators* and mediators (Baron & Kenny, 1986):

...a moderator is a qualitative (e.g., sex, race, class) or quantitative (e.g., level of reward) variable that affects the direction and/or strength of the relation between an independent or predictor variable and a dependent or criterion variable...a given variable may be said to function as a mediator to the extent that it accounts for the relation between the predictor and the criterion...Whereas moderator variables specify when certain effects will hold, mediators speak to how or why such effects occur (pp. 1174–1176).

In other words, a moderator variable influences the direction and/or strength of a relationship between two other variables; a mediator variable explains it. If we would control for the mediator variable (by means of *partial correlations*), the relationship between the two other variables would disappear. With a moderator, the relationship will remain. Its direction and/or strength will only become different.

Because many variables play a role in a game-based training I expected that none of them mediate the two main outcomes of knowledge perception and sensemaking performance. But I did think that they would moderate them. The first and most important moderator is based on the evaluation principle “practice makes perfect”:

**Hypothesis 7** *The number of exercises played moderates the results on the main outcomes; participants who play more exercises will have a higher results on the main outcomes.*

In *Levee Patroller*, the learning objectives are directly coupled to the *game scores*. The game scores are composed of finding, observing, reporting, assessing, and diagnosing failures and taking measures against them (Level 2). This means we could argue that if someone is better in playing the game, that person has more knowledge and skills about levee inspection.

However, game performance is a matter of luck. Player cannot get points if they do not find a failure. In addition, the game’s scoring system is rigid. Those who happen to adapt to this system or who happen to think similar will receive higher scores than those who have other, plausible ideas about what they see. For this reason, I did not expect a large moderation ( $> .50$ ) to occur, but I still reasoned that

**Hypothesis 8** *Game scores moderate the results on the main outcomes; participants with higher game scores will have higher results on the main outcomes, respectively.*

Another reason why I thought that game performance may not reflect knowledge and skills is that players need to have considerable computer skills to play well. Those participants who have the necessary computer skills are at an advantage. They can pick up the game faster and get better scores, simply because they know how to work with it.

The same reasoning applies to game skills. It is often suggested that this requires some skill too and is referred to as *game literacy* (Gee, 2003) or *ludoliteracy* (Zagal,

2010). Participants with significant game experience, especially with First-Person-Shooter games, will have an additional advantage. They are quicker in understanding what is required of them and how to achieve high scores.

In fact, I believed that both may have even been so important that they could “mediate” the game scores. If that would happen and the game scores turn out to moderate the main outcomes, “mediated moderation” would result (Muller, Judd, & Yzerbyt, 2005). This happens when a moderator (the third variable, the game scores) is mediated by another (the fourth variable; computer and/or game skills). To be on the safe side I thought that at the very least

**Hypothesis 9** *Computer skills (Hypothesis 9.1) and game skills (Hypothesis 9.2) moderate the game scores; participants with higher computer and game skills will have higher game scores, respectively.*

I further expected that certain participants would appreciate the training more than others. I thought this was moderated by their *game attitude*, the extent to which participants have a favorable or unfavorable predisposition toward games or game-based training in particular. Certain participants may like to play games and others may not. Those who do not may likely not judge the game-based training as favorable and so that is why I thought that

**Hypothesis 10** *Game attitude moderates the game judgments; participants with a higher game attitude will have a higher game judgment.*

Computer and game skills and a game attitude belong to what I consider the world of Play. From the world of Meaning we are able to identify a number of possible moderators too. Two obvious ones are *motivation* and *expectations*. Motivation is the willingness and the extent to which participants strive to learn the material of the training program (Sitzmann, 2011). This motivation may be affected by playing the game (Garris et al., 2002); participants could become more motivated to learn due to playing the game—it triggers them. However, the initial motivation and subsequent motivations likely differ from one to another and motivation is assumed to be influential in learning processes.

Expectations are closely related to motivation. In fact, according to Vroom’s expectancy theory (1964), expectations are what motivate participants toward an action. If an individual’s perception is such that a certain type of action, such as playing *Levee Patroller*, would not lead to a specific and desired outcome—that is, enhanced levee inspection knowledge and skills—this individual becomes less motivated. Here we also notice the possible close entanglement of the worlds of Play and Meaning. If participants are aware they minimal little computer skills, it may inhibit their expectations. And so

**Hypothesis 11** *Motivation (Hypothesis 11.1) and expectations (Hypothesis 11.2) moderate the results on the main outcomes; participants with higher motivation and expectations will have higher results on the main outcomes, respectively.*

Similar to the outcomes other moderators could be perceived, such as participants' education and the commitment they have toward their organization. Although I measured some of these just to be sure, I did not expect them to play a major role.

### *The Likely Differences*

Whereas the (quantitative) moderators appeared from several characteristics that we could affiliate with the worlds of Play and Meaning, from the world of Reality we are able to identify a number of characteristics that could make a difference in the outcomes among subgroups of participants. This makes the characteristics affiliated with Reality "qualitative moderators."

One clear characteristic is the *type* of patroller. We are able to identify three types: volunteer, regular employee, and expert employee (Level 2). The first two types are not professionally preoccupied with levee failures; the third one is. This is why I assumed that

**Hypothesis 12** *Results on the main outcomes are similar between volunteers and regular employees (Hypothesis 12.1); and the expert employees will achieve higher results compared to volunteers and regular employees on knowledge perception and sensemaking performance (Hypothesis 12.2); but the learning gains of volunteers and regular employees are higher compared to the expert employees (Hypothesis 12.3).*

Before the implementation of the training I was aware that the average *age* among patrollers was relatively high.<sup>3</sup> They are not the target group that has grown up playing digital games. For this reason I expected that they would perform worse (moderated or even mediated by computer and/or game skills) in comparison to those who did. One of the reasons I included the study with students was to test these assumptions and I was not so sure to what extent the training sample population would include younger participants.<sup>4</sup> With this in mind I hypothesized that

**Hypothesis 13** *Students (Hypothesis 13.1) and younger participants (< 40 years; Hypothesis 13.2) achieve higher game scores compared to older participants (> 40 years).*

Another reason to include the study with students is to determine the initial level of patrollers' sensemaking performance. To see if the patrollers have become the sought for "professional Hans Brinkers" is why the study with super experts was included. Both studies would allow us to get a better idea about to what extent the game-based training achieved what is supposed to achieve. Beforehand I considered that

---

<sup>3</sup> The variable "age" could be considered a quantitative moderator too when not the age groups are considered but the participants' raw age numbers. I used both depending on my needs.

<sup>4</sup> The cut-off of 40 years was specified after retrieving participant ages (Level 7).

**Hypothesis 14** *Patrollers' pre-training sensemaking performance is higher compared to students' sensemaking performance (Hypothesis 14.1) and less compared to super experts' sensemaking performance (Hypothesis 14.2); patroller's post-training sensemaking performance will approximate the super experts' sensemaking performance (Hypothesis 14.3).*

The third and final study considered the field exercise. As explained this study involved comparing two groups, a Game Group who participated with the game-based training and a Control Group who did not. In this study I looked into possible affective learning outcomes as a result of the training, which is participants' *confidence* (or self-efficacy) in performing a levee inspection. With many hours of training presumably the Game Group should have more confidence.

I considered in this study another secondary outcome too: communication. As I explained elsewhere, I expected the game to have an impact on communication and with the field exercise the possibility opened up to measure this possible impact. Here too I conceived that the Game Group performs better than the Control Group. The final hypothesis is then that

**Hypothesis 15** *The Game Group has higher confidence before the field exercise (Hypothesis 15.1) and communicates better during the field exercise (Hypothesis 15.2) compared to a Control Group.*

Another clear characteristic from Reality are the *organizations* to which the participants belong. I planned to implement the game-based training at various organizations and many differences exist between them. Each has its own "community of practice" (Level 2). Belonging to a certain community of practice may have an influence on the outcomes too. For this reason, this needed consideration, but I could not hypothesize about this as long as I did not know with what organizations I was going to deal with. Based on my *facilitator notes* I could hypothesize about this and possibly explain for some of the differences between the sessions at the different water authorities.

## Setting Up a Futuristic Scenario

Having explained the evaluation, I will now turn to the training part of the training/evaluation. In the interest of ensuring proper application of the game and because of the lack of its actual use, I decided to design my own training. This requires a "big pond" and many hours of practice with the game ("practice makes perfect"). Those are two principles I immediately realized. Another that did not require much elaboration is that of "Ain't nothing like the real thing." I would of course implement this with the actual patrollers.

This "futuristic scenario" with *Levee Patroller* which I am about to describe may provide an innovative, novel foundation for similar training/evaluation ventures in the future.

## ***Creating a Big Pond***

In creating a big pond it is important to make sure the game remains a “big fish.” With this I mean that the game should remain the focal point of attention and not be relegated as “one of the instructional methods” in a program. Other instructional methods and support should be at the service of making the game more effective.

Creating the right kind of support is not the only concern in arranging a game-based training. My biggest concern was whether they would play at all. I had only seen that the game retained the interest of participants during workshops and there they played it for not even an hour or so. In those circumstances it acts as a nice diversion, a treat. What happens if they engage with it for over hours? Would they even be willing to participate? This I had no idea about. It was a gamble I was willing to take.

But it was not a gamble without any thought to it. I designed a game-based training based on some basic notions of *cognitive load theory* (Plass, Moreno, & Brünken, 2010) and some common sense ideas about the *willingness* of people to participate and their *commitment* in completing the training. Finally, as a game designer I thought of a way to integrate *instructional support* into the game-based training.

### **Notions from cognitive load theory**

Cognitive load theory distinguishes between *working memory* (or short-term memory) and *long-term memory* (Plass et al., 2010). A working memory consists of information that is temporarily stored and manipulated. Long-term memory is our storage of information. In other words:

Although schemas are stored in long-term memory, in order to construct them, information must be processed in working memory. Relevant sections of the information must be extracted and manipulated in working memory before being stored in schematic form in long-term memory...Working memory load may be affected either by the intrinsic nature of the material (intrinsic cognitive load), or alternatively, by the manner in which the material is presented, or the activities required of the students (extraneous cognitive load). Intrinsic cognitive load cannot be altered by instructional interventions because it is intrinsic to the material being dealt with, whereas extraneous cognitive load is unnecessary cognitive load and can be altered by instructional interventions...A further distinction should be made between extraneous cognitive load and germane cognitive load...extraneous cognitive load reflects the effort required to process poorly designed instruction, whereas germane cognitive load reflects the effort that contributes to the construction of schemas. Appropriate instructional designs decrease extraneous cognitive load but increase germane cognitive load (Sweller, van Merriënboer, & Paas, 1998, p. 259).

To develop an effective training, I would need to think of an “instructional intervention” to decrease the *extraneous cognitive load* of learning how to play the game and think of an “instructional intervention” of how to increase *germane cognitive load*. About the *intrinsic cognitive load* of learning about dealing with levee failures I could do not much. In addition, I would also need to think of how to prevent

*cognitive overload*, a situation “in which the learner’s intended cognitive processing exceeds the learner’s available cognitive capacity” (Mayer & Moreno, 2003, p. 43).

Based on cognitive load theory, I reasoned that

- A meeting or session was needed. With good guidance I would be able to reduce the extraneous cognitive load in learning how to play the game. From workshops and other settings I also realized that without guidance and instruction the target group would not be able to play. They are not experienced FPS players;
- To develop schemas or mental models of levee failures participants needed to practice with the game for many hours (“practice makes perfect”) and even a full day seemed short;
- To learn from the game it seemed better to spread out the practice. This would give participants some time to process the material (i.e., encoding information from working memory to long-term memory) and reflect on it (i.e., deeper encoding). It would also force them to be preoccupied with the material over a longer period of time (i.e., frequent retrieval from long-term memory to working memory). I basically hypothesized that by spreading out the practice, germane cognitive load would be increased, which leads to a better construction of schemas in long-term memory;
- This spreading would also be necessary, because playing the game for over two hours straight in a row seemed undesirable, unwanted, and likely to result in failure. The game is intense, especially to “non-gamers.” This means that participants become “empty” and cannot process the material anymore—they will experience cognitive overload. They may also lose their interest, because gameplay is somewhat repetitive.

My intention was not to proof these hypotheses about how to design an effective training with a game. I simply mention this to illustrate the assumptions on which the training was based.

### **Common sense ideas about willingness**

I was concerned about the willingness of patrollers to participate, because of the time and effort required and that I could not offer a large compensation. In fact, what I had to offer was a €12.50 *gift card*. This was basically a small gesture to thank the participants and, consequently, I hoped to take advantage of patrollers’ *intrinsic motivation* to participate (Malone, 1981). Many became a levee patroller voluntarily. I assumed that they wanted to get to know more about levee inspection and that they were excited to finally be able to practice with failures. Anecdotes from the crisis coordinators suggested some of the patrollers were looking forward to playing the game. This strengthened my belief that they have this motivation and that it might work.

Intrinsically motivated or not, under the following circumstances I doubted the willingness of patrollers to participate:

- Volunteers—the largest group of potential participants—would not sign up for a meeting during the day and either volunteers or employees would not sign up for a meeting during the weekends;
- Employees would be less willing to sign up for a training during the evenings, because they consider this work;
- Participants would be less inclined to sign up if the training would take four weeks or longer. A month sounds as a long time to get involved with a new training/evaluation. In addition, the likelihood that people drop out of the training increases too;
- Participants would also be less inclined to sign up (or continue with the training) if they cannot access the game easily, such as when they have to travel somewhere to play it.

The need for much practice, the need to spread the practice over time, and this latter need to increase the accessibility to play the game led to one possible solution: to allow participants to *play at home*. In itself, this is not an innovative idea, as many ICT enabled technologies have made it possible for people to train anytime and anywhere. To make sure participants would receive the necessary instructions to be able to play the game at home and understand its purpose, I decided it was necessary to organize a *start-meeting*. Based on my assumptions about participants' willingness these start-meetings had to be organized at workweek evenings.

From there I was able to decide about the length of the training. Four weeks seemed clearly not an option. Because participants are preferably preoccupied with the material as long as possible and are not “forced” to rush through the exercises at home, I decided on a length of *three weeks*. I admit that this is not rocket science, but considering the circumstances this seemed the most optimal length.

Allowing participants to play at home does have its disadvantages. Players cannot get any assistance and are not facilitated in reflecting on their experience. And as evaluator I have no control about the setting in which participants play the game. But most disadvantageous of all, it bears the danger that participants do not play or drop out. To prevent this as much as possible is why I considered participants' commitment.

### **Common sense ideas about commitment**

Even if participants signed up and are intrinsically motivated to learn more, I was concerned about their commitment. Meister (2002) reported that 70% of corporate learners do not complete online learning programs. Of course, it is theorized that games engage participants and thereby ensure for commitment (Garris et al., 2002), which makes them more likely to be successful, but a *Levee Patroller* or any other “serious game” is not a *World of Warcraft* or any other highly successful entertainment game. My assumption was that people are less inclined to play such serious games without any sort of commitment or incentives to continue playing.

The first solution I generated was to organize an *end-meeting* next to a start-meeting. By organizing this end-meeting participants were “forced” to finish the

home exercises by a certain date. It gave them a goal—a target—to work to. I was otherwise afraid they would not complete the exercises, because people have generally many other activities they could pursue and I was sure that playing *Levee Patroller* was not their first priority.

From an evaluator's perspective, this end-meeting seemed advantageous. By gathering the essential data during the meetings I was ensured to receive this data. I would not be dependent on participants' willingness and commitment to contribute to the evaluation. Also, it ensured that participants would play the game during the exact same time span, which makes it easier to compare the results between them.

But how would I know that people play between the start- and end-meeting? I did not want to become an obtrusive facilitator and question participants continuously about how they are doing. I also suspected, however, that without any incentives or "control" in between the meetings that participants would forget about it or wait with playing until the last minute.

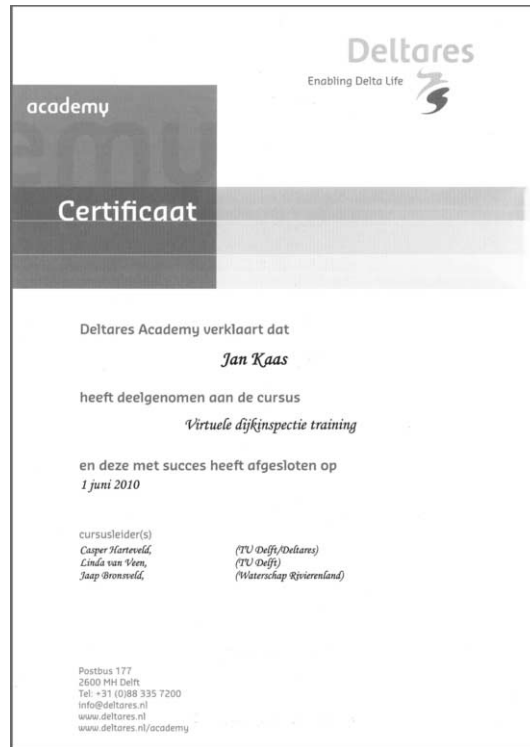
Playing until the last minute was problematic for three reasons. First, participants had to play a sufficient number of exercises ("practice makes perfect") and I hypothesized that they would not be able to play all of them just before the end-meeting. Second, *Levee Patroller* is a complex game due to its reporting system. If participants would not become familiar with its gameplay after the start-meeting, when everything was still fresh, I was afraid that participants would have much trouble in playing it. Third, the insights from cognitive load theory suggest that participants had to play the exercises spread out over the time in between the start- and end-meeting. This would allow them to process the material much better.

This led to my second solution, which is actually based on game design. This concerned the use of *weekly assignments*. Each weekly assignment required participants to play a number of exercises. At the end of each week I then submitted an e-mail to the participants with a *code* to unlock the next assignment. In this way, participants received an incentive to play the game in a structured manner and spread out over the training. They could still accumulate the e-mails and then start playing just before the end-meeting (and some of them actually did this). I further submitted codes sooner to those participants who informed me that they were not able to play during a certain week. I requested them to keep the code a secret to others, because the codes were the same for every participant.

The weekly e-mail further served as a non-obtrusive way for me to maintain contact with the participants and provided a courteous reminder that they were part of a training and that they were expected to complete the assignment of that week. This worked as we can tell by the reaction of Participant #103. He was not the only one in apologizing to me that he did not finish the assignment in time.

A final incentive is that participants received an official certificate for participating (Fig. 3.2). Similar to the gift card this is a small gesture, but it served to illustrate that they did not participate in a research study only. It was an actual training.





**Fig. 3.2** The training certificate of the training, which reads “Deltares Academy hereby declares that John Cheese participated in the course Virtual Levee Inspection Training and successfully concluded this on June 1, 2010.” At the bottom, the course facilitators are mentioned. The name in this example is fictional

### Integrating instructional support playfully

Game designers think alike I thought to myself when reading McGonigal’s (2011) description of *The Lost Ring*, “an online game that would give young adults around the world an opportunity to collaborate at a scale as awe-inspiring as the modern Olympic Games themselves” (p. 281). In this game, played just before the 2008 Summer Olympics in Beijing, players had to reconstruct the “Lost Sport of Olympia,” a blindfold game that the ancient Greeks supposedly banned from the Olympics and who attempted to destroy all evidence of its existence. The designers used two strategies:

...we used the strategy of massively distributing game content in different languages, on localized Web communities, and across far-flung real-world geographic locations in order to make it impossible for any single country, let alone a single player, to experience the game alone...Key online game clues were hidden on regional websites and social networks...physical game objects were hidden in virtually every corner of the world. None of these clues or objects was redundant; each added an important piece of information to the

history of the lost sport...we also adopted the strategy of telling what gamers call a “chaotic story.” Instead of presenting players with a single means of consuming the game story, we broke it into thousands of pieces like a jigsaw puzzle and then diffused it across many different media platforms...This kind of chaotic storytelling forces players to actively make sense of the game content for themselves and for each other (pp. 286–287).

The reason I thought that game designers think alike is that in a way this is how I integrated instructional support into the training. To learn the material well, patrollers were forced “to actively makes sense of” levee failures. They do this in the game, but the game does not give the exact answers. Although it gives feedback about what they need to improve, such as a low score on reporting accuracy, it does not highlight what is wrong exactly.

Participants could already access the handbook from their inventory to find out on how to improve their performance (Level 2). In addition, they could go to a simple *website* that I created for the purpose of this training. This website contains various tips and suggestions about what to pay attention to and what heuristics should be kept in mind, such as when we should speak of a “little” or “large” amount of water that overtops the levee.

The information in the handbook is different from the website. In this training “each added an important piece of information to the” sensemaking of failures. I did this, because I thought that this way of actively putting the pieces of information together would help in constructing mental models of failures. To see if participants made use of the website (and for how long), they had to enter their name before they could access it.

I also added a “fun” section to the website, which participants could access with a code. In this section they could learn how to

- Become Superman and fly around the virtual environment;
- Become Flash, another superhero who has the ability to run and move extremely fast, and make it possible to adjust the speed of the game; and
- Turn into a sheep and walk around eating grass.

I made it impossible for participants to access this section during the training, because some of them are “cheat codes,” which enable players to perform better or finish an exercise more quickly. Enabling them would cause problems in comparing the results.

Besides the website, I created a *manual*, which includes information about frequent errors that can occur while installing the game and information on how to play the game. It also has information about the training procedures such as

Playing one exercise takes a maximum of 24 minutes, but this excludes the time you stay inside the menus. The game is then paused. Therefore, you are likely to spend about 30–40 minutes with a single exercise. Make sure you can play the exercises without too much disturbance during this time. Of course, you can always pause the game to for example grab a cup of coffee, but it is important to not have too much distraction during an exercise.

I did not include the website information to this manual deliberately. This decision relates to the previously mentioned idea of forcing patrollers to actively make

sense of risks. In addition, I could track the website usage and not the use of the manual.

## ***Adapting the Big Fish***

Subsequently, I decided to create a three-week training with *a)* a start- and end-meeting on a workweek evening; *b)* weekly assignments with a number of exercises to complete; and *c)* with a website and a manual as instructional support. What has not become clear is what the participants had to play—what the “big fish” was. This is what I aim to clarify right now. I created a special version of the game which was called the *research version* and carefully selected what to include and what not. Then I carefully constructed a *training program* consisting of game exercises.

### **The research version**

It seemed prudent to allow participants to play at home, but the game *Levee Patroller* was not ready to be played at home. Although the game includes scores, it does not give any further feedback about the performance. Because I was aware that feedback and reflection are important in learning from a game and that no facilitator would be there to discuss the experience, I considered it was absolutely vital to include a feedback system in the game before it was ready to be used in the training.

The feedback system is a screen that players can access after they finished an exercise. They can see where the failures were located in the region, how they look like, and receive information about how they dealt with them. What this feedback system allows for is *reflection-on-action*, which is a deliberation after the action occurred (Schön, 1983).

The research version further differs from the original one in that game options were excluded, such as the scenario generator and full game option. The training exercise option remained. I figured that this option would be useful to participants who experienced trouble in playing the game. They could always access this (and I understood that some participants made much use of this).

In constructing the research version I kept experimental rigor in mind and then in particular the *internal validity*. Internal validity “is concerned with correctly concluding that an independent variable is, in fact, responsible for variation in the dependent variable” (Millsap & Maydeu-Olivares, 2009, p. 25). The independent variable here is playing the game. Now if I would have allowed participants to play whenever and whatever they like, I could have measured what they played and then relate this to the outcomes on the training. However, what they played and in what order affects the outcomes. It affects the outcomes on the exercises and it affects the training outcomes. Some of this can be controlled for, but internal validity is most certainly threatened. It becomes more difficult to compare the outcomes between participants. For this reason, it had to be ensured that players play the exact same

exercises in the exact same order and are unable to play each exercise more than once.

This required another adaptation from the original version, because the original one places failures randomly in a region. Each time an exercise is initiated, it places a failure somewhere in the region and every time at another location. We included this to increase the “replayability” or “replay value” of the game. If this would happen in the research version, the randomization of failures becomes a *confounding variable* (Field & Hole, 2003), because where failures are placed might affect the performance of players. It could for example happen that two failures are placed next to each other. That makes finding and reporting those failures much easier (or harder if the player never gets to that area) compared to a situation where both are spread out in the region.

To ensure that participants played in the exact same order and are unable to play each exercise more than once, the research version made sure that participants could only access one exercise at a time and after finishing that exercise they could not access this anymore.<sup>5</sup> I did not disable the possibility to restart an exercise that was not finished, because it might happen that a laptop runs out of battery or a participant suddenly has to leave his or her computer.

Due to this “escape option,”<sup>6</sup> participants could cheat by ending an exercise prematurely and starting it again. I did not suggest this (of course), but I did suggest to restart an exercise only under special circumstances. I noticed that in the end two or three happened to cheat like this (and they honestly admitted this—they were not satisfied with their scores and used this strategy to improve them). But as with allowing participants to endlessly repeat the training exercise, pragmatism ruled here over experimental rigor.

The research version eventually worked like this (Fig. 3.3). After installing the research version and starting it, participants had to first fill out their name. This had to do with the data collection. Without a name I could not retrieve who played what and because of this crucial importance, participants could not access any content before they filled out their name.

Then the first exercise appeared. After they finished this, the next exercise became unlocked, but only if it still belonged to that week’s assignment. If not, first a code was needed to unlock the assignment for a specific week. The codes were simple words that relate to the subject matter of the game, such as *dijk* (levee), *gras* (grass), and *talud* (slope).

At the start-meeting participants were clearly informed about this procedure to prevent them from getting frustrated. Also, I informed them that at the end of the training they would receive a code that would unlock everything. With this code they would be able to play everything as much as they want.

<sup>5</sup> A simple INI file kept track of the progress of players and configured the game accordingly when started. If problems occurred I asked participants to change this file.

<sup>6</sup> This is a pun, because participants had to actually press the ESCAPE button to leave the exercise.



(a) Entering name



(b) Entering code

**Fig. 3.3** Entering your name and the weekly code with the experimental game version. *Oefening* means exercise and *Bijeenkomsten* are meetings. Notice how Exercise 1 is unlocked and Exercise 2 is still locked

### The training program

What I have not clarified so far is what the participants actually had to play. This is what I consider the “training program.” My first decision was to consider how many exercises they had to play and when. Because I expected that some participants would have trouble in playing the game, I decided that during the start-meeting participants would play the training exercise and the first exercise. This had the additional benefit that I could observe how they played. I was not able to do this when they played at home. For the same reason I included an exercise at the end-

meeting. This allowed me to see the contrast with the start-meeting with my own eyes.

Three weeks passed between start- and end-meeting. I could not demand too much from the participants, but I also needed to ensure that they exercised enough. I decided that two exercises per week seemed reasonable, for a total of eight. This amount includes the start- and end-exercise which they played at the start- and end-meeting, respectively.

Regarding the content of the exercises my highest priority was to create an interesting and effective program. Experimental rigor came second here. For creating an interesting and effective program I considered the quality and variety of the game content. Consistent with how games work, how people learn, and *flow theory* (Csikszentmihalyi, 1991), I ensured that exercises increased in difficulty. The variables I was able to consider are

- *Responsibilities*: Players could have basic responsibilities (i.e., observing, reporting, and assessing) or could get extended responsibilities (i.e., diagnosing and/or taking measures);
- *Weather*: It could rain or not;
- *Failures*: Failures in exercises could vary in type, number, and severity (i.e., reportable, severe, and critical); and
- *Regions*: Different regions were available with each their specific characteristics.

I included full responsibilities, because this makes the game much more interesting, from a gameplay perspective as well as from an inspection perspective. It is simply much fun for players to take measures themselves, even if it requires a simple action. It will further give patrollers an overview of the complete inspection process and this is valuable, even if they do not have this responsibilities in reality.

If it rains, players have less visibility and it becomes arguably more difficult to play (because of an increase in “cognitive load”). I, however, suspected that this would not have a great effect on the difficulty and so I included it to create for variety among the exercises, make the experience more intense, and reflect reality better.<sup>7</sup> If patrollers have to go out they will most likely do this with rain.

The failures required more thought. In my opinion some failure types were not as thoroughly implemented and I neglected these. I decided to focus especially on four types that have been most elaborated upon: the stone damage, small landslide, watery slope, and boiling ditch failure (for details about these failure types, see Level 6). Because the small landslide is the most complex failure, due to its many signals, I included this more than others, to allow players to have more practice with it. To increase the variety I added a few times two failure types that need some refinement: the grass damage and illegal driveway failure.

Then I intuitively estimated that an exercise should have at least two failures and a maximum of five. With one failure two scenarios are possible: players do not find

---

<sup>7</sup> The amount of rain is randomly generated in the original version. Sometimes it pours rain and at other times it just drips a bit. For the research version the amount of rain was always set at medium. It does not pour and it does not drip.

the failure at all and get extremely frustrated or they will find it and see no reason to leave the sight of the failure, because the number of failures in an exercise are known to the player.

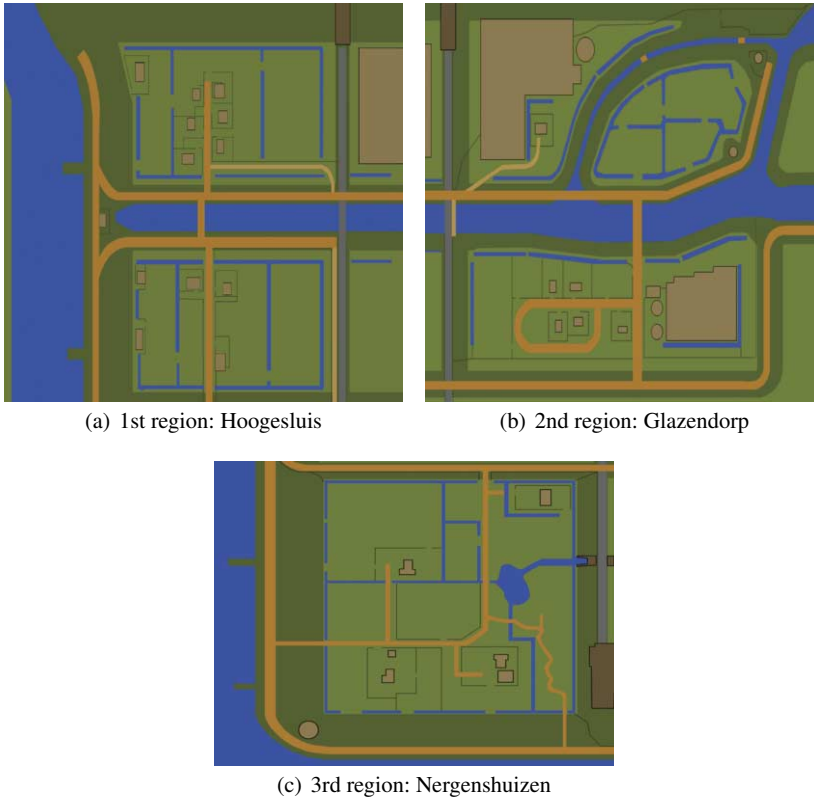
**Table 3.1** Contents of the training program

Variables	Start-exercise	Home exercises						End-exercise
		1	2	3	4	5	6	
Region	1	2	1	2	1	2	3	3
Rain	No	Yes	Yes	No	No	Yes	No	Yes
Failures								
Number	2	2	3	3	4	4	5	5
Severity								
Reportable	1	1	1	0	0	1	0	0
Severe	1	1	1	1	2	1	3	1
Critical	0	0	1	2	2	2	2	4
Type								
Stone damage	1	0	1	0	2	0	1	1
Small landslide	0	1	1	0	1	1	2	2
Water slope	0	1	0	2	0	2	0	0
Boiling ditch	1	0	1	0	1	0	1	1
Grass damage	0	0	0	1	0	1	0	0
Illegal driveway	0	0	0	0	0	0	1	1

*Note.* Region 1 = Hoogesluis; Region 2 = Glazendorp; Region 3 = Nergenshuizen.

With more than five failures I reasoned that players become more busy with finding failures than actually reporting those that they found. It seemed better to configure the difficulty in failure severity. With this in mind, in determining the exact sequence I made every exercise a bit more difficult than the previous by including an extra failure and/or increasing the severity of failures. The final training program of the research version starts with exercises with two failures with small amounts of severity and ends with five failures with large amounts of severity (Table 3.1).

Because certain failures are tied to specific regions I had to include at least two regions: the regions Hoogesluis and Glazendorp (Fig. 3.4). For the first six exercises I exchanged these to ensure for some variety in failures and virtual environment. For the final two exercises I opted for another region, that of Region Nergenshuizen, and I did this to increase variety but especially to “end with a bang.” This region involves a big river and has incredibly large levees. This would lead to a spectacular grand finale for players where they had to show they became “professional Hans Brinkers” that prevent the land from flooding.



**Fig. 3.4** The three regions that were used for the experimental version. The region names translate to “High Sluice,” “Glass Village,” and “Nowhere Homes,” respectively

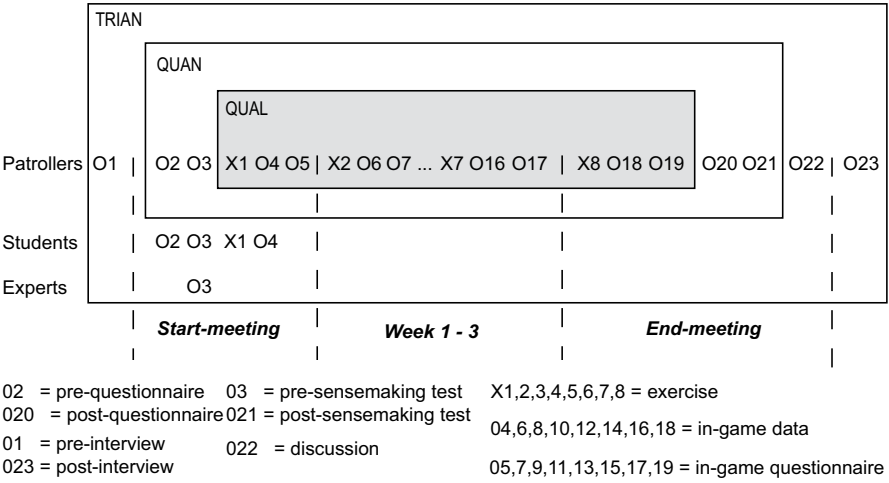
## Lessons Learned

In this level I described the design of the training/evaluation with *Levee Patroller*. The starting point of the design was to proof and understand if and how, respectively, the game improved players in terms of inspection knowledge and sensemaking skills.

To evaluate this, two approaches were used: a quantitative and qualitative approach. The quantitative approach aims to provide for evidence about the effectiveness of the training and the factors that contribute to it. The qualitative approach, on the other hand, aims to provide an understanding of game-based training by considering how participants experienced the training and played the game. This qualitative approach is embedded into a traditional quantitative design, that of a quasi-experimental design.

Various methods were used in support of the quantitative and qualitative approach, but also to triangulate their results. The methods aim to measure the three





**Fig. 3.5** The integrated research design of the training/evaluation, excluding the field exercise study

main outcomes of knowledge perception, sensemaking performance, and game judgments as well as possible secondary ones, such as communication and confidence. They further measure possible moderating and mediating variables. A set of 15 working hypotheses were formulated to describe my initial hunches about these outcomes and likely influential variables and to guide the analysis of the empirical results. I will return to these hypotheses when putting all the puzzle pieces together in Level 11.

Based on the evaluation principles and strategy, cognitive load theory, and some common sense ideas about willingness and commitment, a structured three-week training was developed with *a)* a start- and end-meeting on a workweek evening; *b)* a special research version that includes eight exercises, three regions, full responsibilities, and an increasing difficulty; *c)* weekly assignments with two exercises to complete; and *d)* a website and a manual as instructional support. Unlike presented here in this level, the development of the training/evaluation occurred iteratively, by considering the possibilities and limitations of the game and its possible implementation and by considering the necessities of the evaluation. A number of trade-offs were made, due to the inherent tensions between designing a training and evaluation, such as by making the training more interesting by including various failures and regions.

Figure 3.5 provides an integrated overview of the training setup and the evaluation strategy and methods. It highlights what observations were made (denoted with “O”) and what interventions were applied (denoted with “X”).

## Level 4

### Playing a Futuristic Scenario

*iffreshsoiliscomingwiththewaternoidea—Answer by participant who does not know how to use the SPACE BAR*

*If a disaster occurs, I am ready. But all that complicated talk and stuff, that is too difficult for me and I do not feel like getting involved with that—A patroller who did not want to participate*



In January 2010 I decided to implement the training/evaluation with *Levee Patroller*. I contacted the five water authorities that participated in the development of the game and asked if they were interested in participating with what I called *The Virtual Levee Inspection Training*.

I told them that I would bring my own equipment, because I did not want to risk any issues during the training. For a similar reason I made sure everything worked without Internet. I also explained that I would bring one additional facilitator to assist me.<sup>1</sup> I further informed them about this initial procedure of the start-meeting:

1. *Introduction*: I welcome the participants and I briefly explain the purpose of the training.
2. *Pre-questionnaire and test*: The participants start with filling out the pre-questionnaire and the pre-sensemaking test. I request them to do this independently and to wait outside the training room when they are done. Outside the training room I will make sure food and drinks are available.
3. *Training exercise*: When everybody is ready, we start playing the game. First participants have to play the training exercise. In this exercise, all aspects of the game—from navigating to reporting—are discussed step by step.

<sup>1</sup> The additional facilitators were members of the gaming team at Delft University of Technology and Deltares.

4. *Start-exercise*: After the training exercise, participants could start with their first exercise. In this way they play the first exercise with guidance of me and another facilitator.
5. *Closing*: At the end I explain the procedure of the training. Before leaving every participant receives a map that includes a CD with a special version of the game and a manual.

And I informed them about the procedure regarding the end-meeting:

1. *End-exercise*: At the start of the end-meeting, participants play the end-exercise. Participants who are done could play another game or wait until everybody is finished.
2. *Post-questionnaire and test*: After the end-exercise participants have to fill out the post-questionnaire and test. Like with the start-meeting I request participants to wait outside when they are done.
3. *Discussion*: Depending on the time and the willingness of participants a structured discussion takes place about the training and the game.
4. *Closing*: I thank everyone for participating and hand them a training certificate.

Out of the five water authorities, two declared they were too busy in between March and June and could not participate. The other three assented.

The goals of this level are to describe

- The three participating organizations;
- How I organized the training at each one of them; and
- How the training proceeded.

## The Participating Organizations

My objective here is to provide a “thick description” of each participating organization. This description gives the context of the training and enables to understand some of the results that I will describe in later levels. It is important to emphasize that the description comprises more than some general notions about each organization. It also encompasses the implementation of the training. The training setup differed to start with. With this I refer to how the training was organized: who did what, how participants were recruited, and when and how the training was given.

In addition, my experiences with the training were strikingly different. This may have been partly due to how the training was organized, but it could be attributed to other factors as well. With this in mind, the description provides the context and an impression of the training. Table 4.1 gives an overview of some basic characteristics of the three organizations and how the training was organized.

The three water authorities are heretofore called Organization A, B, and C. They are classified thusly to guard sensitive information. The trainings were sequentially given, first at Organization A, then at Organization B, and finally at Organization C. In this sequence I will also describe the organizations.

**Table 4.1** Overview of some basic characteristics of the three water authorities and how the training was organized

Fact	Organization		
	A	B	C
Area in 10,000 hectares	41	102	201
Area description	Industrial	Mixed	Rural
Prominent levee types	Regional	Mixed	Primary
Levee army personnel	178	650	300
Percentage volunteers	75	70	50
Training time	March	April	May
Training location	Deltares	Guard posts	Headquarters
Participation premise	Voluntary	Voluntary	Compulsory
# of groups	4	5	2
# of initial participants	43	82	35
Training administration	Researcher and organization	Organization	Researcher
Training compensation	Travel costs and €12,50 gift card	€25 per meeting	€12,50 gift card and €25 per meeting
Training assistance	One intern	One intern, two junior and two senior staff	Two senior staff

### ***Organization A: Innovative, Industrial and Inactive***

The first organization that agreed to participate was established in the Middle Ages, after a merger of several smaller water authorities. Nowadays it serves one of the most densely populated and industrialized regions of the Netherlands. Its 41,000 hectares include about 1.4 million inhabitants and 40,000 companies.

The organization has the responsibility over some river and some sea levees. Threats for flooding in its area come for example from the North Sea and a constructed canal to keep the city of Rotterdam and in particular its port accessible to seafaring vessels. However, the region has especially many smaller waterways and so it has to deal above all with regional levees.

In the early 1990s, large regions were flooded in Organization A's area. To prevent this from reoccurring, the organization invested heavily in all kinds of measures, such as extra water basins where excess water could be temporarily held. Consequently, the organization encountered financial trouble, which led them to prioritize certain tasks and responsibilities over others.

Inspecting levees was not a high priority, so it happened that the voluntary levee patrollers did not receive any training for the past three years. I understood they were even thinking about abolishing the use of volunteers.

Contrary to the volunteers, the employees were still yearly trained. In fact, the organization tried to professionalize these trainings and made more use of *Levee Patroller* than any of the other organizations. This is the organization with the “game room” and who organized their own game-based trainings (Level 2).

And contrary to the other two organizations, Organization A invested in the development of the game beyond the yearly payments to improve it. They wanted a special level devoted to “dry failures,” failures caused by drought. The other two organizations were not interested in this expansion as their areas are less prone to such failures. Eventually Organization A trained their employees with this new level while I was training Organization C.

Organization A seems to have two faces with respect to levee inspection. On the one hand, they invest in it and on the other hand they downsize it to such an extent that it is almost non-existent. Their strategy in professionalizing the levee inspection seems to be in investing in employees and (ICT) innovations. In early 2010 they entered into a transition phase and did not know what to do with the volunteers. The only thing they did was sending a mail every year at the beginning of February to ask the volunteers if Organization A could still make an appeal to them as needed. Not surprisingly, every year responses declined and the number of negative responses increased.

With this in mind, my invitation to organize a training came almost as a “gift from heaven.” Organization A could then send their yearly request and this time mention they were organizing something very soon. In fact, the organization told me literally they wanted to make use of this opportunity “to blow new life into their contacts with the volunteers.”

### **Sending out the invitations**

The plan of sending the Organization’s request for volunteer availability together with the training invitation is what happened. First, the availability request was sent with a mentioning of the training opportunity. A week later, the training invitation was sent. Organization A handled the delivery of the mail; I made sure I created an invitation (Fig. 4.1). The invitation was a small folder with four pages. These pages explained what the training was about, who could participate, what was expected from participants, and what they got in return for it. Important to stress is that participation was entirely voluntary.<sup>2</sup>

Organization A did not have much funding for this training and so apart from sending the mail, they could only compensate the travel costs of the participants. An intern attended the meetings to represent the organization. He was also my contact person and helped me with administering the training. All the other aspects of the training were arranged by me, including the location, catering, security, and of course the training itself.

The invitation was sent to 128 volunteers, among whom only 12 people directly registered by sending me an e-mail. Due to this low response rate, we decided that I would call every volunteer who already indicated he or she would be available again. This turned out to be a good “call.” Some confusion resulted among the volunteers,

---

<sup>2</sup> The invitation also clearly mentioned that the training was meant for research and that participation would entail a consent to this research. I repeated this at the start-meeting and stressed that everything would be reported anonymously.

### Where and when?

Unfortunately, all planned training sessions are over. Possibly in the near future, maybe at your local water authority, a new training will be organized. This was the general outline of the activities:

Month	Activity
1	Startmeeting at 19:30
1 till 8	Play 2 exercises at home
8 till 15	Play 2 exercises at home
15 till 22	Play 2 exercises at home
22	Endmeeting at 19:30

Except for the participants of Organization B, in the original invitations people could choose the dates they wanted to attend. You do not need to do this of course.

### Do you want to participate?

Too bad if you want too, because as indicated above, no new training is planned. If you are interested, you can send an e-mail to Casper Hartevelde ([c.hartevelde@tudelft.nl](mailto:c.hartevelde@tudelft.nl)). Maybe he keeps you posted about any possibilities.

Here in the original invitations - except for again Organization B - participants were asked to confirm their participation by sending an e-mail.

# Invitation



## For a virtual levee inspection training!


Delft University of Technology




Delft University of Technology



Challenge the future

Challenge the future

**Fig. 4.1** The front (right) and back (left) side of the training invitation folder. On the inside information was provided about the invitation, the game, the research surrounding the training, who is able to participate, what is expected from participants, and what they get in return for participating

many of whom thought that they were automatically enrolled into the training by telling Organization A they were available again.

A week before the training—this is, three weeks after the request—52 volunteers told Organization A they were available again. Among this total, I was unable to contact two of them, despite of several attempts. Two other volunteers initially agreed to participate but later reneged. One of them told me why over the phone. He was 64 years old and said he “could barely turn a computer on and off.” Four people said no right away. These were each of their reasons:

1. “If a disaster occurs, I am ready. But all that complicated talk and stuff, that is too difficult for me and I do not feel like getting involved with that.”
2. “I have nothing with computers.”
3. “I have no time. I am too busy with work.”
4. “I am too old for this. This requires me to learn again and I do not feel like doing that. I further have no attachment at all with computer games...they are too tedious...and I am also quite busy with my business...if I need to walk [for inspecting levees] then I will be there, but not for something like this.”

Four individuals stated they could not attend any of the meetings but wanted to participate or make an attempt. One of these volunteers had a specific reason: He played the game before and experienced something called *cybersickness*, a form of

motion sickness caused by moving something on a screen while not moving yourself. It gives a nauseous feeling. For this reason, the volunteer rather played the game in a comfortable environment.

Another “stay home” volunteer expressed his doubts about the usefulness of this game. He thought it was “a waste of his time” and “childish.” He also thought it was unnecessary, because “if something is wrong then you will see it.” Eventually these volunteers did not really participate at all. All of them played the first exercise and not more than that. The volunteer with major doubts was so kind to call me and he told me this:

I gave it a try, but I could not find what I was looking for...I give myself an insufficient grade. I wish everybody good luck with this. It is not meant for me.

This left me with 33 volunteers, of whom two indicated up front they could only attend one of the meetings. Considering the short amount of time in which the training was arranged, this was not a bad response rate. I cannot say this about the employees, who were arranged internally. The intern arranged a contact person within the organization and he or she sent an e-mail to all eligible employees. Although employees had to attend the meetings outside the regular working hours, they were allowed to write their hours for the complete training. Despite this arrangement, my hunch was, and this has been confirmed by others, that employees do not feel like doing anything “voluntary” outside the regular working hours. For this reason, the trainings by the Organization were compulsory and took place during regular working hours. In retrospect it may have been a better idea to organize at least one session during the day.

Another reason for the lack of success in employee recruitment was that the contact person went on vacation after sending the e-mail without giving notice to the intern. No follow-up occurred, until shortly before the training, the intern personally asked a number of employees. In the end, six agreed to participate, making the number of participants from Organization A a total of 39 initial participants.

### **Running out of time due to computer illiteracy**

The training with Organization A took place at Deltares, the institute where *Levee Patroller* was developed (Fig. 4.2(a)). I planned only 1.5 hours for each meeting, with two meetings per evening, but this proved to be an inadequate time allotment due to the demands of the program and various delays. I especially underestimated the amount of time it took participants to fill out the pre-questionnaire.

Despite this experience, we decided to not change the end-meetings. I expected fewer problems with time there. I could stretch or shorten the discussion as needed. The expectation about the end-meeting turned out to be true, although 1.5 hours was still too short.

It is worth noting that the low level of computer literacy came as a major surprise. Consider this telling example, a conversation between the assistant-facilitator and one of the participants:

Assistant: Excuse me. I see you do not make use of the SPACE bar.

Participant: The what?!

Assistant: The SPACE bar. You can use this to separate words. I see you do not use it.

Participant: Aha, I was already looking for something like that. Where is it?

Assistant: It is here. *Assistant points with his finger to the SPACE bar*

Participant: Thank you so much. I learned something again tonight!

Later—curious as I am—I looked up what this participant was writing:

iffreshsoiliscomingwiththewaternoidea

A few others also did not make use of the SPACE bar or had other issues (many typed with only one finger). However, participants were enthusiastic and eager to learn. The above-mentioned conversation illustrates this willingness also. Striking in this regard is that except for one participant everybody else insisted on filling out the questionnaire on the computer.<sup>3</sup>

Two participants left the start-meeting, saying this was “something for children.” Both tried to play at home initially but quickly informed me they decided to quit the training. Over the phone one of them told me he was “more of a practical man” who likes to work with his hands and be outdoors, and for that reason he did not like to sit behind “such a thing.” He further told me he found it too complicated and was not able to play it.

Two others never appeared at the start-meeting. I never heard anything from them. At the end-meeting, five people did not appear—they “forgot” or were too busy. On request all of them made the post-questionnaire at home, making their contribution to the research still useful.

### ***Organization B: Ambitious, Big and Clueless***

Like Organization A, Organization B also came into existence after a merger of several smaller water authorities, but this happened a few years ago. With such a recent merger it is no surprise that this new water authority was still looking for how to organize its tasks and responsibilities. The smaller water authorities that existed before the merger had each their own ways of dealing with their area and now they sought to cohere into a single model and apply this to the rest of the organization.

The area Organization B covers is partially industrial and partially rural, and with 101,809 hectares is much larger than Organization A's area. It consists of a number of islands, one reason why it is responsible for many levees: 779 kilometers in total. Consequently, the organization has a different strategy than Organization A. As they told me explicitly, they are very much dependent on their volunteers, which help fulfill a manpower shortage (even in addition to full-time employees). Tellingly, they have 650 patrollers in total, making them have one of the largest inspection organizations in the Netherlands, and they continue to recruit others.

<sup>3</sup> At Organization B, no one filled out the questionnaire on paper. Similar to Organization A, at Organization C only one person chose to use paper and pen instead of keyboard and mouse.



About two-thirds of the total number are volunteers and many of them have been recruited in the past couple of years. Where Organization A became inactive, Organization B did the opposite. They have organized various activities for the patrollers throughout the year and made sure each one participated in the standard levee inspection course by Deltares (Level 2). To further professionalize the organization, they set up a portal where all patrollers can find and retrieve information. They also understand that because these people do this on a voluntary basis and join partially for social reasons that it is important to have a good time. For this reason, to celebrate the end of the storm season in the year of the training, they hired for one evening a complete cinema and invited patrollers to watch *De Storm*, a recent Dutch movie about the flooding disaster in 1953.

Unlike the patrollers of Organization A, the patrollers of Organization B are affiliated with a certain region and levee segment within this region. Organization B has five large regions with each their own guard post.<sup>4</sup> Each levee segment has a leader to whom any findings need to be reported first. These leaders communicate with their affiliated guard post, which are led by a designated commander.

Unlike the patrollers of Organization A, the patrollers of Organization B are affiliated with a certain region and guard post within the area. The area counts five large regions with each their own guard post. Within these regions different guard posts exist who each have their own levee segment. Each guard post has a commander to whom any findings need to be reported. These commanders are mostly employees but some of them are volunteers. I heard that the guard posts organize events on their own behalf as well—which oftentimes happen to start or end at the local bar.

The levee inspection at Organization B is far more organized (and it needs to be) than at Organization A. Organization A's volunteers only know that if something occurs they have to go to a single location, no matter where they are from and where problems are occurring. At this location they are told what to do.

Personnel within Organization B profess to be enthusiastic fans of the game. The inspection coordinator, who was part of the plans to develop the game from the beginning and who remains one of the biggest proponents of its use, told me he used the game to teach at a university. He said “within an hour all of them knew what failure mechanisms are...I have seen that it works.”

Organization B expressed, however, a concern with how they need to train all the patrollers with the game. Before the training, they only distributed a laptop with the game to the five guard posts to enable people to play it when they had a chance. Therefore, for this organization the training was a valuable opportunity to find out if this was a viable way for educating their many patrollers.

### **A luxury problem**

In arranging the training for Organization B we faced a luxury problem. With a base of 650 potential participants, we were afraid too many people would respond

---

<sup>4</sup> In Dutch, the guard posts in regions are called *dijkposten*. Literally this means “levee posts.”

positively to an open invitation. With a similar response rate to Organization A I would have needed to train about 200 people. And Organization B thought it might be a lot more. They received many requests the past years of people asking “when are we finally going to play that game?”

Such a large number of participants was unwanted. Not only did I not have the capacity to run all those training activities, it would have surpassed the purpose for which Organization B wanted to use the training: to see if it is an appropriate way of training their people. This would only require a small portion of the population. In addition, Organization B wanted to invest in this training. They said the gift cards were unnecessary. Instead, they would give each participant their usual compensation of 25 euros per event. As the training consists of two events, a start-and-end-meeting, each participant would receive 50 euros, thus elevating costs.

Unlike Organization A, they agreed to arrange for everything surrounding the training. The administration, location and all the other aspects were this time taken care off by the organization. I only needed to create the invitation and make sure I was there at the right time and right location with the training equipment and one assistant-facilitator. Quite a contrast with Organization A was that five people helped me coordinate this project: an intern similar to the one at Organization A, two junior staff members, and two senior staff members.

Eventually we decided to organize one training per region and let the meetings take place at each of the operating bases affiliated with that region. Then from each region we would allow up to 17 participants to participate on a first-come, first-served basis. Organizing it in this way ensure for a reasonable travel time for the participants and getting input from all five regions. Most importantly, it gave an overview, a structure, with which we could work with and whereby we were guaranteed that not too many people would participate. Invitations were sent two weeks in advance of the training date to avoid too much disappointment. Further, the invitation clearly said that this was an experiment and that if successful, similar training opportunities would follow.

According to the junior staff member who arranged the administration, the responses were overwhelming. Only in one region the response was low. This concerned a region with many employees. Just as those employees at Organization A, the employees of Organization B also seemed to not like to be busy with work outside the regular working hours.<sup>5</sup> Something else may have played a role here too. At this organization employees are part of the inspection on a “voluntary compulsory” basis. This contradictory notion means that certain employees are forced to join. It is compulsory. Yet, any inspection activity takes place in their own time. It is a voluntary activity. Consequently, I was informed, they are generally less motivated.

It further did not help that the meetings for this particular region were planned on Friday evenings. I knew this disadvantage in advance. Due to the many holidays during that time and the upcoming training with Organization C, it was impossible

---

<sup>5</sup> I learned from the sessions with Organization A that it would be possibly better to organize one session during the day. However, as we organized it per region and this already resulted in five sessions, I decided to not organize an additional one for employees.

to plan it on any other day. In the end, we mixed this region with one of the regions with an overwhelming response.

The total number of initial participants came down to 82. This number includes three of the coordinators. They wanted to participate as well. The other two did not find this necessary.

### **Hot and tight**

Based on the experiences with Organization A, the start-meetings were rigorously changed (see “Things That Were Improved” later in this level). In brief, I decided to plan one session per evening, extend the time by one hour, let participants play on their own pace, and include a demonstration session. The changes had the desired effect. Feedback among participants was generally very positive and enthusiastic.

However, certain issues arose. The guard posts were not ideal. Except for the headquarters, all other posts were small. They consisted of five small spaces: two offices, one commanding center (with communication equipment), one meeting room (with a big map of the region), and a kitchen (with a coffee machine). In every post the training could only be done in the meeting room. This space hardly fit fifteen people—and we were mostly with a bit more, about seventeen to nineteen. It was also uncomfortably warm due to the weather, making the training location hot and tight.

That these posts were not designed to have over fifteen laptops running was made clear to us in a rather stressful manner. More than once the electricity went out. And while preparing the sessions, nobody was there to assist us. After we arrived (at about 5 p.m.), the last person on the base gave us the keys and the coordinator that evening arrived just a little before the training with all other participants (at about 7 p.m.). Luckily, we always figured out how to get the electricity up and running again before the training started.

One time, however, was really cutting it close. This post was stationed at a pumping station and we looked in every corner of the building to find the power cabinet. Upon finally finding it, it turned out we needed a key to get to the switches and, of course, we did not have this key. We had to think of something else. In the end, we found lengthy extension cords and tapped electricity from the complete other side of the building. We got everything working just a few minutes before the training had to start.

The administration of the training also did not unfold as planned. I am pleased that Organization B handled the administration. Without them I certainly would not have had so many participants. But in inviting and confirming the training, the organization mixed up the dates for two sessions, and although they largely restored this error by calling everybody, a number of participants (five in total) were not able to attend the start-meeting because of this. For other sessions some participants received the confirmation very late and, consequently, did not attend the start-meeting. A number of participants who missed the start-meeting were still willing to continue the training, but they did not get much further than filling out the questionnaires.

Beyond the dropouts due to administrative errors, one person suddenly had to go abroad for work during the time of the training. Four others participated but quickly expressed a desire to quit the training, for various personal reasons: one estimated the training would take him too much time, another experienced cybersickness, a third had strangely enough an aversion against computers, and a fourth told me he played the game with “blood, sweat, and tears” and that it was not fun for him.

Only four participants were not able to attend the end-meeting (because they were busy or sick). Effectively, the total number of participants of Organization B who contributed to one or more parts of the training was 77 people.

The end-meetings were characterized by a positive atmosphere as well. At the end of three of the five sessions the facilitators received a huge applause by the participants. This contrasted notably with my experience at the first meeting of Organization C.

### ***Organization C: Structural, Critical and Rural***

Organization C also came into existence after a recent merger. Although it is relatively older than Organization B, it has made less progress in professionalizing its levee inspection organization. In 2009, they started to structurally train every year and recruit patrollers. This recruitment is specifically oriented at volunteers. Unlike Organization B the water authorities before the merger worked largely with employees and as a result Organization C had few volunteers to start with. Organization B and C also work closely together and there is much exchange of information and ideas.

But when Organization C copies they give their own twist to what they copy and try even to improve it. They are what the Germans would consider *gründlich*. They look with a critical eye to the possibilities and if they implement it, they do this structurally. They do not recruit their volunteers through newspaper advertisements, something what Organization B does, but ask citizens they are in contact with and of who they know they are somewhat knowledgeable about levees.<sup>6</sup> They further couple volunteers to employees. The idea behind this is that the volunteers have region-specific knowledge; the employees know the organizational rules and specifics.

Also, unlike Organization B they decided to not give away gear to patrollers. Giving away runs the risk of patrollers losing their gear, so they decided to stock the gear at the operational bases and hand it out when needed. Moreover, they implemented a communication protocol between the patrollers and the operator at the operational base. This is the protocol based on *Levee Patroller* (Level 2). They are also the one that adopted the levee patroller terminology by calling their people patrollers and not guards like the others do. Organization B is the other one that uses the terms, but they do so less pervasively.

---

<sup>6</sup> This strategy of recruiting volunteers via-via by Organization C is something Organization B cannot really permit as they need far more volunteers.



(a) Completely prepared for the first session



(b) One of the sessions

**Fig. 4.2** An impression of the sessions of the Virtual Levee Inspection Training

The area Organization C has to inspect is remarkably different than those of the other two organizations. It is completely rural, consisting of large swaths of farmland for cattle and crops. The area is inland and so does not have any sea levees. It does have many river levees as two big rivers cross the region. These rivers caused major floods in the past, also recently, in the early 1990s. After these floods the levees were fortified, a process that required the removal of houses and other structures, which caused lingering resentment among those affected.

The area is divided into six regions with each their own guard post, similar to Organization B. The total area is much larger, counting 201,000 hectare, yet they are able to inspect this with a smaller army: 300 people in total of which 150 are employees and 150 volunteers. Much similar to Organization B as well, each employee–volunteer couple has a predesignated levee segment to inspect.

Beyond the yearly field exercise, the organization provided their own lectures to inform (new) patrollers about failures and the inspection procedure. They were of the opinion they had expertise enough and did not need to involve Deltares to do this. This lecture also did not involve the game. Pretty much Organization C was as far with using the game as Organization B—only some employees played the game. Their attitude only differed. Where Organization B did not need to be convinced, Organization C needed to be. They wanted to use my training to really test whether they could see an improvement. Based on this they would decide to further invest in the game or not.

The training setup at Organization C reflects this purpose. It also reflects their tendency to be *grundlich*. If they do it, they want to do it well.

### **Compulsory and comparable**

The training at Organization C differed in a number of significant ways. First, the training was made *compulsory*. The participants—whether volunteers or employees—had to attend unless they had a valid reason not to. The organization considered the training to be of high importance. In addition, if this was not done it could happen that many employees would not participate, something what happened with the other organizations, and that only the “enthusiasts” would participate. These are the people who have a positive attitude to the use of the game already. Then the results would be skewed and non-representative.

To show the importance further, the organization wanted to reimburse participants not only for attending the meetings but also for playing at home. However, this was deemed impractical, and so it was decided to give participants the 12.50 euros gift card in addition to the 25 euros for each attended meeting.

Another difference was the choice to focus on one of the regions. Organization C saw this as an experiment and wanted to train the other regions if the training was found successful. To judge the success, Organization C proposed to compare this region with another, rather similar region during their next field exercise. I was a bit skeptical about this. Such comparisons are heavily dependent on the circumstances,

but I went along with the proposal in the interest of generating some interesting results (Level 10).

Two training sessions were organized. Both took place at the headquarters of Organization C. I conducted the administration of the training. I received a list with names and addresses and sent invitations to 35 people. Other than the patrollers from the region, this group of people involved all the commanders of the other regions. In case the training would be provided to their region, they would have a good idea what the training is about by participating already. The number also includes the two coordinators, both senior staff members, with whom I was in touch with. They wanted to participate as well. This meant in terms of the number of employees and expertise the training at Organization C was significantly different from the other two organizations.

A number of invited participants reacted negatively toward the training, and for two reasons. First, the start-meetings were planned in a vacation week in that region, something of which both I and the coordinators at Organization C were completely unaware. And as a consequence some could not attend the start-meeting simply because they were on vacation.

Compounding the dissatisfaction over scheduling was the fact that it was made compulsory. The volunteers were especially angered, because they were, after all, volunteers. Organization C saw this differently. In their eyes people voluntarily decide to become part of the levee inspection organization. By being members they are, however, required to participate in activities. The organization needs to be able to depend on their members.

### **Viva la résistance**

On the first evening the discontent about making it compulsory—and during a vacation week—was directly put on the table. The volunteers in the room were of the opinion that the organization should have never used this type of wording. Obviously, this was an issue between the organization and their members, but at that point I had to address it. Together with the coordinator that evening we tried to assuage the discontented participants by agreeing it should have been phrased differently.

However, the damage was irreparable. After this discussion I hardly made clear that this training was an “experiment” and before I knew it a small number of participants, and one person in particular, questioned the training. They did not want to participate in something that was not proven (disregarding that the point was to actually find this out).

This was an unpleasant experience and totally unexpected, especially after all the positive energy of the training sessions at Organization B. As the Organization B training was such an improvement upon that of Organization A, I mistakenly expected the trend to continue at Organization C.

Furthermore, issues of computer literacy were more pronounced in this training. Many participants were old-fashioned, classic farmers, with a beard and all (one with a beard that extended down to his pants). As much as I tried to explain them

the purpose behind the training or help them, they could not understand its purpose. At the end-meeting the biggest rebel of all remained skeptic and said the training was pointless, because “you do not also walk with a mouse over a levee, right?”

The mouse was an object of interest for more of these participants. When the participant with the long beard sat down, he picked up the mouse, looked at it as if he had never seen it and said:

I have seen such a thing before, but I have never ridden it in my life.

After finishing the sensemaking test, he came to me and said he lived all his life along the levees and knows them thoroughly. Despite his apparent expertise, he answered each question with “report failure.” He did not specify what failure. He also left the training without telling me that he does not own a computer. This was a rather clear sign: he really did not want to participate in this training. After a week I called him and when I found out about this, I offered him a loan laptop. He refused this as he told me he was far too busy with his agricultural obligations.

The rebel gave me the same reason for quitting the training, commenting that:

In May every farmer is busy. So am I. And I cannot play this game...not that I do not want to, I would love to be able to use a computer. I just cannot...it would take me too much time to learn it. If this training was hold in November I would have more time, but then still, I question whether I will be able to learn this. I have tried it before and that did not work...but then I wonder, why do I need to use it? Why would I need to invest time and effort?

He seemed to jump on two thoughts. On the one hand, he persisted on saying the training is useless, even at the end-meeting. Generally he was further frustrated with the fact that everything became computerized—now he even had to use a computer for inspecting levees! On the other hand, he indicated that he would like to learn how to use a computer and play the game. He is only aware that this would cost him too much time and effort and that is something he is not willing to do, especially because he has doubts whether he is able to acquire the skills in the far end. So it seems that instead of showing this latter uncertainty and asking for help, he chose to resist any changes. Resistance to change is easier than change. In retrospect it also seems that the game became the object of his frustration with (being forced to use) computers.

To prevent a revolution on the second session as well, I asked the other coordinator to explain the need of this training and the reason why it was compulsory at the beginning of the training. This helped. This session proceeded without any problems.

Eventually only the rebel and the participant with the long beard quit the training. Two of the three vacationers who missed the start-meeting never really participated. On the end-meetings many participants stayed away, but some completed the training in their own time. Others “forgot” the training and like the vacationers did not participate as much.

Remarkable about the meetings at Organization C is that I had to quit the meetings somewhat earlier than with Organization B. With Organization B I sometimes had the feeling I could go on forever, that is how much they liked it. Whether they liked it or not, at Organization C people wanted to go home at some point.



The above paints a somewhat negative image about the training at Organization C. However, most participants participated enthusiastically and expressed frustration with the skeptics.

## Setting the Facts Straight

If we look at the facts, we can confirm that the negative image of the training is not accurate. In fact, it is quite the contrary. Based on Table 4.2, which highlights a number of facts surrounding the training, it is reasonable to conclude the training was a major success.

Being involved in a three-week training on a largely voluntary basis is demanding and I would not have been surprised if the training turned out to be a total failure. Instead, 71% played all exercises. Counting also those who almost played all exercises, it yields a figure of 80% who participated very well. That is a participation rate far beyond my own initial expectations.<sup>7</sup>

The facts further show numbers that one can expect in any training. One will almost always experience people who cannot attend meetings, who decide last minute to not participate, or who decide to drop out. The numbers of this training do not show anything of a particular concern. Consistently at all three organizations, about 5% dropped out, and for various reasons: they were too busy, did not like computers or the game, or experienced cybersickness.

What can be noticed from these numbers is that commitment to the training is indeed important. The number of people not attending the end-meeting is much higher than of the start-meeting. Participants had again various reasons for not attending. In most cases the participants not attending the end-meeting were the participants who played two or less exercises. They may found it useless to attend the end-meeting being unprepared or may have actually dropped out of the training, but without telling me.

From the beginning I knew commitment would be important (Level 3). Therefore, I was not surprised to see that the *non-participants*, those not attending the meetings, did not participate. They have no reason to finish all exercises in time—they have simply no commitment. I think that this is an important learning lesson for implementing a game-based training.

The need for commitment is also confirmed by looking into the numbers of people who played the game after the training. I kept track of who played the game till the end of the year and it turns out that 14 people played it (10%). Eight out of these only played one more exercise. Just two people played it more than three times after the training. This means we can definitely conclude that without any structure or need to play the game, it simply will not be played—despite of the good will of the participants, because many told me they wanted to improve their scores.

---

<sup>7</sup> The actual participation rate may even be higher than 80%, because some participants indicated that they played certain exercises, but I never received those. For calculating the participation rate I based everything on what I received.

**Table 4.2** A number of facts about the training

Fact	Organization, in <i>n</i>			Total	
	A	B	C	<i>n</i>	% <sup>a</sup>
Initial group	43	82	35	160	100
Stay-awayers	2	3	0	5	3
Non-participants	4	2	2	8	5
Participants	37	77	33	147	92
No start-meeting	1	3	2	6	4
No end-meeting	7	8	6	21	14
Exercises at home					
Dropouts	2	3	2	7	5
0–1 exercises	5	2	2	9	6
2–3 exercises	6	5	3	14	10
4–5 exercises	1	7	4	12	8
6 exercises	23	60	22	105	71
Played after	2	8	4	14	10
Visits	0	1	1	2	1
E-mail <sup>b</sup>	17	35	10	62	42
Internet visits	22	47	12	81	55
Loan laptops	4	6	6	16	11

*Note.* Stay-awayers = participants who signed up but did not participate from the beginning and did not receive the training material; Non-participants = participants who received the training material but did not participate; Dropouts = Participants who decided to quit the training.

<sup>a</sup> The percentage of the first four facts is based on the initial total group ( $N = 160$ ). The percentage of all other facts are based on the total number of participants ( $N = 147$ ).

<sup>b</sup> Some participants e-mailed me more than once. The total amount of e-mails reached 75 and not 62.

To foster commitment throughout the training I called a number of participants after the first week. I called every participant who did not play at all and occasionally a participant who played just one exercise. If they did not play, I first always asked if they already installed the game. Then I asked if they experienced any difficulties playing the game or had any questions. Besides fostering commitment, I had two other reasons for making these phone calls: I wanted to let participants know that they can reach me if they experienced problems and I wanted to know if participants experienced any issues already.

During these conversations few told me that they had problems and throughout the training I received a few number of phone calls. I was therefore surprised to hear at the first end-meetings that some participants did not know how to call the Action Center. After hearing that I encouraged the participants of the other two organizations to contact me:

Call me, stalk me. I am here for you. If necessary I will even come visit you. The only thing I ask from you is that you give me a nice cup of coffee...and a piece of cake.

Nobody called me more than once except for two participants. One even called me at 8 a.m. on a Saturday, saying he had to go out to his horses and he still was not

done with the exercise. He asked me if I could guide him through the exercise so he could finish it quickly.

I did not keep track of all the phone calls. However, I would say that I received less than 30 calls in total. Quite a substantial number of these calls dealt with practical issues surrounding the training and not with difficulties in playing the game. I also only ended up visiting but two participants.

Contrary to the phone calls, I did keep track of all the e-mails I received. In total I received 75 e-mails and I categorized these. For every category the number of e-mails I received is indicated between the parentheses.

- *Announcement*: About quitting the training or not being able to attend a meeting. In addition, some participants sent me an e-mail to apologize for not being able to play the planned exercises (27).
- *Game problems*: About changing the controls or not being able to continue to the next exercise (17).
- *Computer problems*: About using an Apple computer or having graphics issues (13).
- *Installation problems*: Questions about installing the game (10).
- *Code request*: To be able to start earlier or to remind me that I should send the code that day. I always did this. Some participants were probably eager to play the next two exercises (8).

This demonstrates clearly that like the phone calls, the majority of the e-mails concerned practical issues. Although it might be that the information in the game, in the manual, and on the website was more than sufficient, it is good to keep in mind that participants do not ask for help to get a better understanding of what they are learning.

I kept track of the website too.<sup>8</sup> Every time participants entered the website, they had to fill out their names. Based on this, I discovered that more than half (55%) visited the website. Of these participants, 33 (41%) visited the website more than once.<sup>9</sup> Throughout the complete training 144 registered visits were made. The average time on the website was 6:43 minutes.

During the discussion some participants indicated that they found the website useful. I also received complaints that participants could not access the website while playing. One participant found a solution: he used two screens to play the game, one for the game and one for the website. But based on the statistics we can conclude that the website did not play a major role of importance.

I handed out *loan laptops* to deal with some of the practical difficulties. By doing this, participants could temporarily get a laptop to play the game. In a certain way one could count handing over these laptops as a visit as well. Frequently I had an

---

<sup>8</sup> I tracked the website statistics with *Google Analytics* and a simple text file that was created after people filled out their names.

<sup>9</sup> According to the website statistics from March 1, 2010, to June 2, 2010, I had 280 visits by 156 unique visitors, a bounce rate of 16%, and 55% new visits. The latter is almost consistent with my findings based on the names that were filled out before entering the website.

extensive conversation with the participants about who they are, what they did, and how they look at the levee inspection.

About 11% ended up using a loan laptop and this is something of concern, especially for organizers of game-based trainings. This means that extra resources and facilities are needed to run a training like this. However, I expect that *a)* this number will decrease for *Levee Patroller* in the nearby future, as people will buy newer and better computers over time; *b)* this percentage will be much different for a younger target group, as they are more likely to have better computers; and *c)* this becomes less of a problem if web-based games are used, something most designers nowadays seem to prefer to use. Despite these expectations, certain people will always experience computer problems. That is almost a given. It is also a given that with a complex training setup problems can be expected.

## Errors and Improvements

Of course, not everything proceeded according to plan. The idea of this training/evaluation was to gather data and for various reasons this did not proceed as one would hope for, because several things went wrong, which resulted in missing data.

In addition, although various bug tests have been performed over time, the game had never been used on such a scale, and, therefore, certain game errors slipped through the cracks. Then the training itself was an “experiment.” It is not strange to find points for improvement after its first implementation. Because it was a training too, several things were improved.

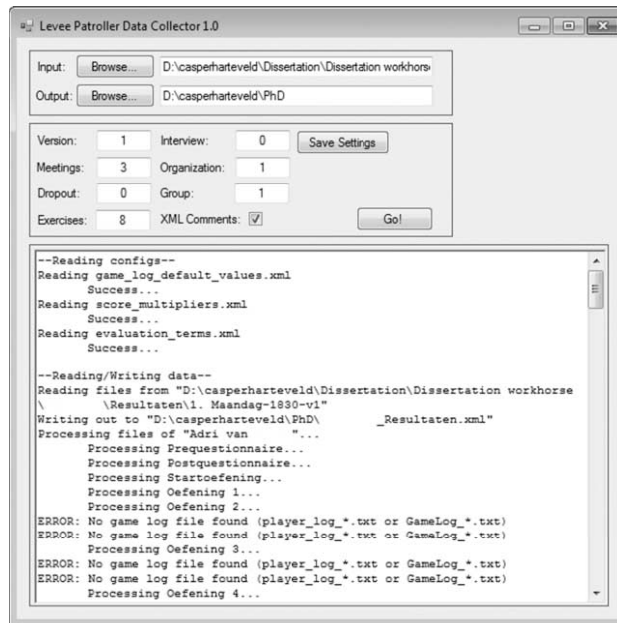
## Things That Went Wrong

To explain what went wrong I need to spend a few words on the data collection. The pre- and post-questionnaire and pre- and post-sensemaking test are websites and participants filled these out on a laptop during the meetings. Its results were saved locally on the laptop—with a “local web server”—in a single file with the name of the participant, the date, and time (e.g., `output.casperharteveld_2010-03-23_21-19-29.xml`).<sup>10</sup> If participants could not attend the meetings, then I asked them to fill out the questionnaire and test at home. This output was saved on my web server space.

The game data consists of two files per exercise: the game questionnaire output (e.g., `evaluatie.casperharteveld_2010-03-09_20-04-01.txt`) and the game log (e.g., `player.log.casperharteveld_2010-03-09_19-15-54.txt`). Both were automatically saved locally onto the computer and submitted to a server. The creation of the files happened as soon as participants started playing an exercise or started filling out the

<sup>10</sup> I used *XAMPP*, a free and open source web server package, to be able to run websites without Internet.

game questionnaire. I therefore did not only receive the completed files. I could see if someone started an exercise and then restarted it.



**Fig. 4.3** *The Levee Patroller Data Collector Tool*, the tool to integrate all the separate data files into a single database. Tool is created by Gert-Jan Stolk

If a person finished the complete training I would receive at least 18 separate files of this person: two questionnaire/test files, eight game questionnaire files, and eight game logs. Based on the number of participants and their effort in the training I calculated that I should have received at least 1,269 files. To assist me in sorting the files and creating one large database, I made use of *The Levee Patroller Data Collector Tool* (Fig. 4.3).

With this in mind we can turn to my top ten list of things that went wrong, starting with what caused me the most headache toward the things that just made me cry a little bit.

1. *Disaster strikes!* Training people to prevent disasters does not mean disasters cannot happen in the meantime. During the training a lightning bolt hit the vicinity of where the game server was located, causing a notable loss of data on Organization C.
2. *Server maintenance and shutdown:* The game server was running on the company's servers and whenever maintenance occurred, the game server shut down and had to be restarted. Quite logical yet not something I thought of at the start and the reason why we did not anticipate it very well the first time. Three days

passed before we restarted the server. Although we made sure to pay attention to these maintenance times after that, in between shutting down and restarting the server it could have happened that somebody played the game.

3. *You have...0 points:* After the ending of the training with the first organization, I was surprised to find out that for about 27% of the exercises played, participants reached a score of zero on everything. I was quite in shock at first, but quickly realized that something was wrong and was able to repair the technical error. However, I did have to calculate all those missing scores by analyzing how they played the game. This reminds me that even if you think you have pretested sufficiently, it is not sufficient. You always overlook something.
4. *Running out of time:* This refers to me completely underestimating the time needed for the first meetings. Instead of 1.5 hours the first meetings took 2.5 hours and this was still cutting it short. Due to this the participants of the first organization went home less prepared. It was a good learning lesson and something I made use of in the next training sessions.
5. *You are not the administrator:* One of the important rules of thumb of software testing is to test on different platforms. This I learned unfortunately after the training already started. If participants were running Windows Vista on their home computer, were not the administrator, and did not say to run the game as an administrator, no log files were written onto the computer. Luckily, relatively few people had this and I was still able to receive the files over the Internet—if they were connected, of course.

In other instances I discovered that for other unknown reasons log files were not written onto the computer. I found this even on a laptop I gave to one of the participants. All log files were present on this laptop except for two. This is quite strange and the only reason I can think of is that the laptop ran out of battery at precisely the point of writing the files to the computer. But that is quite a coincidence. Especially because it happened twice.

6. *We are all connected, are we not?* On many occasions I received half of the files, most likely caused by a failing Internet connection. This should not have been completely unexpected, as most participants live in very rural areas, areas that are known to have this problem. On other occasions, especially if the participants played at work, their network security made sure the files never reached me.

Because of the failing Internet connection I received a number of incomplete files. Some of the game logs I could repair, because the players basically finished the exercise. Just the scores were missing and so I could fix them in the same manner as I did with missing score issue at Organization A (see my No. 3). If the game log was too incomplete, I considered it *incomplete and irreparable* (Table 4.3).

7. *Changing things in between is a bad idea:* Another important rule of thumb is to not change something that “works,” and certainly not during an experiment. I thought otherwise about this, because as I explained earlier, it was a training session too. Therefore, if it seemed an easy fix, I rather fixed this than to see the same frustration and errors with another group. Changing software in between

is risky on the other hand, as it requires to test everything again and it is quite possible to overlook something (see my No. 3).

This happened when preparing the software for the third organization. The layout of the game questionnaire looked a bit different—different enough to possibly choose the wrong checkboxes. Luckily, I discovered this before the training started, but after all the discs were already produced and no time was available to make new ones. We created a software patch and I sent this to the participants right away. I also informed them about this error during the start-meeting.

8. *Is it today or tomorrow?* This concerns the administration of the training at Organization B. Dates of two sessions were mixed and participants received their confirmation letters late. Therefore, a number of participants did not attend the training. Some still tried to participate, but were clearly upset about it.
9. *Goodbye questionnaire:* I set up the training in such a way that I would have a 100% guarantee I would get the pre and post-questionnaires. Besides saving the complete output of the questionnaire, I also saved each separate page, just in case something would go wrong. However, in four cases, the complete output as well as certain separate pages were not saved for some unknown reason. It is too bad but I had worse things to cope with (see everything above).
10. *It does not work:* I expected that the game would not work well on some computers. Some computers were simply too old or did not have the requisite graphics card or compatible operating system. Three participants did not have a computer. For these problems I had a number of loan laptops.

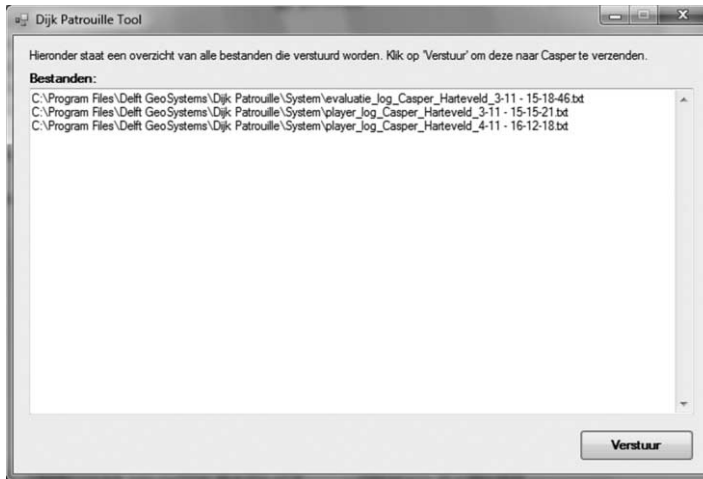
Some of the listed items could have been prevented if I ran a pilot study first and not only a small test session to see if the technology works. If I analyzed the data of the pilot study I may have found for example the missing scores issue (No. 3). I would have definitely learned about the time issues of the start-meeting (No. 4). However, I could not afford the luxury of a pilot study (Level 3).

Looking back I could have ensured for more resilience with the data collection, that is more workarounds to ensure I would receive the game data. We could have enabled the possibility that every time the participant's computer connects to the server all files are submitted or synchronized. Another option could have been to add a button onto the main menu. All files would be submitted by pressing this button and I would have asked participants to do this before going to the end-meeting.

The reason we did not go that far is that I was quite happy that we found a way of ensuring for some resilience by being able to save the files locally. This took away many of my worries. And I underestimated the amount of files I did not receive. I thought that I would visit those participants with many missing files and put these missing files on a USB stick. In hindsight, this presumption seems naïve.

Seventy-three participants had one or more files missing. This is about 50% of the complete training group. To deal with this another tool was created—*The Levee Patroller Log Retrieval Tool* (Fig. 4.4). If participants did not change the destination folders, this tool is able to find the game data files and send them to the server.

I sent this tool to those participants with missing files by means of an e-mail with a URL link to download it. Unfortunately, the response was subpar. Only 12



**Fig. 4.4** *The Levee Patroller Log Retrieval Tool*, the tool to retrieve the log files. The Dutch text says “Below is an overview of all files that will be submitted. Press ‘send’ to submit them to Casper.” Two types of log files are submitted, the “player\_log\_name” files and the “evaluatie\_log\_name” files. The first is a log file that contains information about how the scenario was played (including the scores). The second contains the output of the game questionnaire. Tool is created by Almar Joling

participants made use of the tool. With others the tool did not work, they deleted the files already, or they simply did not respond to my request. Except for this tool, I retrieved files during the training by e-mail; collected them on the end-meeting after asking participants to bring their laptop with them; and put missing files on a USB stick during the post-interviews or other visits. Table 4.3 provides an overview of the resulting missing data of the training.

Despite the setbacks, it is vital to learn from experiences such as these. And these “things that went wrong” were not catastrophic. Furthermore, although I eventually had a significant amount of missing data, a rate of 11% missing data is acceptable, suggesting there is sufficient data to investigate.

### ***Things That Were Improved***

Throughout also game errors or points for improving the training were found. Some errors and points became clear to me during my observation as a facilitator. Most were indicated by the participants themselves. These suggestions were especially made during the discussions at the end-meeting. Some “zealous” participants further suggested improvements by e-mail or over the phone too. A few even provided me with screen shots with clear explanations of what needed to be fixed.



**Table 4.3** The missing data of the training

Missing data	Organization, in #			Total	
	A	B	C	#	% <sup>a</sup>
Meetings <sup>b</sup>					
Pre-questionnaire/test	1	3	1	5	3
Post-questionnaire/test					
Not received	2	5	4	11	7
Incomplete	1	1	2	4	3
Same as pre-test	0	2	0	2	1
game questionnaire	0	6	0	6	2
Game logs	0	3	0	3	1
Home					
game questionnaire	11	28	28	67	9
Game logs					
Not received	7	4	17	28	4
No scores	40	0	0	40	6
Incomplete and reparable	9	12	3	24	3
Incomplete and irreparable	3	10	1	14	2
Total missing <sup>c</sup>					
In #	24.5	60.5	52	137	11
In %	9	9	19		

<sup>a</sup> The percentage is based on the total I should have received based on the number of participants and their effort in the training. The stay-aways and non-participants were excluded from the calculation.

<sup>b</sup> At Organization B I had some issues with data retrieval on the meetings: two participants received the same set of pictures with their sensemaking test; two participants played the first exercise instead of the start-exercise; three participants did not fill out the game questionnaire; and one participant brought his own laptop on both meetings, with Windows Vista and no administrator rights, making it impossible to retrieve his files.

<sup>c</sup> In calculating the total, those files that could be repaired, the “no scores” game logs and the “incomplete and reparable” game logs, were not taken into account. In addition, I counted “incomplete” post-questionnaires/tests as well as post-questionnaires/tests “same as pre-test” as a half, because they were not completely missing.

Seen from a purely experimental point of view, changing the apparatus of the investigation is unfeasible. If changes are made, it means that any effects might be due to these changes. In other words, it means the comparison among participants becomes less valid—its “internal validity” is harmed (Cook & Campbell, 1979). Some purists might argue it is not valid at all. From this perspective, I should not have changed anything throughout the training.

However, in this case, I chose to. This experiment was unique in that it doubled as a training. The participants needed to leave the training with more knowledge and skills about levee inspection. I was there to help them and not to frustrate them. In addition, the purpose of the training/evaluation was also to learn from. By making incremental improvements we learn what works and what does not.

Being aware of the validity issues and the limited time I had at my disposal in between the different groups to fix something, I only improved the game or training if *a*) it constituted a minor aspect, something that would not have a major impact on the training; and if *b*) participants were clearly frustrated with this aspect. If participants were not frustrated I had no reason to change it right away.

Aside from really small changes, such as some textual errors (think of misspellings), the following improvements were made:

- *Knowing the time:* After the training with Organization A a clock was added at the bottom left of the screen. In this way players had a better sense of time and knew when a scenario would end. I noticed that quite a number of participants stopped playing an exercise just before it ended. This meant they had to redo the whole exercise. This was not very motivating.

A disadvantage of the clock is that participants were arguably less immersed into the game. Not having a sense of time is one of the indications of immersion and if a clock is continuously ticking, this is not helping. Another possible disadvantage is that it may increase a feeling of stress, as the clock makes clear to players if they are running out of time. It would have been better if the game gave a rough indication of time, but there was sufficient time to implement this.

- *Hello and goodbye!* Participants widely complained about not being able to pick up the phone call by the Action Center. In this phone conversation the Action Center says hello, tells the player how many failures reside in the region, wishes good luck, and says goodbye. The only real informative part is the indication of the amount of failures, yet this is something players can easily lookup with the statistic tool from their inventory. Despite this, some players wanted to have this conversation so much that they restarted the exercise for that reason alone.
- *Cumbersome conversations:* Participants were generally frustrated with the Action Center. This frustration is part of the play between the player and the virtual Action Center, but I had been hearing so many times that people found it so “cumbersome” that eventually I decided to do something about it. This was just before the start of the training with Organization C. Before the change conversations with the Action Center included a summary of the findings. We included this to emphasize that the Action Center is dependent on the information given by the player and to give the player the possibility to review the findings. I noticed that players did not read this well (if they read it at all) and deleting this part of the conversation would make it shorter and cleaner—and not so cumbersome. In the newest version the players simply say they submitted the findings. To explain this way of communicating, I told the participants at the start-meeting they were using some sort of new technology with which they could report their findings and submit these to the Action Center.
- *I want it on paper!* Players could retrieve information from the game’s handbook tool in their inventory, the manual, or the accompanying website. On purpose I spread the information over these sources (Level 3). The handbook contained basic information about levee inspection; the manual information about how to play the game; and the website provided specific and detailed information. However, the participants of Organization A indicated that they would have liked an overview of all the signals in the manual, something the website did provide. I yielded to this request, although I made sure the website still provided extra information about the signals.

Interestingly, the participants of Organization B who received the improved manual had a similar complaint. They rather wanted to have a manual that in-

cluded an overview of the possible measures, which the website did provide as well. To this I also yielded. The participants of Organization C so happened to be fortunate enough to receive a manual with an overview of signals and measures. They had no complaints about the manual.

- *Only one of each kind, please:* The biggest small change was the cause of the most frustration by the participants. In the original version, players are able to report an unlimited amount of signals. If they report the same signal, the game just adds a number to each similar signal. It could therefore happen that one failure has three cracks with the names “crack (1), crack (2), and crack (3).” We wanted to give players this freedom, as some failures may consist of more than one similar signal.

Two game rules turned this freedom into frustration. First, the game requires that every report is completely filled out; otherwise the Action Center will complain. The player cannot continue until everything is complete. Second, the game ideally requires that if an earlier reported signal changes, a player adds a second report and does not create a new signal. If players do the latter, they have to fill out the report completely. If they add a report to an existing one, they only have to change the elements that are different.

What I observed during the sessions that players entered an almost endless loop with the Action Center if they had one or more incomplete reports. The first reaction of players was almost always to add a new signal instead of modifying or deleting an existing one. Adding a new signal does not solve the problem. So it happened that some players reported sometimes ten cracks or seven settlements! From the game data I saw that at some point some players received a Eureka moment and started deleting everything. Other players gave up and left the failure alone.

Creating a new signal for every observation instead of adding a new report was less problematic, but it did require more effort on behalf of the players and made the game more cumbersome. This is unfortunate because the purpose of the “add report” functionality was to make it *less* cumbersome. Before I could ascertain the size of this problem, I felt that I had to do something to protect the players against themselves.

Inspired by the story of Noah’s Ark, I made it impossible for players to add a signal to a failure if that signal was already reported. For each failure only one of each kind could be reported. After trying to report a similar signal a pop-up message appeared saying it is unnecessary to report the same signal twice and that if that signal changed, this should be reported with the “add report” functionality. Only the lucky ones from Organization C profited from this improvement.

I could not solve and deal with all possible sources of frustrations. One common frustration I could not easily solve concerned the marking of one specific failure. For the game to recognize that players find a failure, they have to place their reporting marker very close to the failure. Because of this plenty of participants were convinced they found this failure, yet the computer did not recognize their marker.

So with much amazement and frustration players constantly had to hear from the levee expert that he “looked at every piece of grass and did not find any failure” while at the same time great amounts of water were running over the levee. There was one player who must have been busy with this failure for about 15 minutes. He tried all possible measures and at least five types of signals and still did not reach any result.

What I did for this particular issue as well as for similar ones is to communicate to participants the problem during the start-meeting and in the weekly e-mails. I further corrected participants’ scores when it was clear that they had found this failure. This only served as a minor patch, as the time invested in trying to deal with this failure could have been spent on other failures.

I made one major exception regarding my rules for improving the training. I made rigorous changes after the training sessions with Organization A. First, I noticed that I was running out of time (see my top ten of things that went wrong) and so for the meetings with Organization B and C one session was planned per evening and taking almost 3.5 hours.

Second, at first we only proceeded to the next step in the training program if everybody was finished. This meant that the more computer literate participants had to constantly wait for the less literate, losing opportunity to train under guidance. Some could not even finish their exercise in time. Therefore, another change was letting participants play on their own pace. If they finished the training exercise, they could go ahead and begin the start-exercise. Only with the questionnaire and test were they still required to wait for each other.

The remaining changes were based on recommendations by participants from Organization A. In addition to the manual I created an one-page *peek sheet* which concisely describes the game steps. With it participants could quickly look up what to do. I also added a new instructional step to the start-meeting. After everybody finished the training exercise, I asked the participants to pause their game. Then I gave a *demonstration* of how to play an exercise and explained some terms and definitions, such as the failure mechanisms.

The differences between the training of Organization B and C were minimal. This cannot be said about the training at Organization A. The changes of the start-meeting made the participants of Organization B and C much better prepared to go home and play the game. The participants of Organization A had less time and explanation at the start-meeting, fewer resources, and more game errors. This should be kept in mind when reading the next levels.

## Lessons Learned

In this level I described how I arranged the training/evaluation and what happened throughout. Three water authorities agreed to participate and the training/evaluation was implemented at each. One water authority, Organization A, saw the training as an opportunity to revamp its relationship with its patrollers. They had not organized

much for years. The second authority, Organization B, was convinced about the game's usefulness but did not know how to implement a game-based training. For them the training was an opportunity to find out if this was a possible way. The third authority, Organization C, still had to be convinced and for this reason they also proposed to compare a Game Group with a Control Group during a field exercise.

The setup differed per authority, in terms of training administration, recruitment, location, support, compensation, and its premise. Especially its premise, whether it was voluntary or compulsory, made a difference. Participants—in particular the volunteers—disliked the fact that it was made compulsory.

Initially, 160 participants signed up for the training, but because some participants did not participate at all, the total number of participants came down to 147. Of this number 5% dropped out of the training and for various reasons—predominantly a dislike regarding the game. However, in general the game-based training can be conveyed as successful. A large majority (80%) played at least five out of six exercises at home, which is a participation rate beyond what was expected.

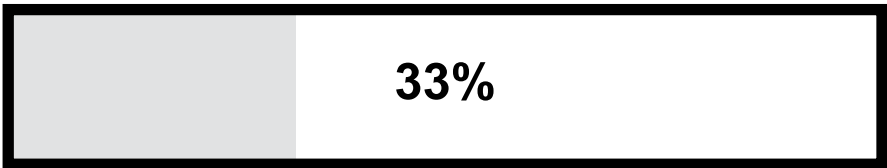
Throughout the training many errors occurred and improvements were made. The errors were especially a result of technical problems and had an impact on the retrieval of the game data and questionnaires. In total an acceptable amount of 11% of the game-related data went missing. Because it was a training too, several improvements were made and in particular after the first sessions with Organization A. It turned out that participants needed more time and support during the start-meeting to be fully prepared to play at home.

## Level 5

### Solving a Crime Is Easier

*Give me a dead body and I will solve that right away. But this...—Participant Henk (#42)*

*Henk, you have never so been so preoccupied with that inspection mess...you are more busy playing that game than solving murders—Wife of participant Henk (#42)*



33%

Two days before one end-meeting, a participant called me. His name was Henk (#42). Henk sounded excited and frustrated. He was playing the final exercises and still had some difficulties. Henk told me he was a criminal investigator and that he had less trouble in making sense of a murder than of a levee failure. His preoccupation with the game drove his wife mad. She continuously stumbled upon manuals and information related to the game in their living room. Even worse, his obsession with the game caused him to neglect her. He told me he would feel embarrassed if he did not get a good score.

Henk is one of 147 participants who played the game, all, of course, with individual experiences. At this point it is my aim to get an understanding of how participants experienced playing the game. I did not have a conversation with every one of them, but through other means I was able to get an idea. Because people played at home, I could not use laboratory types of measurements, such as measuring participants' heart rate and skin conductance. Videotaping seemed too obtrusive and practically impossible to implement.<sup>1</sup> For this reason the “means” I resorted to concerned a short and simple *questionnaire* after every exercise.

My initial idea was to make it open, almost like a diary. To inspire participants to write down their experience my plan was to ask a number of open questions.

---

<sup>1</sup> In the past I proposed to videotape patrollers but a number of them had issues with being filmed. I also looked into videotaping the gameplay and not the player, but without special equipment this would hamper playing the game too much. Even on the most advanced computers at the time the game froze continuously.

On second thought, I believed this would not elicit much information. I became skeptical about whether participants would make an extensive evaluation each time they finished an exercise. In the end, I decided to structure the questionnaire into 15 closed items and five open, making it a quantitative as well as qualitative method.

I further retrieved more “objective” measurements, such as at what days and times participants played and how long they took to finish one single exercise. I was able to retrieve this, because the questionnaire appeared right after finishing the exercise and the questionnaire and the game log had both a time stamp.

The goals of this level are to describe

- For how long, when, and how participants played the game at home (time stamps);
- How players rated their gameplay experience (closed items); and
- What prevalent gameplay responses emerged (open items).

## **Playing the Game at Home**

One of the innovations of the training was letting participants play the game at home. During the training I noticed that some participants played the game at unexpected moments—either very early in the morning or very late at night. I also noticed that many participants completed their exercises close to the deadlines and I heard many participants complain about the length of the exercises.

I looked at the time stamps of the game files to straighten these facts about “for how long” and “when” participants played. Based on the discussion at the end-meeting, my informal talks with participants, and some open question answers I was able to get an idea of “how” participants experienced playing at home.

### ***Almost Two Full Workdays***

Up front I estimated and communicated to participants that it would take about 30 to 40 minutes to play an exercise (Level 3). Each exercise takes a maximum of 24 minutes, but this excludes the time that players are in the menus. The game is paused at those moments to allow players to take their time in reporting the failures. An exercise could finish sooner than 24 minutes if players are able to find and deal with all failures before the time runs out. I reasoned that if they play it perfectly, the first exercises would take about 15 minutes and the later exercises 19 minutes. By telling the participants it would take on average 30–40 minutes I therefore thought I was being on the safe side.

However, these time estimates were highly inaccurate. One participant commented after playing the first three exercises:

I never worked with a computer before. That is why it takes me much longer than what was mentioned to play an exercise. Sometimes it took me even two hours...I do make progression however. I am now going to the next exercise and hope it goes somewhat faster!—GQex3—  
#16

This participant was not the only one with comments about the amount of time. The commenters can be separated into two camps. One camp is like Participant #16. They have problems with working with a computer and probably spent much time in the menus. Clicking on an item might have taken them already a considerable time. Within this camp, some stayed positive and optimistic like Participant #16; others just complained and were pessimistic about improving.

The other camp includes those adept in operating a computer. They found the failures quickly and had to wait till the failures became worse. From the game data I observed that some participants were able to find all failures within five minutes. This meant they had to wait at least another *five minutes* before the failures changed, because the current version does not adapt to the players' progress. The second failure phase always starts after about ten minutes. Waiting for so long is annoying and so it is not surprising that this camp commented on this.

The time stamps make clear that the fastest players spent about 20 minutes on the first two exercises at home and about 30 minutes on all others. The slower players took about 1.5 or more hours. This was about 5 and 10% of the total group, respectively. On average players took 50 minutes for the first two exercises and about an hour for all others ( $SD_{ex1-6} = 18\text{--}29\text{ min}$ ).

One extreme case should be mentioned. For three exercises, one person took about *four hours* per exercise; on the remaining half, he spent about two hours. In fact, this person was Henk, who spent a total of 17 hours on all six home exercises. This is far more than the average participant.

The average participant spent almost six hours on the home exercises ( $SD = 2\text{ hr}$ ). This average is skewed, because of Henk and some of the other players who belong to the computer problems camp. The majority of participants (77%) played in between 4.5 and 6.5 hours.<sup>2</sup> I consider the slow players as those who played over 6.5 hours and this concerns 22% of the total group. The fast players (11%) are then those that played less than 4.5 hours. The fastest participant took about 3.5 hours to complete everything.

Adding the time of the meetings, which ranged from 2.5 to 3.5 hours, we can conclude that on average participants from Organization A spent 10–11 hours and those at Organizations B and C 12–13 hours. These numbers exclude the time participants invested on filling out the game questionnaire, surfing to the website, and looking at the manual, and so effectively I think they devoted between 13 and 15 hours to this training. That is almost two full workdays.

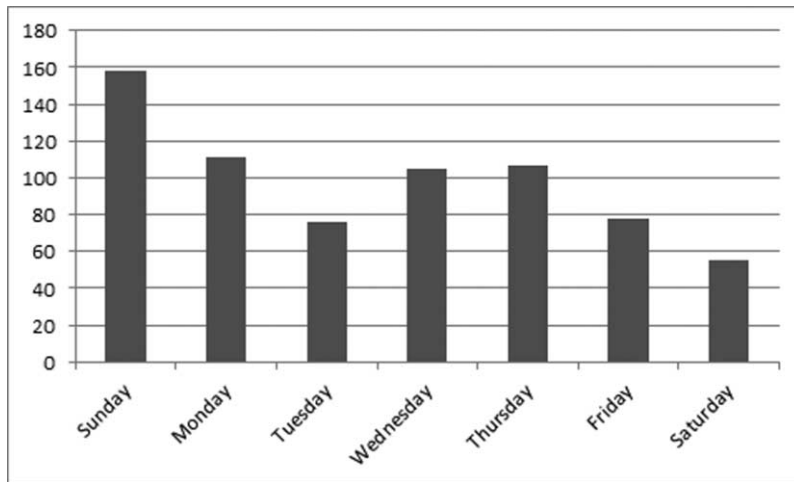
---

<sup>2</sup> This categorization is based on 64 participants, because I retrieved for this number of participants a complete dataset of time stamps. It is also based on comparing the results of the time categories on the two main learning outcomes and on their judgment; see Level 11.



### ***Weekly Assignments Were Useful***

The playing dates and times show that participants played on all days during the week and indeed on all times (Fig. 5.1 and Fig. 5.2). Clear peaks can be seen on Sundays and in between the times 7 and 11 p.m. About 45% of all exercises were played between those times. The time pattern does not change if the weekend days are excluded. During the weekend days participants played spread out over the day and during work weeks a number played throughout the day, probably because some were self-employed, retired, unemployed, had to work in the evenings, or had irregular working hours.

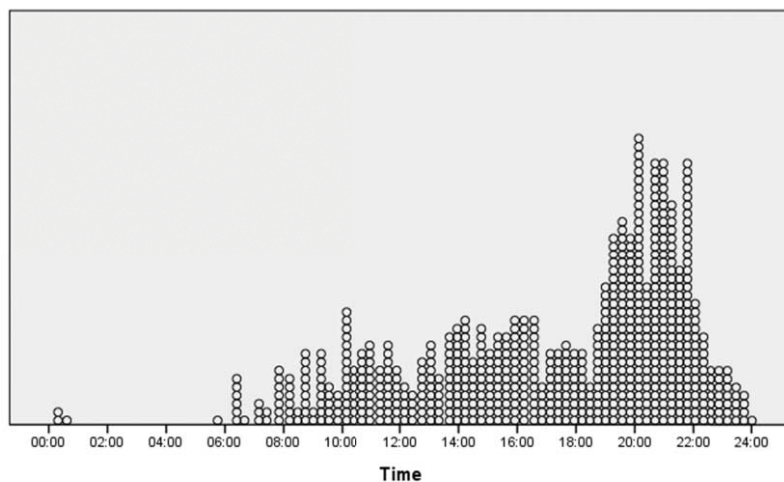


**Fig. 5.1** The number of exercises played over the days of the week. Sunday was clearly the preferred day to play the game

Although I encouraged participants to begin early with playing on the start-meetings, on average participants finished the first exercise just before the deadline of the first week and the second one day after. This confirms my impression that participants completed their exercises close to the deadlines.

The pattern did not change so much for the second week. The third exercise was finished on average one day before the deadline and the fourth on the deadline. However, the last two exercises were finished much ahead of time, about four and three days before the deadline, respectively. This might be an indication that after four exercises participants got into it. It could also be that because the training was about to finish, the sense of urgency among participants was much higher to finish it in time.

These results suggest that the implementation of weekly assignments was successful. They ensured participants played in phases. Although I cannot prove this, I am confident that without these assignments not many participants would have



**Fig. 5.2** A plot of the times at which all home exercises were played. This pattern does not change if weekend days are excluded. If only the weekend days are considered, the exercises are equally spread over the day

played all exercises. Participants would have started playing a week before the end-meeting and would get stuck, because the information and practice of the start-meeting faded away.

The participants were confident about this too. They confirmed the usefulness of the deadlines on the end-meeting and told me the weekly e-mails were a good reminder. Like anybody participants had to pick and choose between many activities. They would have postponed the training if they were not reminded about it.

The participants further told me that it would be much better to perform a training like this during the colder months. Throughout April and May in 2010 the weather in the Netherlands was unusually warm. In such circumstances, people often prefer to be doing an outside activity instead of sitting behind a computer.

### ***Distracting, Liberating and Relaxing***

Many indicated they found the game intense and told me one should not have any distraction and have a clear mind to play well. Among the reactions on the game questionnaire, I read that they should not play the game late in the evening, they should stop surfing the Internet while playing, and make sure they are not disturbed by private matters.

A couple of times I was not paying attention, because of a disruption in the real environment—  
GQex4-#108

The failures were recognizable, but I was sleeping. Too many things happened around me again—GQex4—#97

For a researcher, this is both positive and negative. It means that participants realize they need to create a laboratory type of setting at their home—one in which external influences do not affect their performance. On the other hand, these external influences apparently played a role. At the start-meeting and in the manual I did mention the importance of playing without any distractions (Level 3). Jokingly I always added that they were allowed to take a break now and then, to for example grab a cup of coffee or go to the restroom. Only if a flooding occurred in the real world, they were allowed to leave their computer alone.

Despite these instructions, I was pragmatic here too. For example, I know that at least two participants, both with computer problems, asked their sons for help (#14 and #16). Up front they asked me if this was considered cheating. Although this confounds their efforts, without the help of their sons they would have probably discontinued the training. In addition, the sons basically replaced me, because I offered assistance. I told them it was not an issue as long as they were playing and not their sons.

Another confounding factor for letting participants play at home concerned their computers. On some computers the game worked, but the graphics were portrayed differently. One participant remarked that he found the exercise “confusing because of a cloud of water” (GQex1—#61). When I replaced participants’ computers for a loan laptop, I observed this myself. The water really floated in the air. With others levee segments with a failure received a darker color, making it easier to detect failures.

It is too bad you can see failures from a distance—GQex1—#102

I have no guarantees, but I think that most participants with these issues alarmed me at some point and received a loan laptop. However, it does mean that for these participants their results on the first or first two exercises were confounded by their computer. I never let them replay those exercises. That would have resulted in confounded results too, because they would have known where the failures are located.

Almost always I asked at some point during the discussion how the participants experienced playing at home. Except for one or two, participants said unanimously it was “perfect.”

It is the advantage of making a choice when to devote your time to it. You do have a deadline of one week [to play two exercises], but within that week you can play from as early as six thirty in the morning till eleven thirty at night. It is how you like it...And two times I find just right to learn it. I am a digital illiterate. The only thing I use is [Microsoft] Word and nothing else. I know not any game and I never play them, but I experienced this as very positive—DB5—#1

Besides being able to choose when to play, participants indicated that they liked playing at home because it was more relaxing. They were not bound to any time limits and could look up information as much and as long as they wanted. At the meetings we frequently had to force people to stop playing. We were running out of time and had to go to the next agenda item.

## **Rating the Exercise Every Time**

The game questionnaire produced player ratings about the experience of playing an exercise. In this way, I was able to get a further understanding of how players experienced the exercises and if this changed over time. I decided to make the questionnaire an integral part of the game. The questionnaire appeared immediately after the feedback screen and was made in a style consistent with the game interface. Because I expected that players would not like to fill out the questionnaire every time, they were not able to exit it unless they forced the game to shut down by for example turning off their computer or using the Windows Task Manager. I also made it impossible to skip any questionnaire item.

Although these decisions likely increased the response rate compared to alternatives, it has various drawbacks. For example, I planned on using 7-points items similar to the pre- and post-questionnaires, but the screen looked too cluttered with seven check boxes and resizing them was not that easy. This would have required a complete new interface design. For this reason, I had to switch to 5-points items and this still looked somewhat cluttered.

Another drawback example relates to the game's mouse cursor. This moves automatically, which made it difficult to click on the check boxes and, therefore, made filling out the questionnaire a frustrating experience. Many uttered their frustration with the moving cursor, something I will expound on later in this level. Both drawbacks highlight that the game environment was not developed for filling out questionnaires and the ratings may have been influenced by them.

Much more unfortunate is that I only retrieved a complete dataset for 55 participants (Level 4). A dataset is complete when I retrieved the questionnaires of all eight exercises (i.e., start- and end-exercise and Exercise 1 to 6). Some of the missing data is Missing Not at Random (MNAR), because a number of participants did not play certain exercises or quit the training. The remaining part is Missing Completely At Random (MCAR). These files were not retrieved due to Internet and server issues. For my purposes here, which is to explore patterns in the data and not necessarily to draw strict conclusions, I focused on the complete datasets (listwise deletion) and compared its results with all of the retrieved results on a single item (variablewise deletion). This comparison allowed me to see if the group of participants with a complete dataset is representative of the sample population.

The questionnaire's 15 closed items were divided into three equivalent parts based on the worlds of triadic game design. The first five closed items were based on the world of Play, the second on Reality, and the third on Meaning. I will discuss the rating results per world and conclude by looking into the overall ratings.

## ***An Impression of Playing the Game***

The first set of five items relate to the world of Play. I constructed these items with the following three criteria associated with this world in mind:

**Table 5.1** The closed items of the game questionnaire

Item	Answer	Criterion
<b>Play</b>		
1. I found the exercise:	Fun	Fun
2. I found the exercise:	Difficult	Engagement & fun
3. I found the exercise:	Stressful	Engagement
4. During the exercise I was [answer] of the time:	Conscious	Immersion & engagement
5. During the exercise I had a [answer] feeling of being in a virtual environment:	Strong	Immersion
<b>Reality</b>		
6. I found the failures:	Recognizable	Fidelity
7. I found the failures:	Realistic	Fidelity
8. I found the failures:	Complete	Structural validity
9. I found reporting the failures:	Logical	Process validity
10. The computer judged me:	Correct	Structural validity
<b>Meaning</b>		
11. After this exercise I am better able to recognize failures.	Agree/disagree	Observing
12. After this exercise I am better able to determine the failure mechanism of a failure.	Agree/disagree	Diagnosing
13. After this exercise I am better able to assess the severity of a failure.	Agree/disagree	Assessing
14. After this exercise I am better in reporting.	Agree/disagree	Reporting
15. After this exercise I have a better idea of what measures I can take.	Agree/disagree	Taking measures

***Fun*** A positive emotion as a result of playing a game. According to Lazzaro (2008) this positive emotion could come from tackling a difficult challenge (“hard fun”), the sheer interaction with the imagination of a fictional world (“easy fun”), interacting with other people (“people fun”), or creating something of value outside of the game (“serious fun”).

***Engagement*** The involvement and commitment in playing a game. Involvement relates to the extent players are absorbed by the game while playing (for an overview of involvement types see Calleja, 2011); commitment is about the extent players are thinking about the game while not playing and/or are willing to play the game anytime soon again.

***Immersion*** The feeling of being somewhere else when one is physically situated in another (Murray, 1997). Players suspend disbelief and actively create belief about being somewhere else.

Although in game research literature these three criteria are heavily discussed and debated, all are predominantly mentioned in the process of describing how players experience a game. The items constructed with these criteria in mind asked participants to choose a phrase with a certain concept that provided the best fit of their

perceived experience (Table 5.1). Each item was formulated according to the following ordinal structure:

1. Not ... at all
2. Not ...
3. Moderately ....
4. ...
5. Very ...

For example, for the first item the concept was “fun” and the answers were: not fun at all, not fun, moderately fun, fun, or very fun. The second and third item are primarily related to engagement and in defining these I used the theory of flow (Csikszentmihalyi, 1991). A flow is a state in which people keep on continuing their activity. Being in a flow is another way of saying someone is “engaged.” The activity should not be too easy or too hard to get into a flow. If it is too easy, it becomes boring; if it is too hard, it becomes frustrating. To assess both possibilities, I asked participants to give their opinion on whether they found the experience *difficult* and *stressful*. Finding something difficult does not necessarily lead to a stressful experience. In fact, human beings crave difficult challenges. It is one of the reasons we find playing games fun (Koster, 2005).

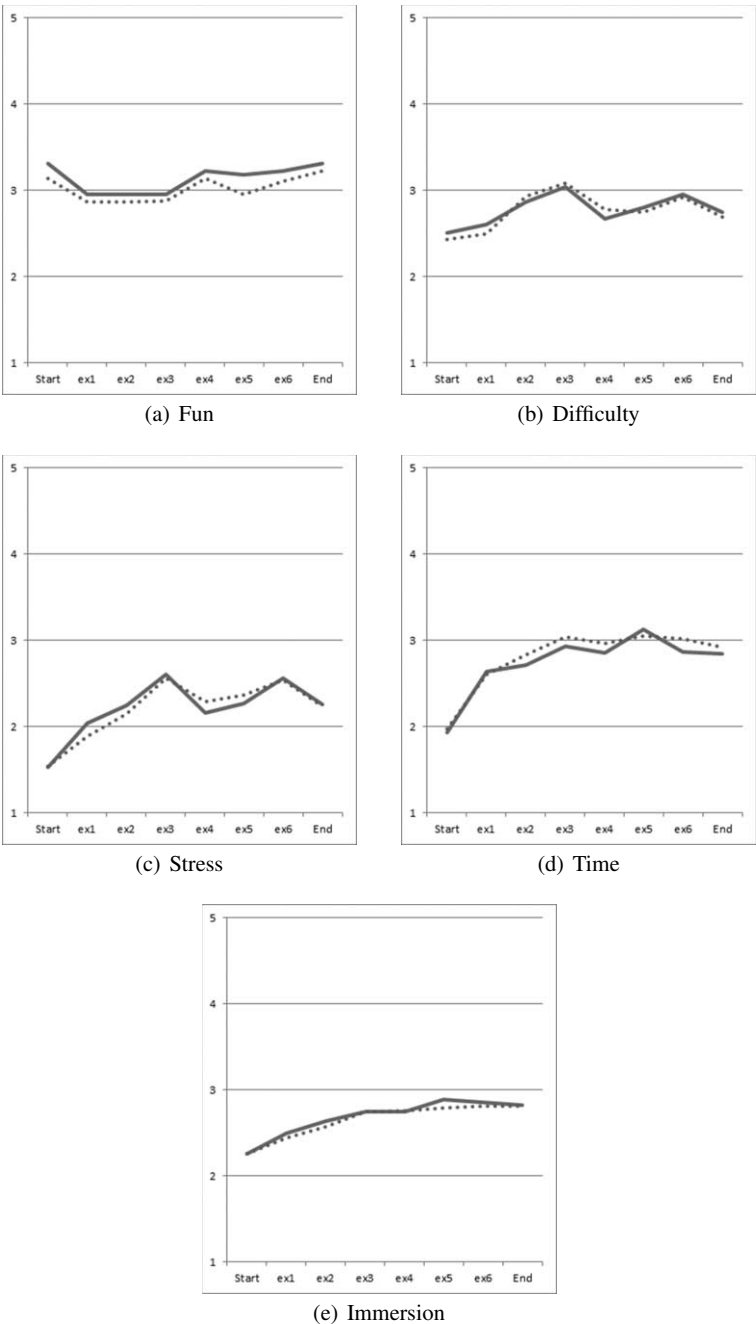
For immersion, I asked whether participants had the feeling of being somewhere else. This may be difficult to answer and so I added whether they were aware of the time while playing. Not having a sense of time is often mentioned as a characteristic of being immersed. It is also frequently associated with being engaged.

Based on the average ratings we see that participants experienced the game over all exercises as moderately fun (Fig. 5.3). If we look carefully it becomes clear that for Exercise 1 to 3 players found it somewhat less enjoyable. I attribute this to the difficulty they experienced while playing. At first, with the start-exercise, they may find it enjoyable because it is new. Then they realize that the game requires quite some effort and they become frustrated because of it. After understanding how to play the game, it becomes more fun. This understanding happened at Exercise 4.

This idea is confirmed when we look at the difficulty data. It is interesting that participants did not think it was so difficult at first. I expected that they would have been overwhelmed, but it seems they underestimated the difficulty (or they received excellent help during the training). The peak at Exercise 3 may be explained by a technical problem, which made it difficult to mark one of the failures. Another explanation is the difficulty of finding failures in this particular region.

Stress follows the pattern of difficulty almost exactly. They only differ in intensity. On average people found it more difficult and less stressful. With both we see a second peak at Exercise 6. In this exercise people enter a new region with many failures and this may explain the increases.

Time and immersion share a similar pattern too. Both elements become increasingly stronger over the course of the exercises. One would however expect that with an increase in immersion, a decrease in time awareness would occur. Being somewhere else makes one forget about the time. What possibly happened is that participants became more conscious of the time in the game rather than the time in the



**Fig. 5.3** The items pertaining to the world of Play. The dotted line represents the average of all responses; the solid line the average after listwise deletion

physical world. Time plays an essential role in the game. If players are not quick enough, a levee breach may occur. In addition, time in the game is very structured. The first failure phase lasts about ten minutes and the other phases about five minutes each. Playing the game repeatedly makes one aware of this time schedule.

An explanation for becoming more immersed over the exercises is that participants may have had to get used to playing the game. They had to know how to interpret the fictional world and learn the controls before they could get immersed. Having to cope with all the troubles of playing the game, I would surmise that players could not allow themselves to feel immersed.

### *An Impression of the Link with Reality*

The second part was based on the world of Reality. With this part I was curious to know how players perceived their experience with regards to the physical world. Here too I was interested in knowing if these perceptions changed over time. This time I used two criteria:

*Fidelity* Concerns the level of realism presented to the player (Feinstein & Cannon, 2001). It measures the degree to what the game is similar to the physical world.

*Validity* The extent that investigation of the model behind a game provides the same outcomes as would investigation in the real world (Peters et al., 1998). A distinction can be made into *structural* and *process validity* (Raser, 1969). The first looks into isomorphism in structure (e.g., the theory and assumptions on which it is built) and the second into isomorphism in processes (e.g., information flows or procedures).

Fidelity is more concerned with the look and feel of a game—how real it is perceived. So how failures look has to do with fidelity. Mud needs to look like mud and grass like grass. Validity is less focused on appearance. It is oriented at the model behind a game. It considers what factors are included, how these factors relate to each other, and how they change over time. Validity examples are what failure elements are included and how the failures develop over time. We excluded some signals and elements, because they were hard to implement. This exclusion makes the game less valid.

The Reality items use the same ordinal structure as the ones associated with the world of Play (Table 5.1). The first two items are about fidelity. The first item asks whether players are able to recognize failures; the second asks whether they find them realistic. The idea behind this set is two-fold. After encountering phenomena multiple times, it is arguably easier to recognize them and so one item was dedicated to measure if participants found failures more recognizable over time. If participants also find failures more realistic over time, it becomes more interesting. We know realism is subjective, but if this perception changes on the mere basis of getting accustomed to virtual objects, it gives us important food for thought in using virtual reality.



Second, virtual objects are representations and some of them are more realistic than others. Their primary purpose in the game is to be recognizable however. Players need to identify failures in the game and use this information to identify failures in the physical world. A not so realistic but recognizable representation may serve this need. For this reason I included both concepts.

For validity I chose two items related to structural validity and one to process validity. For assessing the structural validity I asked players to tell me if they found the failures complete or not. It might be that important elements are missing. I also asked if they were judged accordingly by the game system. If they disagreed with this, it most likely means that participants disagreed with the model of reality behind the game. Regarding process validity I was concerned about the logic of reporting. Although responses on this item are likely influenced by how it is visualized, what steps they had to take and in what order influences this too. Players will notice this if these do not correspond to what happens in reality.

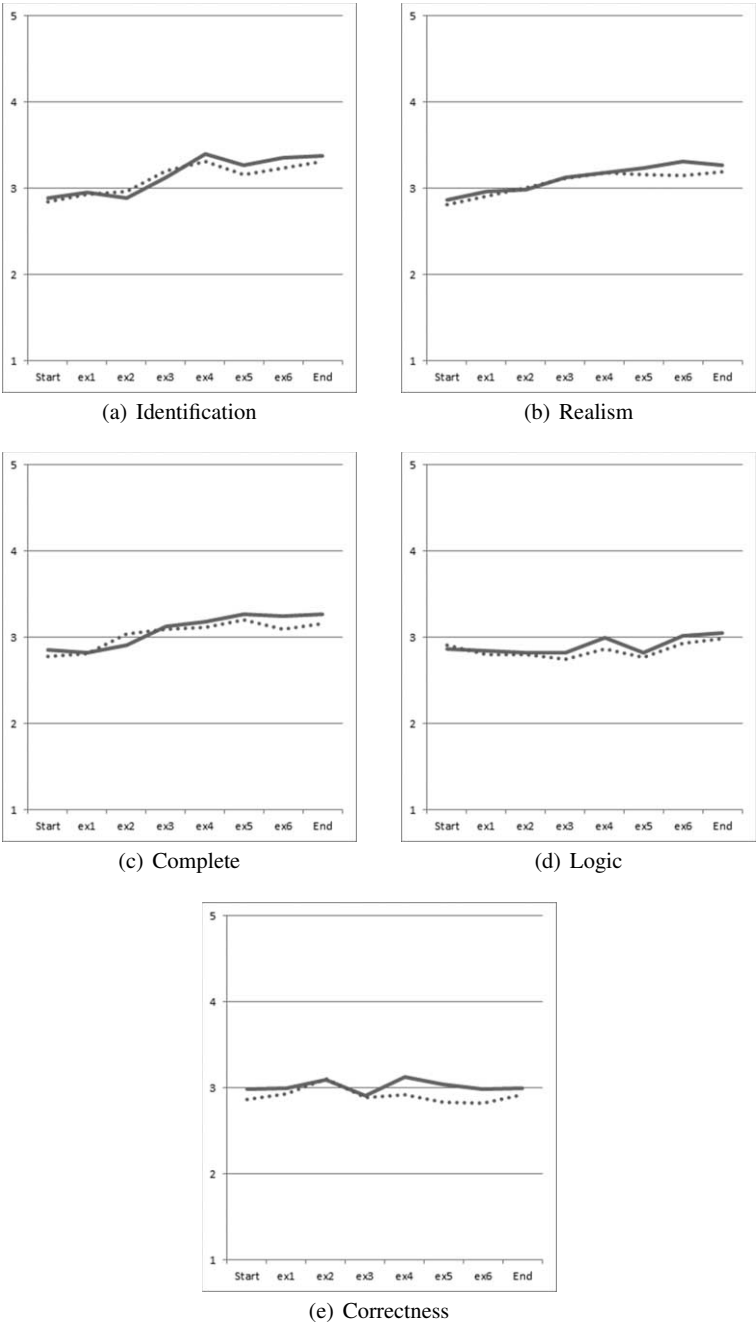
When we consider the ratings, we see a pattern similar to immersion (Fig. 5.4). Over time participants recognize failures better and find them also more realistic. Not only does this confirm the idea that players have to find their way in the virtual world, it shows that important “objective” qualities are dependent on this. We expect something like realism to be consistent, yet the data shows that this is contingent upon the ability to “read” the virtual world. Based on this, one can hypothesize that over time people get accustomed to the virtual world and this becomes their point of reference. This possibility may more likely occur if people are not able to get a point of reference in the physical world, such as with rare failures.

The pattern of the completeness of failures is similar too. This indicates that another “objective” concept like validity is subject to perception as well. This does not hold true for how the failures are reported. The logic of reporting remained consistent over time. One could guess the reverse should have actually happened. Reporting frequently increases the understanding of its process and makes it more logical. One explanation that this did not happen could be the sheer difficulty participants experienced with reporting. Until the very end participants commented on this.

With respect to correctness, responses remained consistent as well. This means that participants thought they were being judged by the game somewhat accordingly over all exercises. So despite being able to read and understand the virtual world better over time, this did not lead to a higher appreciation for why they were being judged in a certain way. Several reasons may account for this. One is that participants did not always agree with how failures should be reported according to the game (see the prevalent gameplay responses).

### *An Impression of Learning from the Game*

The third and final part of the game questionnaire looked into whether participants learned from their game experience. This part relates to the purpose of *Levee Pa-*



**Fig. 5.4** The items pertaining to the world of Reality. The dotted line represents the average of all responses; the solid line the average after listwise deletion

*troller* and touches on the world of Meaning. Whereas with Play I considered three criteria and with Reality two, I considered one with Meaning. I only looked into *relevance*: whether the game achieves its objectives. For *Levee Patroller* I defined five learning objectives: assessing, observing, reporting, diagnosing, and taking measures<sup>3</sup> (Level 2). For each one I included one item (Table 5.1).

The idea behind these items was to see how the learning experience evolved over time. I more or less intuitively decided on eight exercises. Maybe five or six exercises would have been sufficient. By considering the rating patterns we would be able to get an impression about this. To remain consistent with the pre- and post-questionnaire items (Level 7), these items were formulated with a disagree/agree structure:

1. Strongly disagree
2. Disagree
3. Undecided
4. Agree
5. Strongly agree

Except for reporting, the patterns are strikingly similar—among each other and with some of the previous items (Fig. 5.5). We see a small incremental improvement over time, with a steep increment from the third exercise to the fourth. This steep increment might suggest that after the fourth exercise participants perceived themselves to have learned. Reporting may not follow the pattern, because participants achieved rather poor results on this objective in the game (Level 6). This may have led participants to believe that they have not learned that much about it.

The patterns show that learning happened gradually. Although a peak is noticeable, the results tell us also that every exercise was valuable. However, I found plenty of comments from participants who said they did not learn anything new anymore. These comments were made especially after the last two exercises. This indicates that for some, the learning curve came to an end. Others indicated even after playing the end-exercise that they needed to practice more.

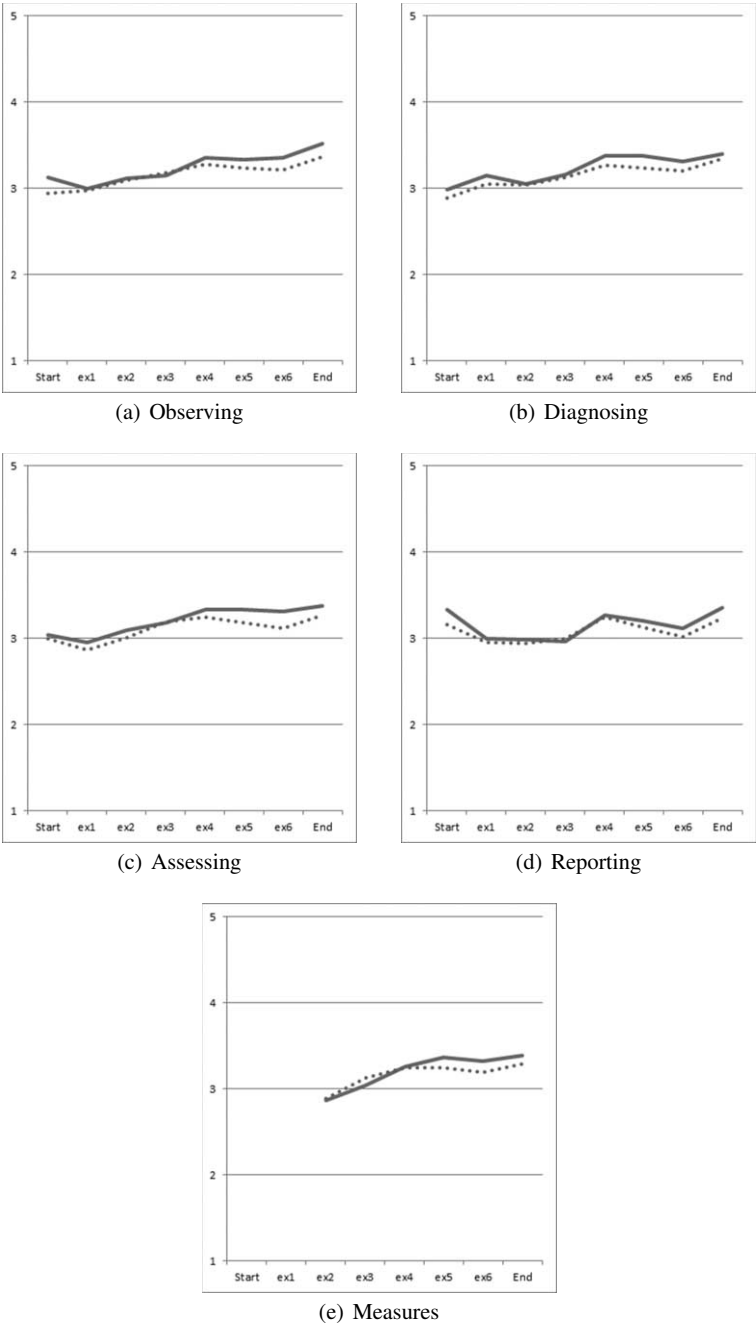
Knowing that the rating three stands for “undecided,” we could reason that on average players hardly perceived to have learned anything during the first four exercises. This indicates that practice was needed. This indicates that practice was needed. Two participants already held this sentiment after the first exercises:

You first have to know the game well before you are able to deal with the failures—GQstart-#137  
 Have to understand the game to work effectively with it. Right now that is not the case—GQex1-#75

A third participant criticized the game after Exercise 2, because in his opinion the game has an imbalance regarding the time involved with the learning objectives:

Something I realize now: learning the “game rules” and the game controls require more time than what the exercise in this game is actually about—GQex2-#143

<sup>3</sup> For the item on taking measures the complete dataset is larger ( $N = 69$ ), because this statement is not asked at the end of the first two exercises.



**Fig. 5.5** The items pertaining to the world of Meaning. The dotted line represents the average of all responses; the solid line the average after listwise deletion

Participant #143 has a point. During the first exercises, many participants are more preoccupied learning to play the game than learning from it. This is because they have to learn how to play the game before they can learn from it. Once they get it, they can focus their attention on the learning objectives.

## *An Impression of the Overall Ratings*

Many participants did not understand why they had to rate these items every time. According to them, they filled out the questionnaires the same way, but differences exist between items, in rating and in how they change over time. Generally the data point to a pattern of improvement. Generally we see a pattern of improvement. This general pattern shows that people have to get used to the game—they need to “read” it. After that they start to enjoy it, find it realistic, and learn from it.

We have to be careful in drawing any firm conclusions about this. The changes are subtle and are based on the data of merely 55 participants. That is why I have been speaking about *impressions*. To get a firmer idea, I first decided to summarize the ratings of each item and see how they relate to each other. I did this because most item ratings follow a similar pattern. Data reduction seemed therefore a good strategy. I further decided to focus this time on the exercises played at home (Exercises 1 to 6). The meetings offer a different setting and this may have influenced the results. Another reason is that with this focus I was able to increase the number of complete datasets from 55 to 60.

Unfortunately, even this number represents the bare minimum for a *principal components analysis* (PCA), which is a procedure to look into data reduction possibilities. A solution exists however. The results with the 55 participants seem to represent the complete sample population. The average of all responses follows the averages after listwise deletion almost exactly for every item. Based on this, *data imputation* becomes a possibility. This is a procedure to estimate values for missing data.

### **In-depth explanation: permuting the data and the PCA**

Data imputation involves dealing with missing data by estimating a value. Because most data were Missing Completely At Random (MCAR), imputation was acceptable. I did a multiple imputation with five iterations and with an automatic imputation method. All game questionnaire variables were used as predicted and predictor variables.

A PCA was conducted on the original data and the five imputations with orthogonal rotation (varimax). This led to low loadings and extracted communalities among a number of variables and inconsistent results between the five imputations. I suspected two variables to be the cause of this: the logic (Item 9) and the correctness variables (Item 10). Both had an extracted communality lower than .70 and this made them a candidate for deletion. After their deletion the results became stronger and more consistent. The results after deletion are shown in Figure 5.2. Both deleted variables loaded consistently on the first component (C1) and so their deletion does not change any of the conclusions.

The PCA with the original data led to undesirable results as expected. Although some of the indicators are acceptable, many extracted communalities were far below the acceptable criterion of .70. Only one variable, the time variable (Item 4), was quite below this criterion (.577), but deletion of this variable led to even less desirable results. Three out of five imputations gave the same result as shown in Figure 5.2; the other two imputations were similar to what was achieved with the original dataset. This confirms at least the existence of Component 1 (C1).

All KMO values for the imputed solution were  $> .69$  and the solution explained 75% of the variance. Kaiser's criterion was used to decide on the number of components.

**Table 5.2** Principal components analysis (with varimax rotation) of the original data ( $N = 60$ ) of the game questionnaire and the permuted data ( $N = 125$ )

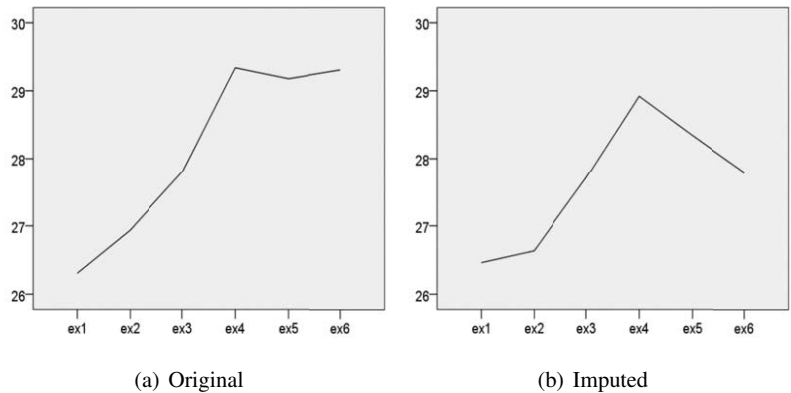
Item	Original data		Imputed data		
	C1	C2	C1	C2	C3
1. Fun	.828		.788		
2. Difficulty		.802		.826	
3. Stress		.818		.864	
4. Time		.693			.646
5. Immersion		.407			.846
6. Identification	.890		.839		
7. Realism	.811		.781		
8. Complete	.874		.825		
11. Observing	.907		.874		
12. Diagnosing	.934		.899		
13. Assessing	.929		.905		
14. Reporting	.834		.838		
15. Taking measures	.841		.840		
Explained variance, %	56	14	54	12	8
Cronbach's alpha	.977	.873	.977	.840	.841
Kaiser-Meyer-Olkin		.873		.914	
Bartlett's Test		$< .001$		$< .001$	

*Note.* Participants had to have played at least four exercises to be included. This rule was implemented to prevent participants entering the analysis with much missing data.

From the analysis with the permuted dataset, three components were derived (Table 5.2). The first (C1) I consider the *appraisal component* related to how valuable an exercise was to the player—how players assessed an exercise in terms of enjoyment, realism, and learning. The second (C2) is about engagement. It involves the two items that were primarily based on this criterion. The third (C3) component, on its turn, is primarily based on the items that were constructed with the criterion of immersion in mind.

The second and third components are not very reliable. They only consist of two items each and including the original dataset, they belong to one and the same com-

ponent. For this reason, I decided to neglect both in order to derive an impression of the overall results. What we do learn from this is that engagement and immersion do not strongly tie-in to perceptions about fun, realism, and learning—at least with this game. Although their ratings increased over time too, they do not relate to the others. One explanation is that the associated items did not directly assess the worth or use of playing an exercise. They merely assessed player’s state of mind when playing.



**Fig. 5.6** The averages of the appraisal component over time according to the original and complete dataset and the imputed dataset (5th iteration). Notice how the appraisal declines with the imputed dataset. With the original it flattens

My second step in investigating the overall rating was to see how this appraisal component changes over time. To find out I summed all the items that make up this component per exercise. Figure 5.6 shows the results for the original and the imputed dataset. An one-way repeated-measures ANOVA confirms that the appraisal changed  $F(4.51, 559) = 6.25, p < .001, \omega < .001$ .<sup>4</sup> Investigation with (simple) contrasts reveals that the last three exercises differ from the first exercise, but with a decrease in effect size,  $r_{1-4} = .36; r_{1-5} = .30; r_{1-6} = .19$ . With the original dataset the results were similar, except that the effect sizes were somewhat stronger and especially between the first and sixth exercise.

These overall ratings provide stronger evidence that the ratings changed over time. They also give further insight into how participants experienced the game. With the first two exercises, the appraisal hardly changed and from there it increased. After the fourth exercise it seems to flatten (original dataset) or decrease (imputed dataset). This tells us that participants started learning from the game after the sec-

<sup>4</sup> The one-way repeated-measures ANOVA had to be corrected with the Greenhouse-Geisser estimates of sphericity ( $\epsilon = .90$ ), because Mauchly’s test indicated that the assumption of sphericity is violated.

and exercise at home and that it became more of a routine after the fourth exercise. Some of the comments confirm this analysis:

I answered everything with “neutral” because I did not learn anything anymore—GQex6-#120

The game becomes predictable after playing all previous levels...most I learned during the previous exercises—GQend-#9

## Retrieving the Prevalent Gameplay Responses

The game questionnaire contained five open items besides the 15 closed items. Three open items asked participants to elaborate on their ratings on the closed items associated with the worlds of Play, Reality, and Meaning, respectively, and were posed after each set of five closed items (“Please clarify briefly your above-mentioned answers”). The remaining two open items were posed at the end of the questionnaire. One asked participants to reflect on how they could improve their performance on their next exercise; the other simply asked participants to write down anything that was not discussed before. Except for this latter open item, they were forced to fill out at least five characters. I was otherwise afraid players would skip these questions.

This fear was ungrounded. My intuition that these participants would not comment much proved incorrect too. Rather, I was surprised by the depth of the open-ended responses. Over all exercises and open items, two out of three participants gave a meaningful response. Non-meaningful responses include “No idea,” “I commented on this earlier,” and “Seems clear to me.” Some participants even wrote essays about their experience, going beyond the given instructions.

I noticed that the comments changed over time, from being general (“It was fun!”) to more specific and detailed (“When I called the Action Center and informed them about a crack I observed, information was shown about soil but I did not report this”). Another change is that at first most comments were about dealing with playing the game (e.g., controls and interface). Then they changed to comments about their experience in the game (“I do not like the Action Center”). At the end participants offered their opinion or suggested improvements for the game. What did not change concerned the mentioning of two Dutch proverbs. One concerns the Dutch equivalent of “Practice makes perfect” (mentioned 32 times) and the other the Dutch equivalent of “Mistakes are often the best teachers” (mentioned eight times).

Some participants thought that they could directly contact me by providing information on the questionnaire. They posed problems they experienced and asked questions. Later they were surprised that they did not get an answer, such as our Henk:

Dear Casper, this is fun but I do not get an answer to my question—GQex6-#42

What struck me most is that the questionnaire seems to have worked as a catharsis for many. The responses on the open items are overwhelmingly filled with frustrations about not finding failures, communicating with the Action Center, and the



game controls. I found a number of comments that said something like “After this exercise I quit” or just simply “I quit.” They only never did. Their frustration must have peaked at that moment.

I detected many patterns among the meaningful responses by grouping them under different themes and then regrouping them until I found what I consider the *prevalent gameplay responses*. These are responses that more than one participant expressed over time and belong to a clear and relevant theme. I will discuss the ten most prevalent ones. Each gameplay response will enlighten us more about how participants experienced the game.

### ***I Am Not a Computer Person or Gamer***

Right at start of the training I noticed some participants had little computer skills (Level 4). Participants had such difficulty, because they do not like to work with computers. They consider themselves “an outdoor person, not a computer person” (GQex2-#16) or simply say they are not fond of computers. Some seem to even “hate” them:

These types of exercises do not appeal to me. If this is the future of levee inspection, I will consider to quit—GQex3-#101

Although many participants complained about the game’s difficulty, only a few made negative remarks, such as the one by Participant #101. I was somewhat surprised by this negativity, because participants decided to participate voluntarily (except for those affiliated with Organization C) and the invitation clearly stated that the training involved sitting behind a computer. These participants may have had different expectations about the training and felt responsible or committed to participate. One participant expressed a valid objection:

Program is so difficult to control that it misses its purpose—GQex2-#74

On the one hand, few expressed specific troubles with the game. On the other hand, nearly everybody expressed having difficulties. This suggests the game may indeed be too difficult, especially for those with minimal gaming experience. Therefore, one can question whether the game achieves its purpose for this target group.

In addition to the dedicatedly computer illiterate, the sample population seems to consist of people who have computer skills but simply do not like playing games. Playing this game makes them realize this:

I am not a gamer and I will not ever become one—GQex4-#146

I surmise that certain game mechanisms did not appeal to him. Another participant confirmed this idea:

The virtual program works. I really have the idea I am being there. However, I am not a computer game person, so I am missing the drive to win—GQex2-#49  
I have the idea that due to the game format a competition element is established and because of this you will lose the learning process out of sight—GQex3-#49

Participant #49 also pointed out that the game format might be a hindrance to learning. It takes away the focus on the learning material according to him. For others not being a gamer is a hindrance:

People with game experience have an advantage against those who never play games—  
GQex1-#34

These testimonials highlight the challenge of using games for educational or job training purposes among those with a general antipathy toward computers or computer games.

### *Too Difficult or Too Easy?*

Another important issue is how participants experienced the pace of the game. Basically, participants experienced the game in two opposite ways. One group needed more time and found it exciting:

I had the feeling that I was continuously short of time—GQex2-#64  
It is quite exciting to play these exercises—GQex3-#112

I was surprised to read some were short of time, because they took on average an hour to play one exercise. I think that some required much time to execute certain game actions, such as clicking on a menu item, which made it a frustrating in addition to an exciting experience. Frustration is what the other group experienced as well. They found it boring and wanted the game to move at a faster pace:

The exercises are somewhat boring and I do not get the impression that they contribute to my knowledge about levee inspection—GQstart-#115  
Because you know from your statistics that only two failures exist and I found these pretty quickly, it took a long time before the game proceeded. You know that you do not have to look further and you have to wait till that one failure changes—GQex1-#53

The participants in this group were young and had excellent computer skills. They were able to finish parts of the game ahead of time. When reflecting upon their performance these participants said the reverse. They indicated that they should take more time in reporting the failures. One telling comment was “first think, then call” (GQex5-#51). Although they might have needed more time, these participants were clearly not challenged enough.

The game is standardized for everyone. The experiences just described are the consequences of making it a one size fits all. If it is too easy, people will find it boring (Csikszentmihalyi, 1991). If it is too hard, people will find it frustrating. For research purposes, it might be preferable to keep the game standardized, but for an optimal training it would be better to adjust the exercises to the player’s performance in real-time<sup>5</sup> or by making players choose a difficulty level beforehand.

---

<sup>5</sup> One easy option for real-time adjustment for *Levee Patroller* could be to speed up the time if all failures are found.

The participant experiences with the game changed over time. A part of the slow group was able to improve their performance, stating, among other things, “It goes better and better every time.” Because over time the exercises increased in difficulty, the game became more interesting to those who found it initially too easy.

The best example of a change concerns a young man named Johan (Participant #69). He was 18 years old, “an all-around carpenter,” played games in his free time, and participated together with his father (Participant #63) in this training. At the start-meeting I noticed that father and son were not too excited about the game, and the father continuously complained about it. This is a (censored) summary of what Johan wrote during the exercises at home:<sup>6</sup>

I did not like this at all (GQex1)...this one was more fun than the previous (GQex2)...this was not any fun (GQex3)...stupid, annoying...I just do not like it all (GQex5)...do not like it...it is boring...very dull (GQex6)—#69

However, in the end they expressed positive feelings about the game. Johan’s father changed because he was improving in it. Sjaak changed because he was challenged more in later exercises.

Not everybody changed. One such person is Pieter (Participant #67), 63 years old and retired. He clearly had some trouble in playing the game. On the end-meeting I was surprised to see that after playing *seven exercises*, Pieter still was not able to navigate through the world. Because most participants were able to play the game without any help, I was able to devote much time to him. Despite my efforts, he simply could not learn the basic ways of navigating. I felt even more sorry after reading his responses on his last home exercise:

It remains hopeless...it stays a problem...it is a waste of your time to have a participant like me—GQex6—#67

The story of Johan and his father and of the unfortunate Pieter have something in common. They show that these people have character. If they commit themselves to something, they go for it.

### ***Being Judged Can Be Confronting***

Many took the game very seriously, so much so that some broke out in sweat when a levee breach occurred. I observed this during the end-meetings. I also saw participants entering the end-meeting embarrassed and telling me they were unable to prevent a number of floods. These are responsible, committed, prideful people who take their (voluntary) job very seriously, even within a game environment. For such people the direct feedback in a game can be confronting, especially if it is about real world aspects, such as the knowledge one has about a subject. Two participants acknowledged this:

---

<sup>6</sup> The results of the game questionnaire of Exercise 4 are not included, because I did not receive this file.

This should be a wake-up call to me—GQex1-#43

Personally I found the exercise quite difficult, because I thought that I would know much more—GQex2-#95

Others said the game made them feel insecure (e.g., GQex1-#40) or made them anxious. They were afraid failures would become worse (e.g., GQstart-#22) or a levee breach would occur (e.g., GQex2-#145). Not everybody had these issues. One commented that he experienced a “fun flooding” (GQex3-#19). Another said it was “a pleasure to look at the animation of it” (GQex4-#93).

I surmise that some of the issues people had with being judged relate to their view on learning and the use of the game. I base this on the many suggestions I received for how to play the game. Some would have liked to have had a handbook with all the answers. While playing they could consult this handbook and provide the right answers. Others would have preferred to first learn from the books and then apply this in the game. For these people the game is more like a final test environment and one in which they do should not fail. They do not see the game as a safe environment where they can learn from their mistakes.

Game-based learning is, however, based on the ideas of trial-and-error, discovery-based learning, and other ideas that involve first trying out something, seeing what happens, and then learning from this. Learning from mistakes is the dynamic that drives continued improvement.

### ***I Really Want to Find Those Failures!***

In terms of performance, one major issue participants felt troubled over concerned their primary task: to find failures. Participants were fairly skilled at this (Level 6). At the very end almost everybody found every failure. Some people were less successful, producing such statements as:

Because I could not find the third failure I felt like I was participating in *Investigation Requested* [i.e., a Dutch TV show in which crime scenes are reconstructed and the help of the public is requested to find the suspects]—GQex3-#40

It is like finding a needle in a haystack—GQex5-#38

In fact, when participants were unable to find a failure, they made this very clear. It seems that not finding all failures severely affected the game experience. One person called the exercise “not fun” because of it (GQex2-#49). This is understandable. Not only concerns it the primary task of patrollers, the scores in the game are dependent on finding the failures and so is the subsequent learning about the experience. As one participant sufficiently put it:

Found nothing, learned nothing—GQex2-#113

Finding failures is part of the game, something that makes it challenging, but some participants make a valid point in suggesting that it might be better to provide (after a time) a hint about where a failure is located.

They may also feel less stressed, because as one participant said, it takes much time before all failures are found, which makes playing the game stressful (GQex4-#74). To reduce this stress, some participants might have changed their strategy in playing the game. They decided to first find all failures and then start reporting them. This is however a poor strategy. The game is largely paused when reporting, which means players do not lose much time by reporting them immediately. If they report after finding all failures one or more failures probably changed their state, which means players cannot get any points for the earlier states anymore.

From the responses, it became clear with what failure participants had the most problems:

I did not find the sand boils. Was it well hidden or was it not there?—GQend-#114

This sand boils is hard to find, especially if it rains. When the rain hits the water, it looks somewhat similar to the signal associated with this failure.

### ***That !@\$# Walking Mouse and Other Annoyances***

Regarding frustration, the leading role was taken by a minor bug. Sooner or later almost every participant commented on the mouse pointer that automatically crawled over the screen. It starts to crawl upwards when users do not move it. As developers we knew about this problem, but we could not fix it. It is a problem in the game engine software and we did not have access to this code. These are some of the many comments:

Again, because of the instable mouse pointer it is very unpleasant to play this—GQex1-#72

The “drifting” cursor ruins a large part of the pleasure—GQex1-#135

The mouse pointer continuously floats away. That is really annoying with for example pointing out the location—GQex1-#135

Can you do something about that mouse that runs off all the time?—GQex3-#11

Everything needs to be perfect. In previous test sessions, participants commented on elements they found annoying, such as missing revetment (Harteveld, 2011) or the mere fact that sheep do not walk around in the rain. These details seem to matter to the players.

In this case the participants’ computer skills probably played a role too. It may sound far-fetched, but it could even relate to a deep-seated fear we humans have when we work with technology: that it starts to get a will of its own. Whatever caused its annoyance, this minor bug shows the importance of considering game interface design.

Participants expressed other annoyances as well. One was bothered by the rain (GQex2-#25). Another found it cumbersome he was not able to measure the width of the ditches easily (GQex4-#136). Some participants were even so annoyed that in answering the question of what they could improve, they wrote that they should feel less annoyed about the errors of the game (GQex3-#57 and GQex5-#43).

In terms of annoyance, a good second is most certainly the Action Center. I suspected this, based on previous experiences but also on for example this e-mail I received from one of the participants who just finished the first two exercises:

During the start-meeting the Action Center already bothered me. I do not think we will become friends—E-mail from Participant #49, 13-4-2010

Others said to not have “good contact” with the Action Center (GQex2-#40), were too much distracted by them (GQex6-#31), or felt that they had not much in common, making it difficult to communicate (GQend-#51). What I find interesting about all these comments is that players immediately attribute human-like aspects to this Action Center, which really is just a piece of code with IF-THEN-ELSE rules. They treat, talk about, and consider this Action Center if it were manned by actual human beings.

Participants were annoyed by the Action Center for another reason too—a reason prevalent enough that I wish to devote a separate section to it.

### ***I Am a Professional and You Should Treat Me Like That***

Some levee patrollers felt they were not being taken serious. In fact, they more or less felt offended by the game and most particularly by the Action Center. This is just a small sample of some of the frustrations to illustrate this:

Lifelike, only disappointing that among the answers I could not always choose to say that the situation is stable. Instead I could choose for “I just felt like having a chitchat.” Come on! These and other rude remarks by the Action Center do not contribute to the game’s effectiveness—GQex1-#51

You are asked to frequently return to the failures. If you do so, you can only choose between worse, critical, or having a chitchat. I am forced to choose the latter, because I cannot report that it is stable. Then you get to hear that you are keeping the Action Center off their work!—GQex2-#92

When everything is found and reported, make the game end if possible. Then I do not have to get that stupid response by the Action Center of why I am calling them for nothing. Throughout the game you have to keep an eye on things and keep on reporting, right?—GQex1-#23

I still got some issues with renewing a report. For some reason the new report does not reach the Action Center and they start being funny about this, while I am actually being very serious—GQex3-#131

Nobody is to blame for these conversations except for me. When we were designing the game, I decided to insert a bit of humor into the conversations. Otherwise these conversations would become somewhat boring I thought. I never expected (and also never experienced this before) that participants would take this so highly.

One explanation is that they are serious and they want to be taken seriously. An additional possible explanation is that they consider themselves professionals and want to be treated with respect.

Some participants had issues with the levee expert as well. This response about what to improve is telling:

Stay calm and do not let myself be influenced by this levee expert. His input is aggravating—  
GQex2-#49

This shows that the game appealed to people's emotions—albeit not always positive ones. What I also noticed from the responses is that participants had difficulty in understanding the roles of the Action Center and the levee expert.

### ***What Happens and Where Am I Now?***

Players got confused in various ways. To start with, many thought the Action Center saw what they were seeing. For this reason, players could not understand why they did not receive points for reports and assessments that were approved by the Action Center. However, the Action Center judges whether the reports and the assessment make a fit based on the information provided by the players. This does not mean this information is accurate. Players may just as well get approval by the Action Center to take measures for a failure that does not exist.

The confusion may arise from the mere fact that everything is displayed on the same screen. Participants were unable to realize that although the Action Center appears on the screen where the failure is shown, they did not see what they were seeing. The Action Center is actually located somewhere else, in an office, far away from where the player is located.

This misunderstanding of perspective may explain the confusion between the Action Center and the levee expert as well:

Answers by the Action Center and the levee expert are in disagreement. Every time the Action Center asks me to take measures, but whatever measure I take, the expert says they are all unnecessary—GQex6-#40

The difference between the Action Center and the levee expert is that the latter inspects the failure himself and then provides feedback. The Action Center never gets to see the actual failure. They base their feedback on the information the player provides them. So it might very well happen that based on the reports the Action Center agrees to take measures and continues to remind the player about taking these, while at the same time the levee expert says measures are not needed.

This confusion could imply that the game's interface is lacking critical information for players to understand the roles of the Action Center and the levee expert. It could also imply something which I previously hypothesized: that players have to read the virtual environment. They are like newborn babies who discover the world and make perceptual errors because they still need to grasp certain rules. The game of Peekaboo is a well-known example. Babies do not understand that if you move an object in front of your face, your face does not disappear. It is just hidden. Similar to babies, players have to learn the rules of the virtual environment.

After learning the rules, some participants got it, such as Participants #58 and #61 who wrote this after the end-exercise:

I determine that for receivers sitting behind a desk (Action Center members) it is not easy to judge what people see in the field. This is most certainly a point of interest for the future—GQend-#58

On the one hand you want to take measures prematurely and the Action Center stops this. On the other hand you may get permission but the expert stops this. A great experience to become aware of—GQex4-#61

Speaking of reading the virtual environment, one of the most mentioned issues—next to the moving mouse pointer—concerns navigating the game world. Players had enormous difficulty with orienting. Some said they were continuously lost (GQex1-#15) or felt they were running like a chicken with its head cut off (GQex3-#143). The difficulties did not only relate to the controls as this participant explains:

Walking circles in your area is more difficult than a levee segment as we know it in our practice. Recognizing where you are is tough....It would help if on the upper left or right side a small screen is visible that shows your location on the map—GQex1-#58

What Participant #58 refers to is that at Organization B (and also at C) patrollers are responsible for a specific levee segment. Such a levee segment is a couple of kilometers in length. They always need to inspect their own segment and so they know where they have to start walking and where to end. They referred to this as their *walking route*. In the game players do not get specific instructions of where to inspect or how to walk. A walking route is missing. The only information they receive is that they have to inspect the region that they are situated in.

Because the virtual environment was new to the players, they had to orient themselves in this environment besides learning its rules. They had to get to know the virtual regions and understand how the map works. Participants had difficulty with this, despite our efforts in placing unique objects within the environments (e.g., windmills, sluices, and towers) and visualizing their real-time location on the map. The difficulty is likely a combination of learning the controls, translating the virtual images to their real world meaning, and understanding the first-person perspective of playing this game. Frequent comments I heard and read is that players especially wanted to know what “north” and “south” was and what levee parts they inspected.

To aid them in orienting, various participants suggested what Participant #58 wrote: to place a mini-map directly on the main screen. They also wanted to see their real-time location when marking the failure’s location.<sup>7</sup> We did not implement all of this because we wanted players to make a strong effort to become aware of their location, which was an important requirement put forth by our clients. However, participants did not see the point of this awareness and saw it as an unnecessary learning experience.

---

<sup>7</sup> When players have to mark the failure’s location they see a map without their real-time location and have to place a red cross on this map. This red cross symbolizes the failure’s location. By using the map tool from their inventory, players can find out about their real-time location and use this information to determine the failure’s location. This makes the map tool a “cheat device,” but using it does increase player’s awareness about thinking of where failures are located.



I further noticed that some participants found the freedom of exploring a region unsettling. They would prefer instead to receive clear instructions, such as a walking route. This shows that a highly appreciated and aimed for quality in entertainment games—the freedom of choice and movement—is not desired by these types of players.

That said, the regions are huge. It takes approximately ten minutes to walk over all levees. Although we placed unique objects, I can understand that players lost track of what levees they inspected. Players took initiative to deal with this. They printed out the region maps and made notes of what they inspected.

### ***Reality Is Not Broken, It Is Much Better!***

Not having a specific walking route is a departure from reality. Clearly some participants had issues with this. This group I consider the *virtualphobes*—those who do not trust virtual reality and do not grasp its utility. Unlike McGonigal (2011), who argued that “reality is broken” and games are a way to fix it, they think reality is superior. They give comments such as:

I do not understand the value of this game. In the real world everything looks completely different. Or are we going to walk around with a laptop in the nearby future?—GQstart-#77  
I have nothing to improve. The virtual images will make me distort reality. I will learn the wrong references—GQex4-#125

The situation outside is much more real. I hope that this virtual exercise will not be confused with reality—GQex4-#38

Virtualphobes are not necessarily the same people who do not like to sit behind a computer. Among the virtualphobes are young people who know very well how to work with a computer. Participant #127 is, for example, 41 years old and one of the younger participants. The mere fact it is a virtual environment brings forward prejudices and disqualifications. To them virtual does not equal real. These are examples of prejudices and disqualifications:

If it rains and you walk over a levee, you always see puddles. Here you see one and that is immediately a failure—GQex1-#68

Changes are (sometimes) difficult to observe, for example if it concerns little or much soil outflow—GQex1-#60

Estimating the width of a crack is hard. Measuring in reality is easier—GQex1-#111

I have to look more around me; however, in the real world I think I do that already—GQex2-#22

As patroller I do not have to take measures. But in the game this is necessary...—GQex3-#112

Some participants are simply incorrect. Not every puddle is a failure. In fact, we put puddles in the game environment to see if players could distinguish a normal puddle from one that is related to a failure. Other participants seem to overvalue the real world; they *idealize* it. In reality, sizes remain a subjective matter of individual perception.

A number seem to think that acting in a virtual world has little to no relationship with reality, such as Participant #22. One participant at the start-meeting who told me he could easily find and identify failures. I observed his gameplay and saw that he only walked over the crest of the levees. He missed all failures, because in this exercise all failures were located on the slopes and not on the crest. Although it is debatable whether this virtual behavior translates to real world behavior, we can safely assume that this player would have looked at the slopes if he had known failures could occur there.

Finally, some participants did not understand why the game was used at all. The comment about walking with a laptop over the levee by Participant #77 illustrates this. Others did see this:

Some participants are plain wrong. Not every puddle is a failure. In fact, we put puddles in the game environment to see if players could distinguish a normal puddle from one that is related to a failure. Other participants seem to overvalue the real world; they idealize it. In reality it is equally hard to say if something is little or much. That is why we make agreements about what is little or much. Even if agreements are made, sizes remain subjective.

A number seem to think that acting in a virtual world has little to no relationship with reality, such as Participant #22. I remember one participant on the start-meeting who told me he could easily find and identify failures. I observed his gameplay and saw that he only walked over the crest of the levees. He missed all failures, because in this exercise all failures were located on the slopes and not on the crest. Although it is debatable whether this virtual behavior translates to real world behavior, we can safely assume that this player would have looked at the slopes if he had known failures could occur there.

Finally, some participants did not get why the game was used at all. The comment about walking with a laptop over the levee by Participant #77 illustrates this. Others did see this:

Very good for practicing. This way you are better prepared for the actual practice—GQstart—#80

The virtual environment is sometimes untruthful and not very realistic, but you do learn much from it—GQex4—#47

The comment by Participant #112 belongs to this category too. I know many patrollers do not take measures, but including this in the virtual procedure gives them overview of the complete process. Again, others did see this:

Taking measures does not belong to my tasks as levee patroller, but it is nice to get a better sense of the situation and what possible measures are needed—GQex2—#148

Nevertheless, these people make valid points. Care and prudence is needed when virtualizing real world practices, and as one participant succinctly noted “a gap will always exist between reality and a ‘game’” (GQex6—#119). For example, as another person noted “inevitably the failures are abstractions and not so diverse” (GQex5—#51). This is true. By translating the real world to the virtual world we had to depart from reality, sometimes inevitably, sometimes on purpose. It is just remarkable how quickly some people tend to disqualify the use of games or virtual reality. This

suggests they have a negative attitude toward virtualness. As stated earlier, those people I call virtualphobes.

Virtualphobia may be cured by getting used to the virtual environment. Most negative comments came from participants in early stages of the training. As one participant said:

Once you are busy, the feeling of being in a virtual environment decreases—GQex5—#66

### ***I Do Not Agree!***

Players freely criticized specific elements of the game:

Only after the situation became severe, actions were taken. That is way too late—GQex3—#100

Revetment was damaged. That is always severe, is it not? Why is it not possible to directly repair it?—GQex2—#61

Gap in pitching stone is critical if the water level becomes higher. Measures are necessary! Program does not allow this!—GQex4—#118

The levee expert is nuts. It is better to repair a small damage than a big one—GQex4—#11

We delayed taking measures in the game to allow players to see every phase of a failure. We preferred this learning experience over what may occur in reality. In addition, deciding on when to take measures is subjective. Nobody disagrees on taking measures in the critical stages and so it was also a safe bet to enable measure only in these phases.

Subjectivity plays a role in assessing the severity of failures too. This is the second critiqued game element:

Damage to pitching stone is as shown not severe—GQex2—#57

I notice that I am adjusting my “inspection technique” according to the knowledge I currently possess of the game. I do not assess the severity based on my opinion, but based on the expectations I have of the right answer in the game. This is a curious development in a training...—GQex6—#138

I am now assessing according to the Action Center. In a real world situation I would react differently—GQex6—#149

Multiple respondents echoed the words of Participant #57. They disagreed about how the game would want them to assess a certain situation. In later exercises I found comments such as by Participants #135 and #146. They still disagreed but decided to play along to get a higher score. This is a perverse situation and one that is hard to resolve, because even the experts disagree on the assessments (Level 10).

Another (related) criticism is that the “game is kind of rigid” (GQend—#116). The following comment elaborates on what the participants mean by the game’s rigidity or “black/white approach”:

It is problematic that the diversity of the daily practice cannot be brought back to the black-white world of the (colored) reality of this realism-based game. About the constructed drive-way discussion is possible, depending on the season, the water level, and whether or not

settlement took place due to the driveway (in reality you would see cracks in the asphalt...) et cetera—GQend-#138

The game does simplify and categorize the real world, by necessity. First, for many topics it is impossible to represent the diversity, variety, and ambiguity of the real world in a game. It would make the game far too complicated—to develop and to play. Second, clarity is needed in a game (i.e. when points are given.) No discussion is possible. Although we could try to lessen the rigidity, the game will always remain “rigid” compared to the real world.<sup>8</sup>

The expert employees had more problems with the rigidity of the game than others. Participant #138, who gave many elaborate and constructive criticisms, was also one of the few that criticized the look of the failures:

The failure with the horizontal movement was not realistic: the whole ditch was filled with soil, but the crest and slope did not show any deformations. It seems to me that the moved soil has to come from somewhere. In addition, the missing pitching stones were hard to see...But maybe this is realistic: in dark waters in the real world it is also hard to see. I think it would be better if at some stage the failure continues above the water. That way you still have a chance to act—GQex4-#138

One needs to have an alternative mental model to criticize and question certain game aspects. This indicates the need for an elaborate discussion of the relationship between the game and the real world and the use of alternative materials during a training.

## ***To Learn or Not To Learn***

A number of comments related to whether participants perceived themselves to have learned or not. One series of comments came from (expert) employees but also from some knowledgeable volunteers. Whereas many volunteers indicated at even the final exercises that they needed to continue practicing and have to look up more information, this is what the knowledgeable participants had to say when they had to elaborate on their learning:

I *am* an employee of the water authority—GQstart-#118, emphasis added

I am busy with this every single day—GQstart-#149

This exercise does not contribute to my knowledge about levee inspection—GQstart-#117

Because I do this as an experienced levee guy I do not learn that much from these simple images. THE knowledge was already there—GQex1-#138

After 20 years of experience there is not much to be learned anymore—GQex2-#139

These are common failures...in my eyes the failures do not contribute to my experience—GQex2-#119

---

<sup>8</sup> One example to lessen the rigidity is to allow for multiple answers. The current version of the game allows only one correct answer. This is too rigid, because failures could be reported in various correct ways.

These comments suggest that the game did not add any value for these knowledgeable participants. In fact, one employee said the game has actually an adverse effect:

The game causes you to doubt findings that are already good. That is not very handy—  
GQex5-#9

This particular group made also positive comments. Although the failure images were known to them, they appreciated the game's systematics and in particular its reporting procedure:

The failures are known to me. Reporting is, however, put into a certain structure and that forces you to think before you report something to the Action Center—GQex1-#119  
I think it is better to see the failures in the real world. You do get to know how to report and what measures are possible—GQend-#27

Even volunteers had this opinion. One stated that “reporting remains the strongest asset of this program” (GQex2-#116). This positive attitude surrounding reporting was hard to reconcile with the many negative comments I also received about it:

It is difficult to make reports—GQex1-#91  
Reporting stays illogical—GQex5-#40

Although the reporting procedure is useful to learn, it is difficult to execute in the game. This execution pertains to choosing what to report and to providing this input to the game. I also have to admit that the reporting system is complicated (or “tedious” according to the participants).

I do think that the (expert) employees may have underestimated how much they learned, even about the failures. One of the expert employees called Adrie (Participant #119) was one of the most outspoken people to say that this game is especially valuable for its reporting. After the end-exercise he repeats this, but indicates that he learned something about the failures too:

To me the big plus is that you learn to think things through and learn to report accurately. After a few failures you go and search for the image you received. You know now how a failure looks like—GQend-#119

This statement suggests that Adrie gained focus about what to look for, which goes beyond thinking things through and reporting accurately. It is about making sense of failures.

Whether these knowledgeable employees learned beyond reporting or not, in general the comments suggest that participants perceived to have learned. These are some of the comments participants gave:

The material becomes much clearer during such a virtual exercise than from a textbook—  
GQex1-#97  
By these means the theory is better learned than in any other way—GQex5-#47  
Good game. Before I started I knew much less about signals, failure mechanisms, and measures—GQend-#101

## Lessons Learned

In this level I explored how players experienced the game. It takes participants some time before they learn from the game and also start to appreciate it. They first need to learn to play and “read” the game. It otherwise does not explain why players find it more fun and realistic over time and also perceive to learn more. Some participants were conscious about this process.

Measures that attempted to measure the experience were strongly related and were reduced to one component, which I called the appraisal component. This component measures how valuable an exercise was to the player and this confirms the idea about a need to read the game. Not until after Exercise 2 the appraisal started increase and it decreased after Exercise 4, suggesting playing became more of a routine.

What furthermore becomes clear is that many have been frustrated throughout the game: frustrated about not finding failures, about talking to the computerized Action Center, and dealing with a mouse pointer that has a will of its own. From this we can learn that a game is a sensitive medium. Little things—think of that mouse pointer again—could disrupt or frustrate a player.

We also find that players engage into meta-cognitive thinking by relating the gameplay experience to the real world or their own standards. They do not take what is in the game for granted. However, I also find that some participants may offshoot into the other direction, because they seem to distrust anything that is virtual (i.e., virtual phobia).

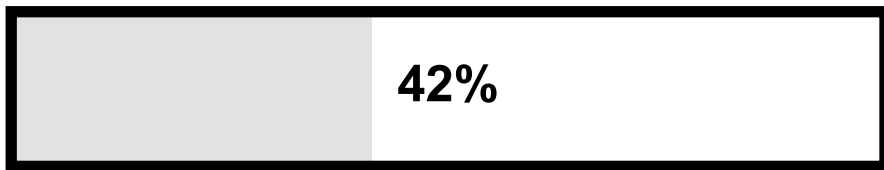
Although players seemed to have different play experiences, the majority played it at the same time. They played it on Sundays or on weekdays late in the evening. In total they spent about two full workdays on the training and on average one hour per exercise.

## Level 6

# Opening the Black-Box

*Garbage fell out of the trashcan. I put it back in—GDstart—#38*

*NO FLUSHING OF SOIL. DO YOU UNDERSTAND IT  
NOW?—GDex3—#74*



The advantage of digital games is that they are capable of tracking anything. With a simple text file with XML code we logged every possible action the player could take in the game (indicated as `<User_event>`). This is for example a code snippet from one of the player logs:

```
<RedMarker>
  <x>-1514.02</x>
  <y>-22118.84</y>
  <failure>LeveeTile15</failure>
  <Name>RedMarker1</Name>
</RedMarker>
```

The code shows that the player placed a reporting marker, which is also known as the “red marker” because of its red color. Certain characteristics are mentioned: the coordinates on the map and in what levee segment (or tile) the marker is placed. The levees in the game are divided into segments and in each segment a failure could possibly occur. If players place the marker somewhere in the identified segment with a failure, the computer is able to recognize that the player found a failure (see also one of the issues discussed in Level 4). For programming reasons the marker gets a name. In this way, other reported elements can be connected to the same failure, such as the failure’s location. Reporting the location is what the player did right after placing the marker, see this next code:

```
<User_Event>
  <Event_name>show_map</Event_name>
  <Time>02:14</Time>
</User_Event>
```

```

<Location>
  <Time>02:15</Time>
  <x>-1515.07</x>
  <y>-22083.85</y>
</Location>
<User.Event>
  <Event_name>RedmarkerLocation</Event_name>
  <Time>02:15</Time>
</User.Event>
<RedMarkerLocation>
  <RedMarker>RedMarker1</RedMarker>
  <x>0.53</x>
  <y>0.80</y>
</RedMarkerLocation>

```

With this code we see that the player first looked at the map, then moved a bit and subsequently decided to indicate the location of the failure. This is illustrative of a player who learned how to play the game, because in the beginning players will call the Action Center without reporting the location or they will try to mark the location without looking at the map. Instead, this player went straight to the map after putting down the marker. For him the virtual inspection procedure became a routine. The code is from the end-exercise so it should have been at that point. The next lines of code continue to illustrate that reporting became a routine:

```

<Diagnosis>
  <RedMarker>RedMarker1</RedMarker>
  <Mechanism>1</Mechanism>
  <Time>02:15</Time>
</Diagnosis>
<User.Event>
  <Event_name>New.Signal.Observation</Event_name>
  <Time>02:15</Time>
</User.Event>
<SignalCrack>
  <Revetment>1</Revetment>
  <CrackType>1</CrackType>
  <CrossCutLocation>0</CrossCutLocation>
  <Name>SignalCrack4</Name>
  <Owner>RedMarker1</Owner>
</SignalCrack>
<SignalReportCrack>
  <LengthOfDamage>2</LengthOfDamage>
  <LengthOfDamageMax>6</LengthOfDamageMax>
  <WidthOfDamage>1</WidthOfDamage>
  <WidthOfDamageMax>6</WidthOfDamageMax>
  <MultipleCrack>0</MultipleCrack>
  <State>0</State>
  <Name>SignalReportCrack6</Name>
  <Time>02:15</Time>
  <Owner>SignalCrack4</Owner>
</SignalReportCrack>
<Conversation>
  <RedMarker>RedMarker1</RedMarker>
  <State>1</State>
  <AssessmentAC>0</AssessmentAC>
  <AssessmentPlayer>0</AssessmentPlayer>
  <SignalCount>1</SignalCount>
  <Time>02:15</Time>
</Conversation>

```

After indicating the location the player—as if running on automatic pilot—moved on to other events. First he diagnosed what failure mechanism is occurring. He thought a “macro-instability” (1 stands for this failure mechanism; other num-



bers relate to other failure mechanisms). Diagnosing at an early stage is remarkable too. Many players forgot about this in the beginning. When players realized they lost points because of this, in later stages it may encouraged them to make it the first thing they did. Another possibility is that players start thinking backwards. This is consistent with ideas related to sensemaking. Weick (1995, p. 26) says that “meanings arise retrospectively.” They start with the conclusion based on previous experiences (“This is a typical macro-instability”) and then traverse back to how they came up with that conclusion.

This traversing backwards is what happened. After diagnosing, the player reported a crack and the characteristics of the report are logged too, such as the length, the width, and whether or not multiple cracks occur. Upon finishing the report, the player immediately called the Action Center to report his findings and to assess the severity of the situation. The Action Center agreed with the player’s assessment. Both found it reportable.

This example is also illustrative of my qualitative approach to understanding how participants played the game and how they made sense of virtual risks in particular. Using the game logs I basically tried to re-imagine and re-construct how players played. The patterns that emerged I consider my *game observations*. My other approach—the quantitative one—concerned looking into game performance. To look into game performance I considered the original overall scores, the *game scores*, as well as the performance based on how players dealt with failures, the *failure correctness scores*.

The goals of this level are to describe

- How participants performed in playing the game (game scores);
- How players dealt with particular failures (failure correctness scores); and
- What prevalent gameplay patterns emerged (game observations).

## Performing in the Game

Performance in games is often depicted by means of a score, which is a performance indicator that uses numbers, letters, or any other symbol of achievement. If someone plays better this should automatically result in a higher game score. Of course, luck plays a role in getting a score. If one player accidentally takes a route leading directly to all the failures and another takes a route that does not, the former may more likely get a higher score. In most games such luck does not (and should not) have a significant influence on the end result. It is the player’s effort that makes a difference (Juul, 2005).

With this in mind, it is not surprising that a discussion exists to what extent game performance reflects learning (Washbush & Gosen, 2001). On the one hand, it is recognized that players need to learn in order to obtain a high score. On the other hand, it is acknowledged that learning does not equate with performance. One may perform terrible and still learn significantly. It may also happen that players learn how to play the game and not learn practical skills from it.

The relationship between game performance and learning is more problematic if the scoring system does not reflect the game's learning objectives. In such games, scores are used to motivate and engage players to do better. With "doing better" I refer to doing better in the game. A relationship still exists, because by continuing to play the objectives may be attained. If the scoring system works like this, we should be careful in inferring any conclusions about learning based on the game scores.

In *Levee Patroller* the scores reflect the learning objectives. To reiterate, the game has five learning objectives: observing, reporting, assessing, diagnosing, and taking measures. The game has seven (quantitative) scoring criteria and each criterion relates to one of the learning objectives. Two learning objectives, observing and reporting, have more than one criterion. For observing we wanted to make a clear distinction between finding a failure and finding the signals that make up this failure; for reporting we understood from our clients that reporting the location correctly is a frequent occurring problem and by making this a separate criterion we gave it the emphasis it needed.<sup>1</sup> Table 6.1 gives an overview of the scoring system.

Each criterion is given a somewhat arbitrary weight for calculating the *final score*. We implemented this to emphasize that some criteria are more important than others. "Location accuracy" is of an obvious lesser importance than finding a failure. Another reason for attributing weights is that for some criteria players have to perform more actions. Without the weights these criteria would gain more importance. For example, players need to fill out more than one report for many failures. By giving reporting a lower weight, finding the failure does not become under appreciated.

The final score is therefore a percentage of a summation of all the individual scores on each criterion multiplied by their weight and then divided by the total possible score in an exercise. Mathematically, one would describe this as:

$$\text{Final score} = \left( \frac{\sum \text{Criterion} \times \text{Weight}}{\text{Total}} \right) \times 100\%.$$

This final score is what I consider the game score, but it does not determine game performance entirely. For a satisfactory performance players need to ensure no levee breach occurred. If a levee breach occurred players' performance is insufficient, even if they reached a final score of 80% or higher.

Out of the 975 exercises that were completed during the training at least 170 levee breaches occurred.<sup>2</sup> Considering that in about half of the exercises no levee breach could occur, this means that in about one out of three exercises a levee breach happened.<sup>3</sup> We should be glad it was all virtual!

Later (in Level 11), I will analyze whether these criteria reflect actual learning. What I will discuss in this level is how the participants performed: how players performed on "average" and how they "progressed over time," on the individual criteria and on the final score.

<sup>1</sup> Players disagreed with the need for location accuracy, see Level 5.

<sup>2</sup> The actual total number of levee breaches is probably higher. The number 170 is based on all the files that I retrieved.

<sup>3</sup> Some failures never lead to a levee breach. In addition, the first exercises did not have any critical failures. For this reason, the start-exercise and exercises one, two, and five could not lead to a levee breach. That is exactly half of the exercises.

**Table 6.1** Overview of *Levee Patroller*'s scoring system

Criterion	Description	Learning objective	Weight	<i>M</i>	<i>SD</i>
Observed failures	The player indicates that he or she has found a failure	Observing	10	.77	.24
Location accuracy	The player specifies the location of the failure	Reporting	1	.60	.29
Observed signals	The player reports what signals are part of the failure	Observing	5	.53	.25
Reporting accuracy	The player fills out a report for each signal	Reporting	1	.32	.19
Assessment accuracy	The player makes an estimate of how severe the situation is	Assessing	2	.41	.24
Diagnose accuracy	The player determines the failure mechanism behind a failure	Diagnosing	5	.46	.29
Measure effectiveness	The player takes an action to prevent the failure from becoming worse	Taking measures	5	.58	.33

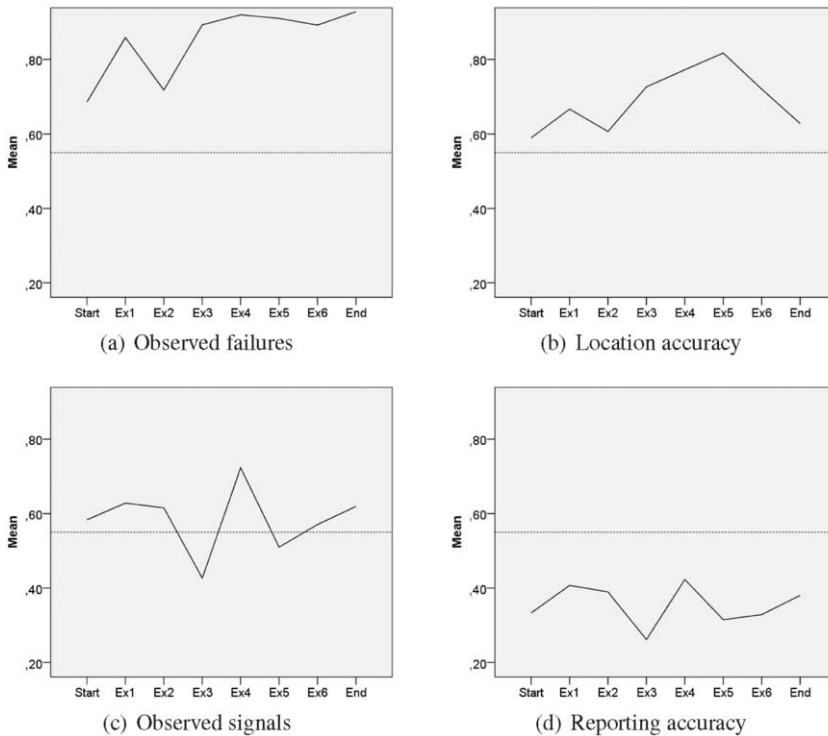
*Note.* The mean and standard deviations are based on the relative scores of all participants ( $N = 147$ ). The relative scores are the actual scores on a criterion divided by the total scores possible. For measure effectiveness the sample is less ( $N = 132$ ), because a number of participants did not play any exercise in which they had to take measures.

My first step was to calculate the average relative scores per criterion based on all the exercises that were played. These relative scores are the actual scores divided by the maximum score per criterion. The results are depicted in Table 6.1. From these numbers we can draw three conclusions. First, players did not have much difficulty in finding failures. Second, in contrary to finding failures players were not so good at reporting them. Third, with standard deviations ranging from .19 to .33 it becomes clear that performances differed widely.

Further investigation of the third conclusion reveals that participants scored on the complete possible range for every criterion, from no points at all to an almost perfect score. This hardly changes when we only include participants that played at least four or five exercises. This means that until the very end a large variety in scores were achieved and that some players were never able to get a high score.

On average participants achieved a satisfactory final score. If we neglect the levee breach criterion, players need to achieve a score of 55% to reach a sufficient mark and as a matter of fact, the relative average final score is exactly .55 ( $SD = .22$ ). Based on this, we could conclude that the scoring system is well balanced.<sup>4</sup>

<sup>4</sup> A game would ideally allow everyone to get a sufficient score at the end. For this to happen, a game should enable a variation in difficulty. In this way, players could play the game according to their own skill level. The current version of *Levee Patroller* does not vary in difficulty and from a research perspective this is also unwanted. It makes it harder to compare the results.



**Fig. 6.1** Overview of the scores on failures, locations, signals, and reports. The scores are expressed as a division of the maximum score in an exercise. The horizontal, dashed line represent a score of .55

My second step was to consider the average scores per criterion and the final score over time. Figures 6.1 and 6.2 show these developments graphically.<sup>5</sup> These are my findings:

*Observed failures* Participants become better with finding failures over time. At the end-exercise almost everyone found every failure. The development is not completely linear. In Exercise 2 people had more trouble finding the failures than with Exercise 1. This is because Exercise 2 had three failures that participants had trouble with finding at first. A reasonable explanation for the decline at Exercise 6 is that they entered a new region and needed orientation time.

*Location accuracy* The small decline at Exercise 2 is easy to explain as a result of the performance on the criterion “observed failures.” If players do not find a failure, they cannot report the location. A more interesting result is that the accuracy dropped to its initial performance with entering the new region in Ex-

<sup>5</sup> Deviations (or errors) are not considered in Figures 6.1 and 6.2, because they have a wide range and this makes inclusion useless. It should be kept in mind that the results varied widely among participants.

ercise 6. Although orientation could be a factor of influence, other factors may have played a role too, because unlike with “observed failures” a recovery does not occur. One possibility is that players were so preoccupied with the many failures that they spent less time and attention on reporting the location.

*Observed signals* With this criterion we do not see a real increment. Three exercises have a clear deviation: Exercises 3, 4, and 5. Exercises 3 and 5 are similar to each other. They have the same region with similar failures. A probable explanation is that these exercises contain failures that players had trouble with. Later in this level we will see that players performed worst on the failure “watery slope” and this failure only appeared in these two exercises and in Exercise 1. Exercise 4 has a peak on almost every criterion. Participants did well in this exercise in general.

*Reporting accuracy* This criterion follows the observed signals pattern almost identically, only with an inferior performance. The pattern is identical, because for each signal at least one report should be made and if that signal is wrong, the report is wrong too. The pattern has a lesser performance, because if the signal is correct, the report could still be wrong. To be considered accurate, at least two-thirds of the reporting items need to be correct. Another reason it has a lesser performance is that participants may have missed one or two reports. One signal could have as many as three reports.

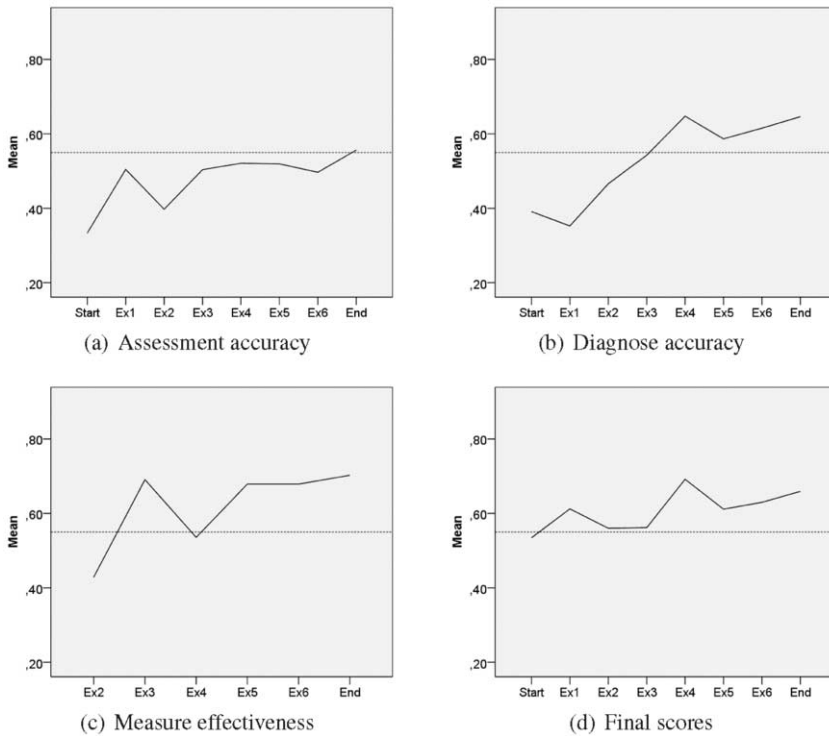
*Assessment accuracy* In terms of assessment, a clear improvement can be noticed over time. The small decline after Exercise 1 is due to not finding failures. Although the accuracy seemed to stabilize after Exercise 3, a firm improvement is noticeable at the end-exercise. What may have happened is that players started to accept the game’s answers at this point to get a better score (Level 5).

*Diagnose accuracy* With this criterion we can certainly speak of an improvement. The accuracy almost doubled from Exercise 1 to Exercise 4. It did not increase after that and I think this happened because a number of people have been guessing the failure mechanisms.

*Measure effectiveness* Measures were not necessary in the start-exercise and Exercise 1. Many participants did not take measures in Exercise 2, because the critical phase of the failure in that exercise is hard to see (i.e., this is the “stone damage” failure). This explains its low performance and that of Exercise 4. The latter exercise has this failure too.

Much like the individual criteria, the final score increased over time. A clear peak is noticeable at Exercise 4. During this exercise, it seems that most players got used to the game, that they were able to “read” it. From the previous level we also observed that a peak was to be seen at precisely this exercise when it comes to the appraisal of the experience. So it looks like performance and appraisal go hand-in-hand.

Multiple reasons exist to explain why the final scores increased dramatically at Exercise 4 and then declined. We know from previous runs with the game that players have more difficulty with the second region than with the first. This second region, which is used in Exercises 1, 3, and 5, has many curves, making orientation



**Fig. 6.2** Overview of the scores on assessments, diagnosis, and measures. The fourth subfigure gives an overview of the final scores. The scores are expressed as a division of the maximum score in an exercise. The horizontal, dashed line represent a score of .55

harder. Another impact is that this second region contains a failure with which players did not perform well (the “watery slope” failure). Then, in Exercise 6 and the end-exercise players, were introduced to a new region, and it seemed players needed some time to adapt to this. In addition, these two exercises contained many critical failures, which may also have influenced performance.

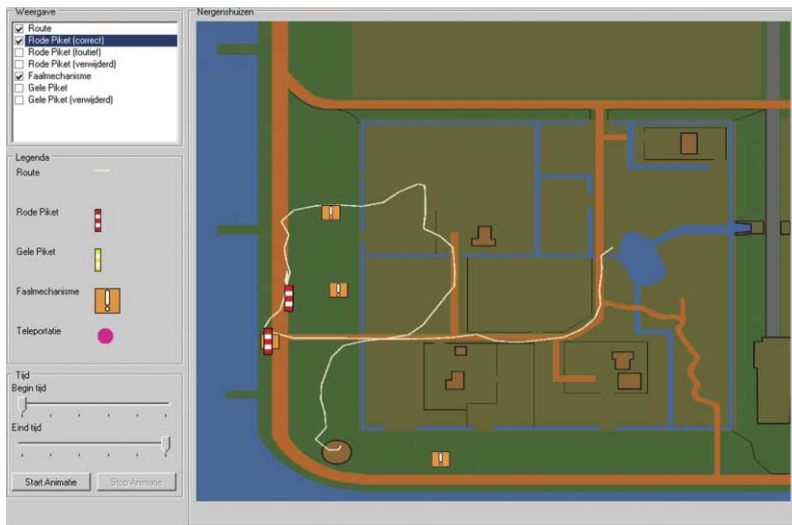
Statistical analysis (Repeated Measures ANOVA) confirmed that participants changed their performance over time. The results further suggest that variety within participants is less strong than the variety between them.<sup>6</sup>

<sup>6</sup> On all statistical tests for the scoring criteria, Mauchly’s Test of Sphericity indicated that the assumption of sphericity had been violated,  $p < .001$ , and, therefore, a Greenhouse-Geisser correction was used. With or without this correction, all Repeated Measures test results, within-subjects as well between-subjects, indicated a value of  $p < .001$ . This tells us that participants scored differently. Effect sizes for within-subject results ranged between .078 and .20 and for between-subjects between .82 and .97. These conclusions do not change if we consider the performances for each region separately.

## Dealing with the Virtual Failures

I used the game data to look into the performance on the individual failures too. This allowed for a deeper understanding of how players made sense of the virtual failures and it made me open the black-box completely. I looked into almost every detail, from reporting the crosscut location to mentioning the type of revetment.

My initial plan was to process the log files with the *Levee Patroller Logging Tool* (Fig. 6.3). This tool reads one or multiple log files and is able to show all elements from the log file in a user friendly way. Its nicest feature is to show on a map how players walked during the exercise and where they placed a reporting marker and measuring marker.



**Fig. 6.3** Screen shot of the Levee Patroller Logging Tool. Created by Arjan Peters

I ended up not using this. Besides minor bugs and issues with reading files from the server<sup>7</sup>, using the tool proved to be more time consuming than analyzing the raw XML files. I also decided against extending the *Levee Patroller Data Collection Tool*, which I used to automatically extract the game scores and questionnaire data (Level 4). Although automatic extraction is more failure-proof, I would have missed many observations and gained less insights with automatic extraction.<sup>8</sup> The manual procedure I opted for enabled me to hypothesize about what the players thought. It allowed me to re-imagine and re-construct how players played.

<sup>7</sup> If files were sent to the server, they did not have a neat XML structure anymore. The *Levee Patroller Logging Tool* could not handle this. I first needed to restructure the files before I could use them with this tool.

<sup>8</sup> I used data validation techniques to limit errors with the manual procedure.

This was my method of working in analyzing the game data:

- I analyzed the game data from player to player. I first looked at all exercises made by one player, starting with the start-exercise and ending with the end-exercise, before I proceeded to the next player. This gave me a sharp idea of how the player progressed over the exercises.
- For each exercise, I wrote down notes for every included failure and for the exercise in general. After looking at all exercises of one player, I wrote down my thoughts about this player.
- Parallel to reading the log files I extracted facts and wrote these systematically down into a spreadsheet. Besides the previously mentioned notes, I noted:
  - *If participants had played the exercise before*: If participants had played the exercise before: Previous experience will influence the results. Players had to play an exercise for more than five minutes before I considered them to have played it before.
  - *If participants found a failure and at what time*: This was to find out if certain failures were found more often than others. I wrote down the time to see if certain failures were persistently discovered earlier than others.
  - *The signals and their reports and whether they were correct*: For each failure I identified beforehand which signals players had to report and what reports. I wrote down whether they observed this or something else. If they reported this correctly, I included how they filled out their reports.
  - *All other aspects of a failure, such as the assessment, diagnosis, and measures taken*: Choices on these aspects were written down and noted whether or not they were correct.
  - *If participants reported non-failures*: Non-failures are failures reported by players, but which are not considered failures in the game. If players reported non-failures, I wrote down their coordinates and the first reported signal reported—if the player reported one.

Table 6.2 gives an idea of this systematic notation. It shows a part of reporting the first phase of the small landslide failure. In this phase players will find two small parallel cracks. The table shows participants who did not find the failure at all (Jan and Ton), a participant who did not think it was a crack but a settlement (Ingrid), and three others who did find the failure and said it concerned a crack (Hans, Peter, and Ben). The procedure resulted in 1.470 variables.

The procedure also resulted in the “game observations” and “failure correctness scores.” The game observations are the prevalent gameplay patterns that emerged from studying the log files. I will discuss these later in this level. The failure correctness scores are what I will discuss first. The failure correctness scores are the unweighted scores players achieved on the specific failures in total and on each of the learning objectives for specific failures.

Before I will detail the failure correctness scores, it is necessary to review some of the game elements and mechanisms:



**Table 6.2** A part of the systematic procedure for analyzing the log files

Player	Found	Time	Asses	Signal	Revetment	Type	Length	Width	Multiple
Hans	Yes	5:52	Reportable	Crack	Asphalt	Parallel	Medium	Small	Yes
Peter	Yes	10:11	Severe	Crack	Grass	Parallel	Small	Small	Yes
Jan	No								
Ingrid	Yes	7:33	Reportable	Settlement					
Ton	No								
Ben	Yes	16:47	Critical	Crack	Asphalt	Perpendicular	Large	Large	No

*Note.* The original file worked with numbers and not with words, because this would allow for a quantitative analysis with statistical software. For ease of interpretation I used words here. The original file also indicated whether the answers were correct.

- The research version contains six failures: stone damage, small landslide, watery slope, boiling ditch, grass damage, and illegal driveway. We identified three failure phases: reportable, severe, and critical. Some failures have two phases; others have three. In each exercise a number of failures are placed and this number is increased over the exercises. At first the failures do not always develop into their critical phase. At the very end almost all of them do.
- In dealing with the failures players have to mention their severity (assessing), signals (observing), and failure mechanism (diagnosing). A failure has one or more signals. For each signal and every change of this signal a report must be made (reporting). If eventually the failure becomes critical, players have to take a measure (taking measures).
- One important design choice was to let players work with exact measurements (Level 2). With the measuring (or yellow) marker they are able to retrieve these measurements. In reporting, however, they have to choose every time from six ranges. The ranges represent if the size of the signal could be considered very small, small, medium, medium large, large, or very large. I will use these representations in representing the reporting outcomes.

### **In-depth explanation: calculating the failure correctness scores**

Every failure has the following general correctness variables: assessment per phase ( $A1$  to  $AN$ ), signal observations ( $O1$  to  $ON$ ), diagnosis ( $D$ ), and measure ( $M$ ). Failures also have specific correctness variables. These are tied to the various reporting elements ( $R1$  to  $RN$ ). For every signal the amount and type of reporting elements differ.

All failure correctness scores are displayed in percentages. This makes it possible to compare the scores, although we should keep in mind that some failures are composed of more elements than others. Except for the overall performance, the performances are calculated per failure phase: for the reportable ( $x = 1$ ), the severe ( $x = 2$ ), and the critical phase ( $x = 3$ ). In calculating the total correctness scores I ignored the learning objective taking measures, because players had to implement measures infrequently. It does not count for every failure.

For determining the performance of assessing in a phase of a specific failure, the assessment correctness variables for that phase ( $AN$ ) are summarized for the number of

failures  $n$  and then divided by this same  $n$  to achieve an average. To calculate a percentage, this average is multiplied by 100%.

$$\text{Assessing}_x = \frac{100\%}{n} \sum_{i=1}^n AN_i$$

Calculating observing is less straightforward. Some failure phases have more than one signal. If that is the case, multiple observing correctness variables exist per failure phase (from  $O1$  to  $ON$ ) and these have to be summarized. It also means that to calculate the average the number of failures  $n$  needs to be multiplied by the number of signals  $N$ .

$$\text{Observing}_x = \frac{100\%}{n \times N} \sum_{i=1}^n O1_i + O2_i + \dots + ON_i$$

The reporting formula is similar to observing, except with this calculation multiple items always exist. One report could have as many as ten reporting items. I only included the results by players who had at least one signal correct in a phase, because if the signals are incorrect, the reports are incorrect too. In this way, the reporting results are not biased by the signal observations.

$$\text{Reporting}_x = \frac{100\%}{n \times N} \sum_{i=1}^n R1_i + R2_i + \dots + RN_i$$

The last two learning objectives—diagnosing and taking measures—are calculated in the same way as assessing, except that they are not calculated per phase. Both have one possible answer for every failure. What is considered correct differs between the two. In the game players can change the diagnosis continuously. The last choice is decisive. I applied this principle with the correctness scores too.

$$\text{Diagnosing} = \frac{100\%}{n} \sum_{i=1}^n D_i$$

Taking measures works in the opposite way: the first choice determines the score. I only considered this calculation for failures for which a measure had to be taken.

$$\text{Taking measures} = \frac{100\%}{n} \sum_{i=1}^n M_i$$

For calculating the total scores per phase the scores on assessing, observing, and reporting are averaged. I used a different reporting score—one without the restrictions. With the total score the idea is that everybody is included. In addition, some phases do not include a new signal observation. This means that observing does not play a role. If that happens the sum needs to be divided by two ( $k = 2$ ) instead of by three ( $k = 3$ ).

$$\text{Phase}_x \text{ total} = \frac{1}{k} (\text{Assessing}_x + \text{Observing}_x + \text{Reporting}_x)$$

In calculating the total failure correctness score, two rules were applied. The first rule is that each learning objective contributes equally to the outcome. The second rule is that each phase contributes equally to the outcome. To ensure the latter observing needs to be considered separately. As explained above, some phases do not include this. That is why the number of phases  $k$  involved with observing could be different than the number of phases  $n$  a failure has.

$$\text{Failure total} = \frac{\text{Diagnosing}}{4} + \frac{1}{k \times 4} \sum_{x=1}^3 \text{Observing}_x + \frac{1}{n \times 2} \sum_{x=1}^3 \text{Assessing}_x + \text{Reporting}_x$$

Table 6.3 gives an overview of the failure correctness scores and shows average performances on the learning objectives per phase, the totals per phase, and the overall total score for each failure. Having established my method for understanding

how players dealt with the virtual failures, we can now move on to discussing the failures, including the non-failures.

**Table 6.3** An overview of the failure correctness scores for each failure, per learning objective, phase, and in total

Score, in $M(SD)$	Stone damage	Small landslide	Watery slope	Boiling ditch	Grass damage	Illegal driveway
Found	88(17)	90(17)	85(23)	70(28)	90(27)	91(24)
Reportable						
Assessing	NA	70(32)	75(34)	70(37)	NA	75(40)
Observing	NA	98(10)	59(28)	90(25)	NA	64(21)
Reporting	NA	77(9.8)	55(19)	86(14)	NA	40(24)
Total	NA	82(12)	62(23)	79(23)	NA	59(18)
Severe						
Assessing	49(33)	57(31)	72(36)	45(45)	42(45)	NA
Observing	95(18)	64(42)	NA	NA	55(20)	NA
Reporting	86(8.8)	43(20)	87(13)	67(18)	44(17)	NA
Total	77(15)	45(22)	54(30)	58(26)	46(19)	NA
Critical						
Assessing	77(38)	96(15)	83(35)	66(41)	92(21)	88(29)
Observing	NA	74(38)	27(42)	NA	NA	91(29)
Reporting	69(19)	54(27)	69(20)	56(11)	68(16)	84(15)
Total	73(23)	70(19)	46(23)	57(23)	80(14)	82(17)
Diagnosing	82(31)	70(33)	59(40)	88(24)	67(41)	54(45)
Taking measures	82(34)	79(38)	72(39)	79(34)	85(30)	92(23)
Total	77(15)	69(13)	56(19)	72(18)	60(16)	67(18)

*Note.* The sample ( $n$ ) differs widely among scores because some players did not play all exercises; players may have not reported or done something; and of the missing data. For the reporting objective players had to have reported at least one correct signal. NA = Not Applicable.

### *The Disappearing Stone Damage*

The stone damage is probably the simplest failure. It has one signal and a maximum of two phases. However, with this failure, at some point, when the water level rises, it becomes difficult to inspect the failure. Because of this players missed or forgot about the failure and did not inspect the second, more critical phase. In the three instances the stone damage failure became critical, on average 67% of the participants who found the failure reported this compared to 95% who reported the severe phase. That is still more than half of the participants.

In the first phase of this failure missing stones are to be observed (Fig. 6.7). These missing stones concern pitching stone, because they are neatly ordered. If they were not, we would speak of rip-rap. The water does not reach the failure and flushing

soil cannot occur. In the later phase, the water does rise and if the soil starts to flush, it results in a much bigger damage.

Other than the difficulty of seeing the failure in this phase, some players complained that they did not understand when the failure would become critical and when not. In both situations the water level rises, meaning water reaches the damaged revetment. That should always result in degradation—at least, according to some players. That is a valid point but it may also happen that the remaining revetment is strong enough to prevent that.

As a result of this difficulty, many players did not or were not able to check if the failure became worse. It seemed that many players automatically filled out the reports, especially in the later exercises.<sup>9</sup> They either assumed the failure became worse or they filled it out just in case. To them, prevention was better than cure. This shows that they gained a clear expectation of what could happen and acted with this failure development in mind.

Despite the difficulty, players performed very well and from the beginning. Although some individuals chose at first crack, liquefaction, water outflow, human activity, and especially settlement, at the start-exercise the majority (85%) chose rip-rap/pitching stone already. At the very end everybody chose this. Other choices are not immediately non-sense. Take for example human activity. In my talks with patrollers I heard frequently that stone damage is inflicted by young people who for unknown reasons remove stones from levees.

Almost everybody (> 90%) had the reporting items crosscut location, revetment type, and stones loose and missing correct over all exercises. The answers for these items may have been obvious. This is less true for all other reporting items. Concerning the length, a small majority (49%) reported the damage as very small at first, but after Exercise 4 a majority (70%) reported one of the correct answers: small. With width, no changes occurred over the exercises. About half were correct all the time.

In indicating the length and width in the critical phase, participants were less correct. On the one hand this is rather surprising, because more answers are considered correct. On the other hand, in this phase it became difficult to measure the failure. Many players may have guessed. Consistent with this idea, answers were spread out over all exercises and so eventually about half had it right. This difference explains largely the underperformance of this phase compared to the severe phase. As for the flushing of soil, the majority (70%) was right in most circumstances.

Regarding assessment, a subtle change occurred. Initially, more people (53%) thought it was reportable against those who thought it was severe (43%). This reversed at the end. For the critical phase no changes happened. About 77% had it

**Reporting stone damage**

*General*

Signal:	Rip-rap/Pitching stone
Crosscut:	Outer slope
Revetment:	Pitching stone
Stones loose:	Yes
Stones missing:	Yes

*Severe*

Length:	Small or medium
Width:	Small
Soil flushing:	No

*Critical*

Length:	Medium large or large
Width:	Medium to very large
Soil flushing:	Yes
Soil quantity:	Much

<sup>9</sup> Clear evidence of this is the number of people who took measures for stone damage failures that did not become critical. At the start-exercise only 5% decided to take a measure. For Exercises 4 and 6, this percentage quadrupled.

correct in every exercise with a critical stone damage. Participants did better with diagnosing and taking measures (both 82%). Players opted for various alternatives, but clearly the majority had it right all the time.

The specific game results with stone damage confirms the supposition that this was an easy failure. What struck me is that players performed well on this failure from the onset. This does not mean no improvements occurred. Until the fourth exercise players markedly improved their performance on this failure.

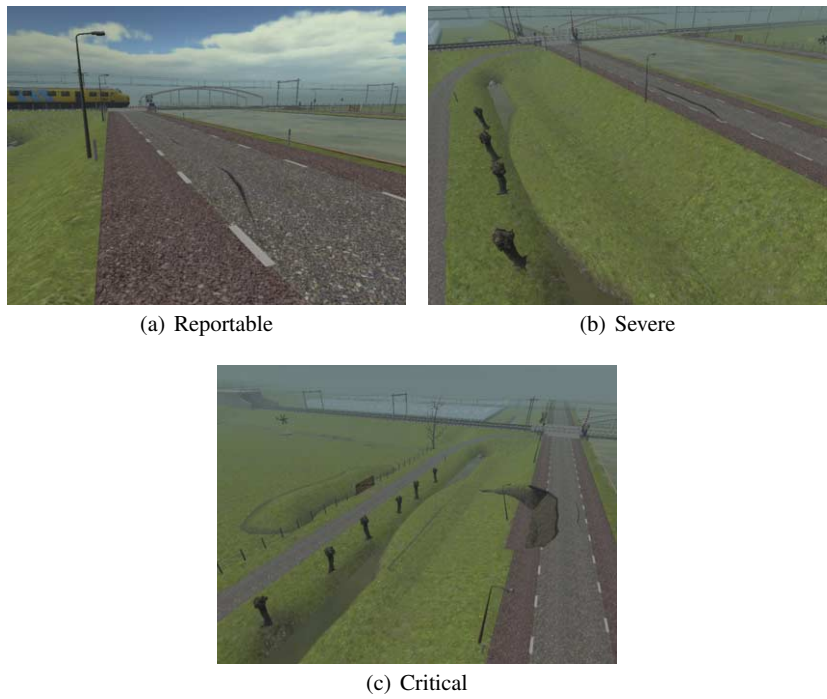
### *The Scary Small Landslide*

The name may suggest otherwise, but the most spectacular failure in the game is the small landslide. It is called small because the failure is relatively small compared to the landslides that may occur in mountainous areas. It also starts small. First, two cracks are seen on the crest of the levee (Fig. 6.4). Later the two cracks become one large crack. In this critical phase players have to report another signal: horizontal movement. This movement happens at the toe of the levee, in the hinterland. A large part of a ditch is closed off because of this movement. If the failure continues to develop a settlement occurs. The inner slope slides away and if no measure is taken other levee parts will slide away too.

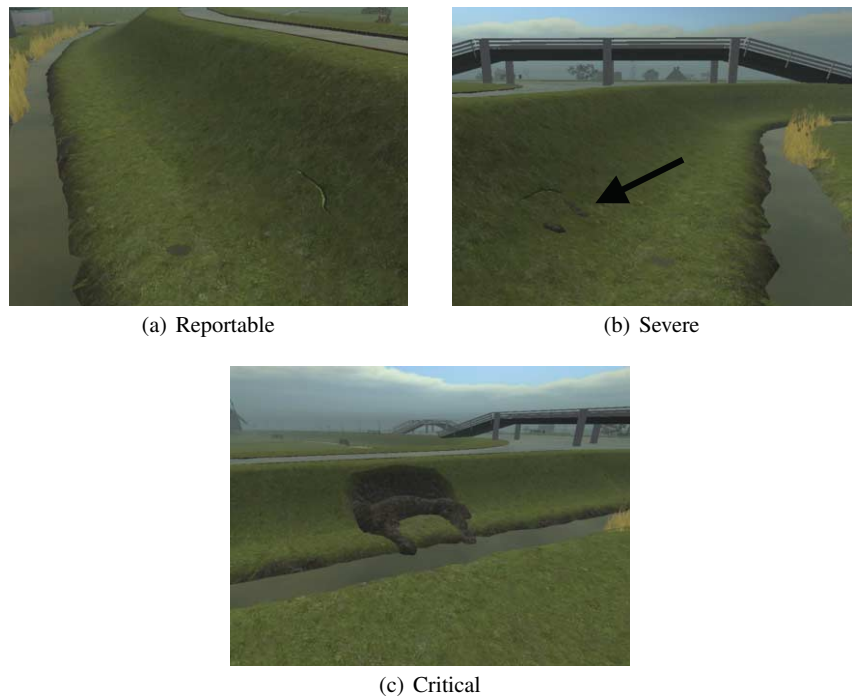
In the game two different versions exist. One is the primary levee version, the other the regional levee version. They develop and look similar, except that the scale of the failure is smaller for the regional version. All length measures are therefore different between the two. Because of these two versions, this failure is the only one that occurs in all regions—one of the reasons it is included in most exercises. The other reason is that the small landslide is more complicated than the other failures, in terms of phases and signals, and for this reason the research version provided players with enough experience of it.

If the failure changes, almost three-fourth (71%) of those finding the failure found it during its reportable situation. Almost everyone reported a crack here. Except for providing measurements, players had little trouble in filling out the reports. Over all exercises players answered the crosscut location, type, revetment, and number correctly. A small majority (58%) picked the right length for the regional as well as primary situation.

For the width, something curious happened. At the start players were especially split between choosing very small and small. The remaining players chose medium. Over the exercises more and not less players chose very small, the incorrect answer. This most likely happened because of the introduction of the severe phase. Although the width remained the same, players reported larger sizes here. Almost three times as many chose medium compared to the reportable situation (8% compared to 22%). It seems that reporting is contextually bound. With the increase of the length, players assumed that the width increases too and that in an earlier phase it must have been smaller.



**Fig. 6.4** The phases of the small landslide failure



**Fig. 6.5** The phases of the watery slope failure. The arrow point to some flushing soil

Except for the width, the crack reports in the severe situation show two other differences. First, a difference between the regional and primary version can be seen. In this phase, about half (52%) had the length correct, whereas in the primary situation many more (72%) did. Players with incorrect answers had a tendency to report larger lengths in the primary situation than with the regional one. This is another indication of the importance of context in making a decision.

Second, with number of cracks the reverse was filled out. Now a large majority (85%) said no, which is correct, because in this phase the two separate cracks have become one and the same crack. Beforehand I had my doubts whether players would take notice of this. They did and to me this shows how neat and precise they filled out the reports.

Many (59%) who reported the crack in the reportable situation added a second report in the severe situation. Others did not report the reportable situation or did not report a crack during the severe situation. The latter only happened in Exercises 4 and 5. One-fourth of participants noticed nothing more than the horizontal movement. Suprisingly, the majority also reported this signal correctly from the beginning. I was surprised about this, because horizontal movement is not a term that is used on a daily basis. It needs some explanation. The other frequent chosen option concerned one that is more well-known: settlement. Throughout the exercises, the number of players reporting the signal correctly increased (from 60 to 80%).

However, the number of players observing the horizontal movement severely decreased. Whereas first the proportion of players reporting a crack and horizontal movement were more or less the same, after Exercise 5 the proportion started to become increasingly in advantage of people reporting the crack. At the end-exercise twice as many players reported a crack as the horizontal movement. It looks like players started to forget about this signal!

Those reporting the horizontal movement were less accurate than in reporting the crack. Just a small majority (63%) had its crosscut location right and about half its length (49%). With the width answers were spread out. The most chosen answer concerned small rather than medium, the correct response. Participants had less issues with classifying the revetment. As with the crack, almost everyone was correct.

Something I noticed with reporting the severe phase is that most of the time players first found the crack, made its report, and then called immediately the Action Center. They almost never looked around to see if any other signals could be found. Some continued to look for other signals after they called and if they found the horizontal movement, they always increased their assessment and wanted to take a

<b>Reporting small landslide</b>	
<i>Reportable</i>	
Signal:	Crack
Crosscut:	Crest
Revetment:	Asphalt
Type:	Horizontal
Length:	Medium (small)
Width:	Small-medium
Multiple:	Yes
<i>Severe</i>	
Signal:	Crack
Length:	Large (medium large)
Multiple:	No
Signal:	Horizontal movement
Crosscut:	Hinterland
Revetment:	Grass
Length:	Medium
Width:	Medium
<i>Critical</i>	
Signal:	Settlement
Crosscut:	Crest
Revetment:	Asphalt
Direction:	Land side
Length:	Very large (medium)
Width:	Very large
Height:	Very large

measure. Apparently, it seemed as if the failure worsened in the meantime or they may have realized that it might have been worse than they initially thought.

Whatever their reasoning, this might explain why participants were much more ambivalent initially about the assessment of this situation compared to the reportable phase. During the reportable phase about two-thirds to three-fourth continuously chose reportable over severe. With the severe phase responses were split fifty-fifty between severe and critical. After exercises four and five a solid majority (71%) chose severe over critical. Most likely—and this is the second reason to explain the mixed results—this is one of the places where players said they went along with the game instead of expressing their own opinion. Some players could not believe that they did not have to take a measure here. They tried all possible measures twice.

With the critical situation, hardly anybody doubted its state. This time most participants (75%) reported settlement instead of horizontal movement. It is interesting that some players chose crack; in Exercise 6, about 7%, and in the end-exercise, 13%. These players reported a huge crack and then took a measure. In some way, the situation could be seen as a huge crack, but this seems to stretch the definition of a crack.

Those reporting the settlement provided mixed results and some for a good reason. The settlement starts at the crest, which is made of asphalt, but the larger part of it concerns the inner slope, which has grass as revetment. The majority eventually chose the “wrong” answers for the crosscut location (69%) as well as the revetment (58%). With the length and width responses were surprisingly poor. It is clearly a *very large* settlement. Yet although most chose this, not a convincing group did so (46% and 40% for the length and width, respectively). All other responses were spread out over the options. With height, a clear majority (83%) chose correctly.

The small landslide failure is a clear example of macro-instability, which most players (70%) observed correctly throughout the exercises. Because some may thought it was indeed a small landslide, they may have opted for micro-instability, which was the most chosen alternative. In case the failure became critical a few (7%) decided it must be erosion inner slope. The best way to deal with a macro-instability is by placing sand on the slope to provide for a counterweight. The majority (78%) seemed to agree, and the others were of the opinion that with some sandbags, such a counterweight could also be provided.

### *The Difficult Watery Slope*

The watery slope only occurred in Region 2; that is, Exercises 1, 3, and 5. It shows one of the more difficult failure mechanisms: micro-instability. Hereby the levee erodes from the inside out. In the game, it starts with a small crack (Fig. 6.5). If one would observe carefully, they would also discover a small puddle in its vicinity. This is a subtle indication that water is flowing out of the crack. Both signals—the crack and the water outflow—need to be reported in this phase.



During the more critical phase, another subtle hint is given. In addition to the puddle, players should see flushing soil, a clear indication that the situation is becoming worse. If the process continues, the critical phase is reached. In this phase, a large part of the levee bulges, because the slope is too watery, something that is referred to as liquefaction. Players have to report this liquefaction and take measures as soon as possible. They could place sand or cover the inner slope with foil.

After the training session with Organization A, I decided to show how to play the game during the start-meeting (Level 4). I secretly played Exercise 1, which contains this failure and the previous one, the small landslide. In this presentation I made sure to highlight that they had to report two signals with the watery slope failure in the reportable phase. This improved the performance for Organization B and C, yet especially in the beginning. A little more than half (56%) reported at least two signals during the start-exercise, but this dropped to about one third (36%) in later exercises. This drop cannot be fully explained by participants finding the failure until very later, when it is critical and the signals are no longer visible. In Exercise 5, one of the watery slope failures remains reportable and it has the exact same percentage of participants reporting two signals compared to the watery slope failure in that exercise that does become critical.

A large majority (82%) of those finding the failure reported the crack. A much smaller number (37%) reported the water outflow, although a good portion spoke of liquefaction (13%). The remaining people signalled above all grass damage or settlement in addition to or instead of the crack. All alternatives are understandable. In the end the failure shows liquefaction; the grass is damaged; and due to the crack's location on the inner slope it may seem it is sagging.

Among those reporting water outflow during the severe phase, almost all (90%) stated little soil was flushing. Few participants (15%) reported the water outflow during the reportable as well as the severe phase. Most did so during the critical phase, probably because it took players a significant amount of time to find the failures. Filling out the remaining reporting items for the water outflow signal turned out to be simple. Hardly anybody made an error. This is similar to reporting the crack. Players only had some trouble with reporting its width. About half got it right.

Regarding the critical phase, many more thought they were seeing a settlement (50%) rather than liquefaction (29%). This is understandable, despite the fact that on several occasions—in the handbook and on the website—it is clearly stated that if soil is wet, it cannot be a settlement. But when the failure bulges, most participants assume it is a settlement. Others (12%) added another water outflow report, this

#### **Reporting watery slope**

##### *Reportable*

Signal: Crack  
Crosscut: Inner slope  
Revetment: Grass  
Type: Horizontal  
Length: Very small  
Width: Small–medium  
Multiple: No

Signal: Water outflow  
Crosscut: Inner slope  
Length: Very small  
Width: Very small–small  
Water velocity: Slow  
Water quantity: Little  
Multiple: No  
Soil flushing: No

##### *Severe*

Signal: Water outflow  
Soil flushing: Yes  
Soil quantity: Little

##### *Critical*

Signal: Liquefaction  
Crosscut: Inner slope  
Length: Medium large–large  
Width: Very large  
Soil flushing: Yes  
Soil quantity: Much

time indicating that much soil was flushing. This is a logical pick considering water outflow had to be reported earlier and it is not necessarily wrong. Liquefaction is just a more accurate description given the situation. Much to my surprise and similar to the small landslide failure, the remaining few reported a huge crack and then took a measure.

For reporting liquefaction, most (93%) said flushing soil occurred, and to a large extent (82%). Measurements were spread out, more so for the width than length. With the length about half gave the correct answer, but only a few (21%) correctly named the width. The phase with liquefaction was furthermore assessed correctly, much similar to the reportable phase. A notable number of participants thought the severe phase was either reportable or critical, but over the exercises an increasing majority said it was severe.

Despite the difficulty of understanding the failure mechanism micro-instability, a small majority (59%) was correct in diagnosing the watery slope failure. Some rather thought it was its big brother, macro-instability. This especially happened with the watery slope failures that became critical. That is not surprising, because a sort of landslide occurs, a typical characteristic of macro-instability. Another frequent choice concerned erosion inner slope.

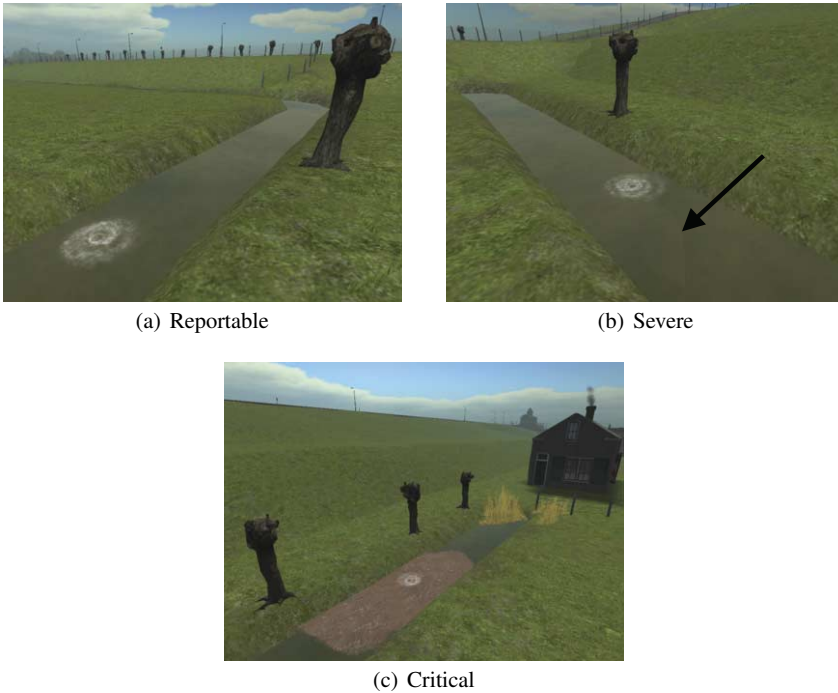
Participants were better in picking the right measure (72%), but two answers were correct. Players had a tendency to choose placing sand on the slop (61%) over covering the inner slope with soil (12%). The other participants chose placing sandbags at the toe as their preferred measure. Although this measure does provide a counterweight, it is not as effective as the other two measures.

### ***The Hard to Find Boiling Ditch***

As mentioned earlier, some players indicated they had some trouble in finding the boiling ditch failure. Unlike the other failures, this failure occurred somewhat farther off the levee, in what is called the hinterland. It has a subtle signal that is especially hard to discern if it is raining. The one time it rained, during the second exercise, less than half (42%) found the failure.

The severe phase is another aspect of this failure that is difficult to discern. If players would look carefully, they would see at the bottom of the ditch the flushing of soil. I hinted at this during the start-meeting and on the website. Nevertheless, most participants never reported this phase. On average, 21 participants reported in every relevant exercise this phase, and of those, about half reported it as if it were reportable. Even those who did report the flushing of soil in this phase, they stopped doing this in later exercises. The highest average score in this phase is also achieved during the second exercise. It seems that if a signal is unclear, players quickly forget about what is occurring and return to their natural habit of observing the phenomenon.

Players skipped many phases of this failure. Many did not find this failure until its critical phase. Others were often too late in reporting it. Especially during the last



**Fig. 6.6** The phases of the watery slope failure. The arrow points to the soil at the bottom of the ditch

two exercises, I noticed that a number of players started to report this failure when the levee was about to breach. They reported signals as overtopping and settlement. It is probably the failure that causes most levee breaches. All of this explains largely why the total performance score of boiling ditch is relatively low.

On a more positive note, the majority reported many items correctly. Over all exercises, the majority chose the correct diagnosis (+/- 90%)—sand boils—and measure (>75%)—sandbag containment ring. Also from the onset, the majority (>85%) always reported water outflow. However, players did indicate that they had some trouble in finding the right term for describing the signal of this failure:

I thought the exercise was hard, because I could not make the right choice in all circumstances. For example, I had to look for a good replacement for seepage—GQex2-#43

Two Dutch terms are well-established to describe this signal: *kwel* and *wel*, the Dutch equivalents for seepage. Although they are well-established, they are incorrectly used by some interchangeably. *Kwel* refers to the general phenomenon of groundwater finding its way to the surface. *Wel* refers to a highly concentrated amount of groundwater finding its way to the surface. Sometimes soil is flushed with this process and that is when the failure mechanism sand boils may occur. With the boiling ditch failure we are clearly dealing with a *wel* and not a *kwel*.

Since both were not listed, participants had to look for an alternative and it seems that most ended up with the right choice.

Participants further struggled in picking the right crosscut location and indicating whether more than one water outflow occurred. With the remaining reporting elements results varied among the phases. For length and width players performed rather well during the first two phases and poor during the critical phase (only about 20% were correct). Players continued to indicate small sizes and, therefore, seemed to neglect the large mud puddle that came into being.

What is interesting is that whereas players neglected the mud puddle, players seemed to assume that because the failure became worse, the water velocity and water flowout became fast and much, respectively. This is not the case. The velocity and water flowout remained the same over all phases. This observation is one of the reasons I constructed the “context is influential” gameplay pattern.

As for the flushing of soil, practically everybody said no during the reportable phase and clearly everybody said yes during the critical phase. With the severe phase, the results are mixed: some said yes, some said no. This confirms the idea that some players reported this is as if it were in its reportable phase. This pattern is repeated with the assessments. A clear majority had the assessment right during the reportable (80%) and critical phase (75%), but for the severe phase the results were mixed.

#### **Reporting boiling ditch**

##### *General*

Signal:	Water outflow
Crosscut:	Hinterland
Water velocity:	Slow
Water quantity:	Yes
Multiple:	No

##### *Reportable*

Length:	Very small
Width:	Very small–small
Soil flushing:	No

##### *Severe*

Length:	Very small–small
Width:	Very small–medium
Soil flushing:	Yes
Soil quantity:	Little

##### *Critical*

Length:	Small–medium
Width:	Medium–medium large
Soil flushing:	Yes
Soil quantity:	Much

## ***The Frustrating Grass Damage***

The failure grass damage only occurred twice, in Exercise 3 and 5 (both Region 2). It led to much frustration. For marking the failure, players had to place their report marker in a very limited area; otherwise, the game would not recognize they found it. After I realized this, I informed players about this limitation, during the start-meetings and in the weekly e-mails, but players still struggled with it. After many attempts some were able to finally mark it correctly; others gave up and continued their search for other failures or waited till the exercise ended.

To account for this, I corrected the game output by considering the “failed” attempts. This is not a complete correction. Players still devoted an unnecessary amount of time and effort into this failure—time and effort that they could have better spent on some of the other failures.

Most did not have problems with finding this correction, especially with the critical phase. The reason why scores are low for the second phase is that many did

not mention the grass signal. In Exercise 3 twelve participants reported this signal and 20 did so in Exercise 5, many fewer than the 81 participants who reported the overtopping signal in Exercise 3 and the 78 in Exercise 5.

Those few who mentioned the grass signal reported it well. Although some reported liquefaction or settlement, most reported grass and a solid majority (+/- 60–70%) had the length, width, and crosscut location right. The choice for the crosscut location is a bit ambiguous, because it starts at the crest but is largely on the inner slope. With this in mind, it is not surprising that the other participants (+/- 30%) opted for the crest as crosscut location.

The exact reverse happened with the overtopping/waves signal. About 70% chose the crest, the correct answer, while the remaining 30% chose the inner slope. The crest is correct in this situation because the water runs over the crest and then onto the inner slope. The grass signal is foremost on the inner slope, which is why this is considered the correct answer in that case.

A clear majority (90%) spoke of the overtopping/waves signal; the remaining participants thought of water outflow. Of those choosing overtopping, about 10% thought they were witnessing waves, and the majority decided that the area was accessible and that the water infiltrated the soil in both failure phases. Both these reporting elements—about accessibility and water infiltration—are arguably hard to answer, and whether the game’s answers are correct can also be contested. At first the situation is clearly accessible, but when the water quantity increases most patroller guidelines prescribe it would be better to stay on a safe distance: the area becomes inaccessible. Yet, most players thought otherwise.

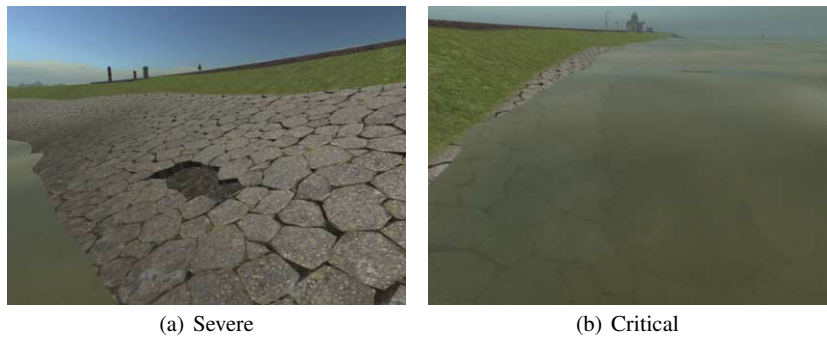
**Reporting grass damage**

<i>Severe</i>	
Signal:	Grass
Crosscut:	Inner slope
Length:	Medium–medium large
Width:	Medium–medium large
<i>Signal:</i>	
Crosscut:	Crest
Type:	Overtopping
Infiltration:	No
Accessible:	Yes
water quantity:	Little
Soil flushing:	No
<i>Critical</i>	
Signal:	Overtopping/waves
Accessible:	No
water quantity:	Much
Soil flushing:	Yes
Soil quantity:	Much

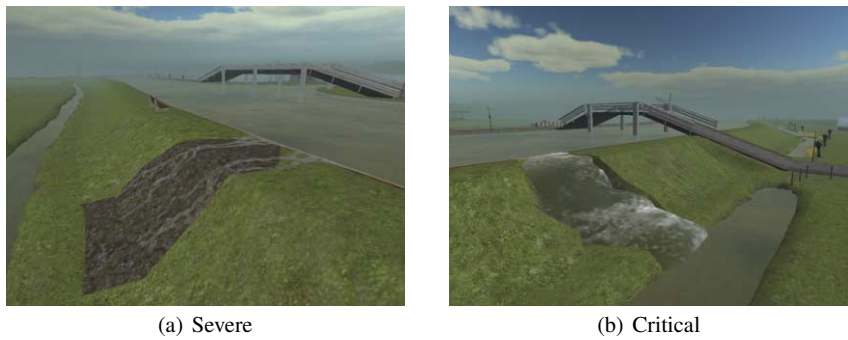
Infiltration is the process by which water on the ground surface enters the soil. This never happened. The water ran only over the levee and thereby damaged the revetment. However, I can understand that players may have pictured the failure differently. They may have thought that the resulting grass damage represents liquefaction of the inner slope. This is where we see—again—the limitations of the technology used. The game’s visualizations are not accurate enough to show the subtle distinctions.

What further becomes clear from the results is that about half of the participants thought the severe phase was already critical. Most also thought soil was flushing and half of these participants said the quantities were huge. At first even half of the participants said that huge quantities of water were flowing over the top! With the critical phase results were less mixed. The majority answered these reporting elements correctly.

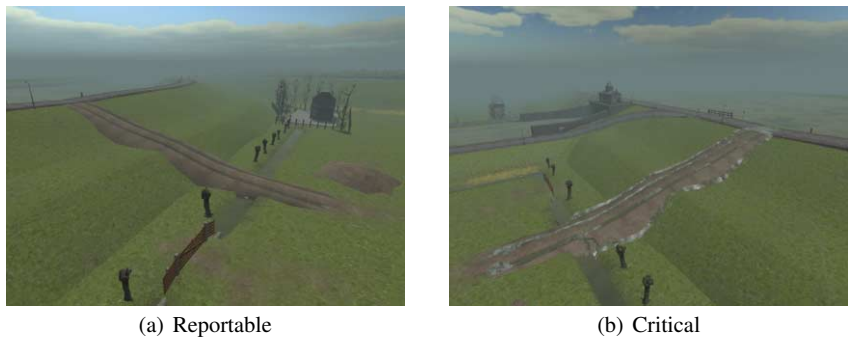
Most diagnosed correctly and took the right measure. The failure depicts the process of erosion of the inner slope. Unexpectedly, the most frequent alternative concerned macro-instability and the failure has really nothing in common with this



**Fig. 6.7** The phases of the stone damage failure. Notice the difficulty of seeing the critical phase



**Fig. 6.8** The phases of the grass damage failure



**Fig. 6.9** The phases of the illegal driveway failure

failure mechanism. Two correct measures could be taken. On average 62% decided to place sandbags at the top and 23% decided to cover the inner slope with soil.

*The Not-So-Surprising Illegal Driveway*

The illegal driveway is a failure in Exercise 6 and end-exercise (both Region 3). It depicts a temporary driveway from the hinterland to the crest of the levee. Such a driveway is by itself not problematic. In fact, some patrollers pointed out that if a permit is given and the season is right, such a driveway should not be a problem at all. However, in the game players inspect with the possibility of highwater. If water runs over the levee, the levee may quickly degrade with the presence of a temporary driveway. Normally the grass protects the levee from eroding, but this grass is not there anymore. They should report a temporary driveway for this reason.

The failure has two phases, a reportable and critical phase. The reportable phase has two signals. The first is the driveway, which should be recognized as a human activity. The second is a small settlement on the crest. The idea behind this is that heavy trucks made intensive use of a part of the levee and this caused a settlement. This settlement is difficult to observe in the game. Players can only see it if they look at the levee sideways.

During the critical phase, water starts to run over the levee. This means that the signal overtopping should be reported. Because the driveway is built out of sand, players should report that this overtopping is paired with much flushing of soil. The game does not visualize this. It is an assumption based on what is shown.

At first I thought that players would not recognize the failure. Unlike the other failures the illegal driveway looks like it is part of the scenery. It was not introduced until Exercise 6, so participants may have thought that by then they had seen all possible failures already. Eventually, the majority did end up finding the failure, although about one-fourth only did so when the failure became obvious: when the water started to run over the levee.

I think two important reasons account for participants finding this failure. The first relates to gameplay. Players wanted to find new failures. This is a response of one of the players after the sixth exercise:

I am happy to see a new image. Too bad not all failures from the handbook were used. I would have very much liked to send sheep of a barren levee, close down a pond in construction, or remove heavy trees from the outer slope et cetera. Many opportunities remain possible. My general impression is that this game is a great initiative to educate patrollers, but that to achieve a better learning effect the exercises need to be more diversified. This increases the motivation and enlarges the view on reality—GQex6-#138

**Reporting illegal driveway**

*Reportable*

Signal:	Human activity
Crosscut:	Inner slope
Type:	Earthwork
Signal:	Settlement
Crosscut:	Crest
Revetment:	Asphalt
Direction:	None
Length:	Very small–small
Height:	Very small

*Critical*

Signal:	Overtopping/waves
Crosscut:	Crest
Type:	Overtopping
Infiltration:	No
Accessible:	Yes
water quantity:	Much
Soil flushing:	Yes
Soil quantity:	Much

The other reason is mentioned by Participant #138 as well. All possible failures were mentioned in the handbook. Careful examination of the handbook by players would establish which ones they could find. Among these was an image of the illegal driveway.

During the reportable phase participants reported either the human activity and/or the settlement. Most reported the human activity (about 85%) and fewer the settlement (about 54%). Although less people reported the settlement, I am still surprised about the number of players who did mention it, because of the difficulty of seeing it. I think that players were able to assume that the levee would settle. About 15% of those reporting human activity also mentioned that it concerned traffic. A greater number (31%) stated changes were made to the levee. About half gave the “correct” answer, which is earthworks.

Regarding the settlement, most (81%) indicated correctly that the settlement was over the entire levee and had asphalt as revetment (93%). Measurements were less accurate. A small majority (64%) was still correct with the length, but the height was spread out over many answers. Apart from settlement and human activity a few reported grass. This is not necessarily wrong. The grass is damaged by the driveway. It would be better to report human activity. Human intervention is the primary cause.

Regarding the critical phase, nearly all (95%) spoke of overtopping. The others said it was water outflow. Although people had to assume the flushing of soil, a small majority (64%) said yes and of these, 80% said much soil was flushing. A similar number (83%) thought the water quantity was much too. In terms of accessibility and infiltration, the majority said yes as well (with 85% and 67%, respectively). For similar reasons as with the grass damage, no infiltration takes place. But contrary to the grass damage, the area remains accessible. The amount of water is relatively less compared to the critical phase of the grass damage.

Players properly assessed both phases. They did a better job with the critical phase (88%) than with the reportable one (75%). Surprisingly, participants had more trouble with diagnosing the failure. Many thought they were dealing with macro-instability. Yet, still 58% was correct with mentioning erosion inner slope. Similar to the grass damage, here too the sandbags on the crest (78%) and covering the inner slope with foil (14%) are considered correct.

### ***The Not-So-Many Non-Failures***

We purposefully inserted a number of non-failures (or “false positives”) into the game. Non-failures are not failures, but may be perceived so by players. This encourages players to think hard about what they see; otherwise any deviation may easily be signaled as a failure.

Due to time limitations, we did not implement many non-failures in the first version. We only included two: small puddles and molehills. Over time we observed that players found other non-failures. This decreased the necessity to include any other non-failures and so far this has not happened.



The data from the training confirms this. Participants classified a surprising variety of elements as a failure. Before elaborating, I need to stress that overall, participants did not mark many non-failures. Depending on what region they were situated in, either one out of three (Region 1) or one out of six participants (Regions 2 and 3) found one non-failure.

#### **In-depth explanation: determining the non-failures**

To find out about these “false positive” non-failures, I retrieved manually all X and Y coordinates of any non-failure from the game files and wrote down the first signal they reported. I noted up to three non-failures per player and wrote down non-failures that were deleted as those that were not. Per region I ordered the non-failure data to see patterns.

With the patterns in mind I walked around the regions in *debug mode*. This is a mode for designers to identify and remove any errors. In this mode various information is offered. Most important to determining the non-failures concerned the X and Y coordinates (measured in *Unreal Units*, the unit of measurement of the *Unreal Engine*) of the location of the player. With this information, I was able to walk to certain coordinates. Then I still had to decide what they reported. In many cases this was obvious. Based on player comments and observations I already had a firm idea of what players typically marked as a non-failure.

The number of non-failures did not markedly increase or decrease over time, although I did notice that participants were quicker in understanding whether it was a non-failure. In the beginning, not every participant made use of the statistics tool. With this tool players can easily see if what they found concerns one of the failures that they need to look for. If not, players deleted the report marker and continued their search.

I further noticed that a small group of players was responsible for most non-failures. One or two players marked about eight non-failures in just one exercise. Quantitative data confirms this: 18% of the participants marked in total more than three non-failures over all exercises. Marking of non-failures happened if players could not find a failure. Then they started to try out several possibilities. For others I felt it was their strategy of playing the game. They consistently had more non-failures than others. By placing many markers, they hoped to have a better shot.

The majority hardly marked a non-failure. About 36% never left a marker at a non-failure. They either deleted the marker right away or possibly never put one down. Moreover, 25% only had one non-failure over all exercises. Returning to the metaphor of shooting, this tells us that most saved their ammunition and only used it when they were quite sure it concerned a failure.

Another (very probable) explanation is that to most it was pretty clear when something concerned a failure and when it did not. What makes this idea very likely is that it did not happen too often that participants placed a marker and then deleted it. And about one third of the reported non-failures were not really non-failures, for three particular reasons:

1. A number of markers were *failure-related*. Players positioned the marker just outside the imaginary box of a failure. This happened especially in Region 2 with the grass damage failure, which has a very small imaginary box. Even if

**Table 6.4** An overview of the non-failures found over all exercises

Non-failure	Region 1 <i>n</i> (%)	Region 2 <i>n</i> (%)	Region 3 <i>n</i> (%)	Total <i>N</i> (%)
Puddle	125(43)	0(0)	0(0)	125(21)
Molehills	3(1)	4(2)	0(0)	7(1)
Sheep	3(1)	21(10)	1(1)	25(4)
Roadwork	14(5)	18(9)	44(42)	76(13)
Little boat	5(2)	6(3)	4(4)	15(2)
Pipes	8(3)	5(2)	1(1)	14(2)
Human activity	5(2)	19(9)	5(5)	29(5)
Unpolished elements	26(9)	50(24)	13(13)	89(15)
Failure-related	2(1)	61(29)	21(20)	84(14)
Overtopping	77(27)	2(1)	0(0)	79(13)
No idea	20(7)	24(11)	15(14)	59(10)
Total	288(100)	210(100)	104(100)	602(100)

*Note.* Only the first three reported non-failures by each player in every exercise were considered.

- players placed the marker next to the failure, the game would not recognize that they found it.
2. In another region (in Region 1), the water slightly trickles over the regional levee. This happens when the water level is raised to its maximum. Various players reported this, but in the game they only need to report local incidents and not something that occurs everywhere. The Action Center and others will keep track of the water level and other general events. Players reported this as either water outflow or overtopping.
  3. On many occasions, it was not apparent what the player had in mind to report. Possibly they used report markers strategically, something I observed during various playtest sessions. By placing markers players could easily teleport from one to another location. Others may have tried something and forgot to delete the marker. I labeled all non-failures that I could not relate to something as “No idea.”

Table 6.4 gives an overview of all non-failures. Apart from the aforementioned, basically eight types of non-failures could be distinguished. Among these, the puddle seems the most “successful”; the molehills the least. In the middle are some recorded from previous game observations: the sheep that walk around, the roadwork, and the little boat. Unlike the puddle, these latter three could in certain circumstances be considered a risk. Too many sheep will damage the revetment; roadwork increases the likelihood of erosion; and if a boat starts floating it may damage any levee revetment, so players had reasons to be cautious with these.

The three remaining types of non-failures were new to me. The first concerned the drainage pipes that connect the various ditches. Pipes and drainages are notorious for causing levee problems and so it should not be surprising that players reported them. One participant, an employee, explained his decision to report this:

At the previous exercise I did not mention that I thought it was strange that the drainage pipe could not be closed off. It not being closed off should maybe have been a report—GQex3-#61

One player even went so far to report a pipe that was hard to see. It was located underwater and far off from where most of the “action” in the game takes place. The second new type are several human activities, such as graffiti and improper trash disposal. The last type of non-failure combines all kinds of observations that are caused by some of the more unpolished models in the game. Some players thought that some of the grass represented cracks. Others reported a variety of signals for locations where the models deviated from the standard look of the game. This especially happened at locations where grass was missing and in areas where levee segments meet. Because Region 2 has many corners, there players especially reported such non-failures. One player stated:

With some images I thought to observe an uprise along the water line, but unfortunately that was not correct—GQex5-#97

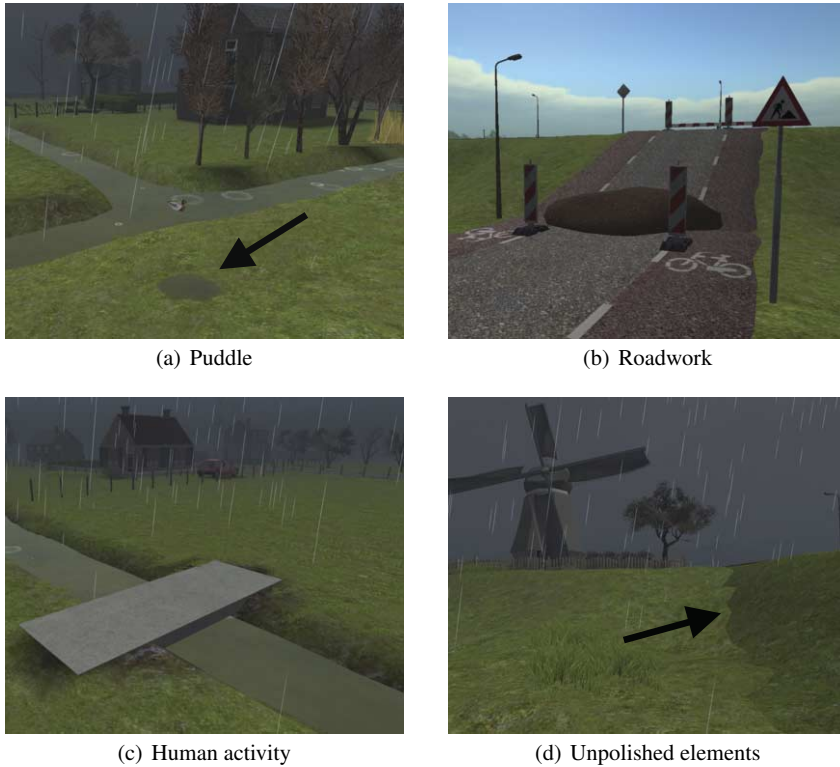
Because of this unpolished quality, players said the game was “too virtual,” meaning it did not appropriately represent the real world. In general, players were able to quickly pick up when an element looked somewhat strange but was not a failure. At the start-meeting I received many questions and after a while the participants already answered their own questions by saying “that is probably because of the game, right?”

Notably, the non-failures differed greatly among the regions. Most of these differences are easily explained. Region 2 has those temporary bridges; the other two have very few. If the water rises in Regions 2 and 3, it does not trickle over the levee as in Region 1. Then roadwork plays a more prominent role in Region 3, explaining why players reported this far more there. The puddles, on their turn, are prominently visible in Region 1. For sheep I do not have a simple explanation. All three regions have them and it is surprising that players reported these especially in Region 2. It might have to do with prominence of sheep in this region or that players had a harder time to find the failures and started to consider possible alternatives.

## Retrieving the Prevalent Gameplay Patterns

We discussed how participants performed in playing the game and how they dealt with the particular failures. What remains is to discuss my “game observations.” While analyzing the game data, I wrote extensive notes throughout. I wrote these down per exercise and per participant. Once I realized I was seeing a general pattern, something that more than one player illustrated, I wrote this down too. This is how I constructed what I consider the “prevalent gameplay patterns” and some of these may have already become clear after reading how players dealt with the failures.

After I established these patterns, I used the quantitative data for confirmation and support. I further retrieved help in understanding player behavior from an unex-



**Fig. 6.10** Four examples of non-failures that players reported. The arrows point to the non-failures

pected source: the *notebook tool*. This tool is part of the game’s main inventory and it allows players to write down notes while playing. We included this tool, because it was easy to add and we wanted to give players the opportunity to write down notes by typing.<sup>10</sup>

Many utilized the tool and their writings were useful for understanding what happened. It made for example clear to me that some had fun in playing the game, such as Participant #38 who made a note that he put garbage into a trashcan (GDstart-#38). It also made clear that some were frustrated and why. Participant #74 clearly disagreed with the Action Center about whether flushing soil occurred (GDex3-#74).

My final step was to group the patterns and it turned out I could identify four themes. The first is about their “performance” and the second about the “errors” they made. The patterns affiliated with these themes explain for the game and failure correctness scores. The other two themes are about how players played. In general I found players to be “serious, consistent, and persistent.” However, on occasions I

<sup>10</sup> The notebook tool also keeps automatically track of any measurements with the measuring marker.

noticed some remarkable gameplay behavior. I think this happened, because players were “focused and goal-driven.”

## *The Surprising Performance*

If we look at how players dealt with failures, it turns out—much surprisingly—that their performance is good, they improved little over time, and they were actually good from the start. Considering the game scores presented earlier in this level, my conclusion is that performance improved game-wise and not failure-wise. These are the patterns I retrieved and affiliated with the theme of “performance.”

### **Performance is good**

Table 6.3 shows that when players encountered a failure their overall performance was satisfactory. If we take 55% as the cut-off for sufficiency, then with each failure the players did well. They performed worst with the watery slope failure (56%) and best with the stone damage failure (77%). Although the performance does not differ too much among the failures, it indicates that it is dependent on the difficulty of the failure and less on how much players practiced with it. For example, players performed almost equally well with the driveway failure (67%) as with the small landslide failure (69%). They only had much less opportunity to practice with the first than with the latter (two vs. eight times).

The overall score is based on an equal share of four of the learning objectives: assessing, observing, reporting, and diagnosing. With assessing, it becomes clear that the performance differs among phases. During the reportable phase a bit more than two-thirds of the participants had it right. Except for the watery slope failure, only half had it right during the severe phase. This phase seemed to be most ambiguous. The critical phase was not. Apart from the boiling ditch phase, a clear majority were correct. A rough conclusion would be that in extreme situations players can estimate the situation. Anywhere in between, their estimates are as good as a guess.

With observing performance is excellent, unless a phase requires players to observe more than one signal. In those instances performance is still sufficient, but not as good as when players only have to report one signal. Performance also lowered in some critical situations. In fact, with the watery slope failure performance was dreadful (27%). I will elaborate on this later on, but the ambiguity of what to report and the fear of getting a levee breach are factors that may have played a role here.

The performance with reporting is insufficient too when a phase requires players to observe more than one signal. This is not surprising, because if players do not report a second signal, they cannot make a report for this and are only able to get a maximum score of 50%. To even get such a score, the one report they did make needs to be perfect. Also similar to observing is that the performance in the severe phases is lower than elsewhere.

In diagnosing, participants had no problems with the stone damage and especially the boiling ditch failure. The latter was expected because the failure mechanism term “sand boils” is widespread and well-known. The stone damage failure is the only failure on the outer slope, which explains why this did not pose too much problems to the average player as well. The remaining three failure mechanisms have been confused with each other. The confusion between micro-instability and erosion inner slope are known. We as designers sometimes still struggle with that. What I noted is that players had a tendency to call everything they had some doubts about macro-instability.

Players had no issues with taking measures. For each failure the performance is very good. It is noticeable that for failures with overtopping the performance is slightly better. That is no surprise either. The most well-known levee measure is to put sandbags on top of the levee and that is what in these cases the players had to implement. It conforms to a mental model that probably every Dutch citizen has of taking measures.

To conclude, it is apparent that a player’s performance is good when they were able to deal with a failure. The average game scores are not very promising (Table 6.1), but we can be reassured that on average patrollers seem to know how to make sense of failures.

### **Improvement is incremental and subtle**

On average the player performance on (almost) every failure phase for every failure improved, but only marginally. Nowhere is a drastic improvement noticeable and especially on phases and failures players only encountered twice. For failure phases, players encountered more than twice, performance incrementally changed with about 10% difference between the lowest and highest performance score.

Such a differentiation seems significant, but we cannot draw this conclusion. This requires statistical confirmation and unfortunately the number of participants who dealt with a similar phase more than twice is as little as two and as “large” as 48. This makes it difficult to rely on statistics. With so few people of the target group it seems unjustified to impute missing values. Statistical results (with a Kruskal-Wallis Test) with these few participants do highlight something interesting: independent of the number of participants all failure phases that have been encountered six times or more are considered significant. All failure phases that have been encountered less are insignificant.

We have to be careful in drawing any conclusions, but the current evidence seems to suggest that improvement occurs in subtle ways.

### **Performance was good from the start**

What we can be sure about is that player performance was good from the start. When players found a failure, they immediately dealt with it in an appropriate way,

which was unexpected because one assumes that over the exercises players become better and better. They did if we look at the game scores. However, if we consider the failure correctness scores, it shows that players were good at it from the start. This may explain why improvement was incremental and subtle. If performance was good from the start, it is difficult to improve. And although perfection is desirable, with the ambiguity in the visualizations and reporting systems it cannot be expected that this is fully achieved, unless players get to see the exact answers that they need to provide. That is of course unwanted. Players would then just learn the answers.

In a way, players could still practice this unwanted behavior, by making use of the statistics tool, and a number of them did. They used it in particular to find out what failure mechanism belongs to a failure. First they chose a failure mechanism, checked the statistics to see if they received points, and then chose another failure mechanism if they did not. This process continued until they were sure enough they had the correct answer. I expected this to happen. It is a trade-off we made for wanting to provide players direct information about their performance.

I can provide two complementary explanations for why players performed well from the start. First, it is not that hard. The failures are relatively easy to distinguish and find, as the numbers also show, and most signals are easy to identify. Second, the reporting procedure is completely automated and assists in reporting. The game's reporting system could be regarded as a decision-support system that helps players in reporting failures. Players still have to think and decide, but they do not have to think about what they need to consider. This thinking is done for them.

### **Performance improved game-wise and not failure-wise**

This warrants further explanation, because the average general game scores are, except for observing failures, rather poor, between 32% and 58%. The failure correctness scores seem markedly better. Only one in five scores is 58% or lower.<sup>11</sup> If the failure performance is good, why do we still see poor game scores? What also requires an explanation is that unlike the failure correctness scores, which turn out to be good from the start and subtly improved, the game scores improved significantly over the exercises.

Both differences are caused by exclusion. In calculating the correctness scores, players who did not find the failure were excluded. Missing a failure has however large consequences game-wise. It means that players do not get any points for finding, observing, reporting, assessing, diagnosing, and taking measures. Over time players missed fewer and fewer failures. This one reason why *a*) game performance is lower than failure performance and *b*) game performance improves and failure performance (hardly) does not.

Except for the total failure correctness score, players who missed a phase were excluded too and players missed a lot of phases. To highlight, with the stone damage

<sup>11</sup> In calculating the number of failure correctness scores below 58% I considered the assessing, observing, and reporting scores per phase and included the diagnosing and taking measures scores. This reaches a total of 52 scores of which 12 are below 58%.

failure players reported on average 1.60 phases out of 2; with the severe small landslide 1.61 out of 2; with the critical small landslide 2.07 out of 3; with the severe watery slope 1.39 out of 2; with the critical watery slope 2.10 out of 3; with the severe boiling ditch 1.08 out of 2; with the critical boiling ditch 1.69 out of 3; with the grass damage 1.27 out of 2; and with the driveway 1.51 out of 2. Over the exercises players improved in reporting phases, giving another reason why game performance did improve over the exercises.

Two other exclusions were performed on purpose. In calculating the reporting failure scores participants were only included if they reported at least one needed report. Otherwise the scores would not reflect player's actual abilities in dealing with a failure. In addition, for calculating the diagnosing and taking measures scores, participants were only included if they provided an answer. Many participants—for unclear reasons—forgot to diagnose failures or take measures. This did impact their game scores.

Because players only made subtle improvements in reporting the failures, improvement in game scores must have come from finding failures, reporting phases, and providing answers in diagnosing and taking measures. Thus, participants improved their performance game-wise and not failure-wise.<sup>12</sup> This does not mean players did not learn anything. These differences in performance indicate that players had to learn how to play the game—to make sure they find the failures in time, check them often enough, and realize what steps they need to make.

### ***The “Errors” Players Made***

A number of my other pattern observations speculated on why players made an “error” that caused them to not get a perfect score. It turns out that missing items and ambiguity play a major role. I speak of “errors,” because the resulting performances are likely an underestimate of player capabilities.

#### **Missing items and ambiguity explains imperfection**

Exclusion explains largely why the game scores are poor compared to the failure correctness scores. However, failure performance was not perfect. Some scores were even unsatisfactory. These are the contributing factors in order of magnitude:

1. *Missing a phase*: This is the foremost reason for why the total failure score are not perfect. Many participants did find a failure, but did so later in an exercise.

---

<sup>12</sup> An alternative explanation for the difference in performance has to do with game balancing. The regions in all exercises are of the same size and the number of failures and their severity were increased over the exercises. The first two exercises have two failures; the final two exercises have five. It is harder to find two failures than five in a region of a similar size. In addition, with fewer possibilities to report, incorrect answers have larger consequences.



This was especially noticeable with the two failures that overtop. Before the overtopping the failures were not noticed. On other occasions players skipped a phase or did not reach a failure in time to prevent it from causing a levee breach. On average, each phase was reported by 69% of the players.

2. *Missing a signal*: If a failure (phase) consists of one signal, missing this signal has consequences game-wise, not failure-wise. If a failure (phase) has more than one signal it does have consequences failure-wise. Scores for phases with two signals are dramatically worse compared to those with one signal. With the small landslide 59% of those who had the opportunity reported both signals; with the water slope 50%; with grass damage 39%; and with the driveway 34%.
3. *Ambiguous signals*: Some signals were clear-cut, such as the crack. Others could be interpreted in different ways. This made participants report them in different ways. However, one means of reporting was considered correct.
4. *Ambiguous reporting items*: Similar to signals, some reporting items were clear-cut and other were not. This forced players to guess, rely on assumptions, or use other cues to base their decisions on. These items are more likely to result in errors. For signals with many reporting items this has not so much consequences. Players have to get 70% or the items correct and so they can make one or two errors. For signals with few reporting items an ambiguous item will have much consequences.

The first two factors are about exclusion too. To see to what extent this made an impact, I calculated the scores for a restricted group. This restricted group reported all correct signals and reported at least 50% of all occurrences of a single failure. For the reportable phase this group reached an average score of 85% (regular group = 71%); with the severe phase 78% (regular group = 56%); and with the critical 79% (regular group = 68%). This means that about 16% of the scores is caused by missing a signal or reporting an incorrect one. About 19% is caused by ambiguity and random errors.

The last two factors relate to ambiguity, which I will discuss next.

## Dealing with ambiguity

I discovered that some players had creative solutions to deal with the game's ambiguity, such as to include all possibilities. To describe the critical phase of the watery slope failure Participant #107 decided to report liquefaction together with horizontal movement and settlement. This is an example of dealing with ambiguous signals.

To deal with ambiguous reporting items, players did something similar. They created different reports with the same signal. With the driveway failure ambiguity exists about what type of human activity it concerns. To deal with this, Participant #76 wrote down the two main possibilities: earthworks and traffic. He even adjusted the crosscut location for both of them accordingly. These creative solutions work,

because players are not punished for adding extra signals and reports. They are only rewarded for providing the right signals and reports.<sup>13</sup>

Much of the ambiguity is caused by how the failures are visualized. Due to graphical limitations, some of the processes are not shown. When the levee overtops at the driveway failure, players cannot see the soil actually flushing. They have to assume this. Similarly, with the stone damage failure, they could see stones vanishing but they will not see any flushing of soil. They again have to assume this happens. Other visual ambiguities are about quantities and in particular if it concerns “little” or “much” water outflow and/or flushing soil. I gave rules of thumb to the participants (e.g., one or two fists full of soil is little), but I understood many still had issues with deciding on the amounts.

What became clear as well is that occasionally players had a complete different interpretation of a visualization. What especially happened is that if the visualization was supposed to show soil on top of the levee, players interpreted this as a gap or hole. At the watery slope failure, where at some point two chunks of soil flow out, one participant noted “Two holes appeared next to the crack” (GDex3-#127).

Although the game’s ambiguity accounts for some of the errors, the majority of players frequently chose the correct answer, verifying our choices and assumptions. They seem to deal with the ambiguity just fine. The very best players ensured nothing was left in doubt and found creative ways to deal with its ambiguity.

### **Context is influential**

I further noticed that when ambiguity exists, players were influenced by the context in deciding what to report. With the context I refer to environmental cues surrounding a signal or the state of the failure situation—whether the situation is reportable, severe, or critical. If cues or the state tend to be worse, players might be persuaded to report other items as worse too. So it happened that players assumed the crack at the watery slope failure became larger after soil started flushing. This crack only never got larger.

The example of the assumed larger crack was incidental. Only a few participants did this. More evidence is provided by the boiling ditch failure. This has two ambiguous reporting items, the water velocity and quantity. Although the failure becomes increasingly worse, both items remain the same throughout the development of the failure. We decided that the first is slow and the second is little. Over the exercises the majority (63%) chose slow during the reportable phase. Much to the contrary, a firm majority (74%) chose fast consistently during the critical phase.

The pattern repeats itself with the water quantity. A majority (69%) chose on average little as answer for the reportable situation. The same firm majority (74%) opted for much during the critical phase.

<sup>13</sup> The solution of adding a similar signal more than once was not possible with the version Organization C received. They could however add as many reports to one signal and in this way players were still able to creatively deal with ambiguity.

Many participants probably made these decisions because they assumed that if more soil flushes, the water velocity and quantity increase too. This is not necessarily a wrong assumption; the game only did not visualize this. Admittedly, the game did not visualize many other things too and in those instances players had to make assumptions as well. Nevertheless, I want to stress that player decisions can be influenced by the context, especially if the choices are ambiguous.

### **They know it but do not do it**

The notebook tool provided evidence that some participants found failures and knew very well what to report, but did not know how. At the first watery slope failure one participant noted “soil is flushing” (GDex1–#84). He did not report this however. At the same failure two other participants noted the puddle (GDex1–#74 and GDex1–#45). The first (#74) also did not report this; the second (#45) reported liquefaction instead of water outflow. A fourth participant (GDstart–#4) even noted the correct signal for observing the stone damage failure but chose another.

Although these latter participants did not report the failure perfectly, they at least reported it. Some others did not even get that far. They only used the notebook to report their findings. Participant #48 wrote in Exercise 4 for example about parallel cracks on the crest of a levee that become larger at some point. He also spoke of a settlement. It seems that this participant clearly witnessed the development of the small landslide failure and is able to describe this, yet he failed to report this. With Participant #146 this happened as well, but then with the grass damage failure:

Close to the water-basis, water flows over the levee. The situation is very critical...[and later]...The situation gets out of control. The polder is getting filled with water—GDex3–#146

Important to add is that Participants #48 and #146 have little computer literacy. At the start as well as end-meeting they showed considerable difficulty with playing the game and also told me so. Others who used the notebook as a replacement for placing report markers tended to have issues with computer literacy. They used the notebook tool as alternative for reporting the failures.

The notes show that the game scores probably underestimate the actual capabilities of the participants. In addition, they show that some participants still had to learn how to report what they see.

### **There is (almost) always a logic to it**

The game’s scoring system is very rigid. In certain cases multiple ways of reporting are possible, but the current scoring system does not allow this. This rigidity implicates that if participants did not give the desired response, they did not have to be necessarily wrong either. What I discovered is that in many cases players were not necessarily wrong. There was (almost) always a logic to what they reported.

Consider for example the decision by Participant #45, who noticed the puddle at the watery slope failure in Exercise 1. In his note he said “Also water outflow close to the ditch.” However, in addition to the crack, he reported liquefaction instead of water outflow. Later when the soil starts to flush, he creates a second liquefaction report indicating that soil starts to flush. He therefore captures this phase as well—but he did not get any points for it.

Although water outflow remains a better choice, it could be argued that liquefaction would be a better choice, because *a*) no flow of water is noticeable; and *b*) the failure leads to liquefaction in its critical phase. So this player’s choice has some logic. In many other cases I could find some logic for why participants decided to report a failure in a certain way. Other common “mistakes” are using water outflow instead of overtopping and interchanging horizontal movement and settlement.

This does not mean no illogical reports were made. For example, one player (GDex1–#63) decided to report the little puddle as overtopping. It is not completely impossible, but a rather unlikely conclusion. Although I encountered more of such mysteries, players reported illogical signals especially when they became desperate. If the Action Center did not accept a report or a measure could not be taken, players started to experiment with what to report.

### **You cannot teach what is hard to observe**

Another reason why errors are made is that the game suffers from *vagueness*. Something is vague when it is difficult to form any interpretation. I knew two types of vagueness and I gave hints to players on how to deal with these. The first concerns the reporting of the first two phases of the watery slope. Here two signals need to be reported and to report them correctly, players should understand that the puddle of water is a result of a flow of water from the crack. Because I noticed after the sessions with Organization A that players had difficulty with this I showed the participants of Organizations B and C on how to report this failure.

This explanation may have made some difference. About 86% of players reporting both signals were from either Organization B or C, whereas they account for 75% of the sample population. Independent of their organizational affiliation the players who reported both signals did not continue to do so in the future. On average, 47 participants reported every time both signals and only 21 did so for three exercises. Exactly 13 participants reported both signals for all exercises. Although other factors may play a role here, it seems like it is impossible to teach what is hard to observe. Most participants end up reporting the crack and forget about the water outflow.

The other vagueness is the second, severe phase of the boiling ditch failure. If participants looked carefully, they would have seen an accumulation of soil at the bottom of the ditch. This is hard to see, which is why I provided several hints to report it. Nevertheless, of those finding the failure only one-third reported this phase (35%) and less than half (45%) noticed the flushing soil. The others reported it as if it were in its reportable phase. Those who already found it could not see a difference:

See page is the same as....one hour ago—GDstart—#22

In addition, here few participants reported this phase many times. In fact, only six participants reported the phase three times or more, showing that with this vagueness players did not stick to what they reported before too. This is remarkable, because generally I found players consistent in what they reported.

### ***Serious, Consistent and Persistent***

The meticulousness of play surprised me; they evidently took this game very seriously. Other prevalent gameplay behavior is that players based their reporting decisions strongly on what they decided before. They were also strongly convinced about what needed to be done and persisted in what they thought was the right thing to do.

#### **Players are serious**

I expected that sooner or later players would start to become less precise, reckless, and sloppy with filling out the reports. They never did. Over all exercises reporting items with obvious answers were filled out by everyone near to perfect. As for obvious answers, you can think of the type of revetment and the failure's crosscut location.

Another unexpected indication of their seriousness is that a great number of players kept on measuring the failures. Sizes of failures are not dynamic and so essentially, if players would measure it once and remember the size, they would not have to measure it again.

The occasions when players were "sloppy," it was not their fault. Although players frequently forgot an item, in helping the participants during the meetings, it became clear that players thought they clicked on an item, but they did not do this accurately enough and so the computer never received their answer. Because players did not read the subsequent warnings, they never understood why they could not proceed and why the Action Center gave them the same response over and over again: "Your report is incomplete."

These user interface problems with reporting may explain some of the players' frustrations discussed earlier with the mouse pointer and the interactions with the Action Center among others.

#### **Players are consistent**

Players tend toward consistency, generally staying with their initial choice the first time they encounter a failure. This is a common human tendency and error. We like to sit where we sat before—in classrooms or meetings. We also go by what works.

I further observed that if players were inconsistent, they were so with the more ambiguous reporting options. For example, it is difficult to judge if the water velocity of the boiling ditch failure is fast or slow. Another ambiguous choice is if the surrounding environment of a failure is accessible. Not knowing what to choose players may have answered such ambiguous reporting options randomly and this very likely increased the possibility of choosing another answer on another occasion.

It also became clear that if players changed correctly, this often did not happen until much later in the training. Participant #102 consistently reported uprise as signal for the boiling ditch failure until finally, at the end-exercise, he reports water outflow. This is an extreme example, but the game data shows that players *a)* stick to what they choose first; unless *b)* the choice is ambiguous; and *c)* they encounter the situation many times and realize (somehow) they need to report something else

### **Players are persistent**

Players are not only consistent; they are persistent as well. This became especially clear if players were convinced a measure was needed. In those instances players could not think of the possibility that a failure may not need to be taken care off. They neglected feedback from the levee expert and Action Center and continued to try out various signals and measures. Some were even so persistent that they tried out five different signals and eight measures. When confronted with the same situation in another exercise, they tried to take a measure again.

This persistence may arise from a firm belief in the need to take a measure. This happened especially with the severe phase of the small landslide and grass damage failure, which appear quite critical and in practice measures would probably be taken in these failure stages also. However, it highlights that players have such strong beliefs and are hard to convince otherwise.

### ***Focused and Goal-Driven***

In addition to the previously discussed player behaviors I found a number of remarkable gameplay patterns that illustrate player behaviors that are different from how the game is supposed to be played. Unlike the tendency of being serious, consistent, and persistent, these patterns are not displayed by the complete sample population. What the patterns share in common is that participants became increasingly focused on certain aspects of the game and were driven to prevent a levee breach at all costs.

## Being pragmatic

An exception always exists and so too with reporting logically. One somewhat illogical practice was done at a larger scale: to make it large. In total 18 participants decided to report the critical phase of the small landslide in particular and also the watery slope failure as a large crack. Four of these participants did this more than once. Besides making a larger crack, some players decided to make larger grass damages before taking a measure. It requires some imagination, but in some way the situations can be seen as a crack or grass damage, which definitely stretches the meaning of a crack or grass damage. They are not the most likely choices.

A significant number of participants engaged in such behavior. It might be that these players already made a report of a crack or grass signal and it would require them less effort to add another report than to create an entirely new one. Although they may have opted for this reporting procedure out of easiness, I suspect that they did it so out of efficiency. They encountered these failures in their critical phases and wanted to achieve the end-result, taking a measure as soon as possible and with whatever means available.

Another confirmation of players being pragmatic in times of danger are the measurements. Unlike other phases, no majority had the correct answer. In fact, answers were pretty much randomly chosen. Most of the times sizes were large or very large. Apparently players did not want to think this through or take the time to measure the sizes. This pragmatism seems to contradict the seriousness that I discussed earlier. There I mentioned that players were very precise and accurate. Apparently they were not always precise and logical. I think players traded these values in favor of being pragmatic. All of the 24 occurrences happened in the last three exercises where players had to deal with many failures and a levee breach could very likely happen. They would have rather prevented a levee breach than fill out the reports perfectly.

## Is it paused?

Pragmatism seems a plausible explanation for the “make it large” and “random measurement sizes” patterns. It is only a bit odd that players needed this pragmatism in a game that is paused while players are in the menus. If players enter a menu at the bottom it says in red “paused.” The clock also stops ticking. Because of this pausing, players had all the time to report what was needed. We wanted to give players this opportunity, because in the end the game is not so much about saving the region but about learning how to deal with failures.

The “make it large” and “random measurement sizes” patterns highlight that pausing the game did not completely succeed in this. Another pattern confirms this too. This is the strategy to first find all failures and then start reporting them, which I discussed in Level 5. This is a poor strategy because by the time players find all failures, some failures have likely changed and the player is unable to report the earlier phases.

These patterns show that players were not aware of the game being paused when they were in any menu. Aside from a possible bad user interface design that caused this, it may be based on bad experiences. More than once I saw a player leaving the reporting menu to measure or take another look at a failure and right in front of them the levee started to breach. The game may have been paused, but if the levee breaches right after leaving the menu it does give the (unjust) feeling that the game was not paused.

I believe that it is especially caused by the imagination. Imagination in games is very strong. To play the game, players have to create the belief that they are put into a dangerous situation where a flooding could occur at any time. To imagine such a situation, it is hard to switch to a “reality” where in the heat of battle, players could just take their time and not worry. The game has therefore a tension between this fast-paced quality and this more educational mode where time is given to think things through. It seems that some players were unable to make the switch. The imagination wins. The game scores and learning lose.

### **Reporting after the fact**

Another pattern confirms the hypothesis that players felt they were in a rush or had an urgent need to take care of things. Quite frequently it happened that after the fact—when the measure was taken—players reported the correct signals and/or added other signals. One of the players (GDend-#38) who reported the large crack when observing the small landslide failure reported a settlement after the measure was taken. This settlement is the correct signal. Another player (GDex5-#130) first took a measure at the grass damage failure to prevent the overtopping and then later reported the grass signal.

Reporting after the fact is rewarding. Players do not get points for reports from previous phases, but they do get points for reports belonging to the critical phase and all relevant signals. Considering this, first taking a measure and then looking at the situation is actually a smart idea. Players do not need to be worried about a levee breach and are still able to get points. The measures do cover up some or all of the signals, so it is only a proper strategy if the player already knows what to report. I further stress that this most certainly does not explain the “make it large”-phenomenon, because not all 18 players reported like Participant #38 something else after taking a measure.

Players are smart. They will try to find or adopt strategies to improve their scores; this reporting after the fact may be one such strategy. In this situation I only suspect it is more driven by fear of having a levee breach than a conscious strategy. I base this on all other observations that indicate that players like to take care of things.



### **One at a time**

Failures in a 3D environment allow players to look at failures from multiple angles and, therefore, get a more complete mental model of them compared to looking at a picture. Up front it was made clear to the participants that failures could consist out of one or more signals and I thought that players would ultimately look for a second (or possibly third) signal immediately after they found the first. This barely happened.

What did happen is that the majority of the players who reported more than one signal reported the signals one at a time. The time in between these reports is significant. Players first walked away from the failure and later, when they returned, they reported another signal. This happened with the watery slope failure. Player would first see the water outflow and later report the crack or it would happen the other way around.

It happened with the severe phase of the small landslide failure too. Here a frequent pattern was that players first noticed the crack, indicated that the situation became worse, and then much later discovered the horizontal movement. At that point they decided the situation is critical and they wanted to take a measure. Both signals appeared at the same time, but players may have thought that the horizontal movement occurred much later.

This behavior did not change, showing that the game did not teach them the desired behavior of exploring the failure region to look for more signals. It seems the opposite occurred: players acquired a very narrow focus on failures. Such a narrow focus may have been fostered by the gameplay. Players knew more failures were waiting around the corner, so after having dealt with one, they may have felt it was better to move on.

What strengthens the idea of having a narrow focus is that the percentage of players reporting both signals went down after inspecting a situation more than twice. During the first two exercises in which players could report the severe phase of the small landslide failure this percentage came down to 71%. It averaged to 52% for all other exercises. With the watery slope it went from 58% during the first two exercises to 45% for the remaining three exercises.

### **Players develop expectations**

Although the game did not teach players to look for other signals, they seemed to increasingly know what will happen. They developed failure expectations. One clear example concerns the stone damage failure. In Exercise 4 players were confronted with two of them, one who becomes worse and one who does not. Players realized at some point that these failures could become worse and because they could not clearly see if this happened, they reported both as if they became worse.

With other failures, similar types of failure expectations could be observed. Exercise 5 includes two watery slope failures. One becomes critical; the other remains reportable throughout the exercise. At this latter failure one participant noted "I see

one crack and no flushing soil” (GDex5-#54). This note highlights that this player expects flushing soil, which would be the case if the failure became worse.

My most favorite example concerns the cracks with the small landslide failure. In the reportable phase players see two cracks equal in length. They need to report one of them and indicate in this report that multiple cracks are occurring. During the severe phase these two smaller cracks become one. In the last two exercises a number of players kept this in mind and decided that it would be better to report one large crack from the start. They expected those two cracks would become one and the same and acted accordingly.

Too much reliance on expectations is not good. If failures develop differently, automatic responses are not appropriate. But having expectations about failures at all is desired and something this game aims for. It gives players a certain focus on the object of their attention.

## Lessons Learned

Although on average player performance was sufficient, the variety in scores among players is enormous. It reaches from one extreme end to the other. On average, players improved over the exercises—but not dramatically.

When decomposing the overall score, it becomes visible that players improved above all on diagnosing and performed generally poor on reporting, which can be partially explained for by the many dependencies this score has with others. For example, if players do not find a failure, they cannot make any reports and, therefore, receive a low score on reporting.

However, regarding finding failures players had little trouble (except for the boiling ditch failure). They reported also surprisingly few non-failures. If we look at the individual failures, we encounter different results, also per learning objectives and failure phases, highlighting that differences in difficulty exist, certain failure features are better implemented in the game, and/or certain failure features are easier to learn. For example, it becomes clear that players were very good in taking measures and that they had the most trouble with the watery slope failure.

If we consider the “failure correctness scores” instead of the game scores, which are the scores of how players performed on the particular failures, we find—much surprisingly—that players already performed well from the beginning. An explanation concerns the reporting procedure, which structures the reporting process and helps players to identify what they see. In addition, what players do “wrong” according to the game is still often a plausible choice. The game only allows for one correct answer, whereas in reality many more answers could be possible. Players made further errors with ambiguous signals and reporting items.

Players also engaged in activities that we designed to prevent or hoped to encourage for, such as that they forgot that the game was paused and that they did not consider reporting second signals. Although player behavior was often logical, occa-

sionally it was irrational and this happened noticeably in critical situations. Possibly players favored pragmatism in dealing with the failures over an accurate report.

However, in general players were meticulous with their reports. They took this activity very serious. This is something I already noticed from the results in the previous level, but when delving into how players reported the failures, it became clear that players did this very precise.

## Level 7

# Knowing the Pen-and-Paper Generation

*When it storms I go on purpose to the beach. The raging water...I find that wonderful to watch—IPpre-#5*

*I am of the pen-and-paper generation—IPpre-#5 again*



50%

This level is focused on exploring in-depth who these levee patrollers really were—their background and preferences—and what they thought of levee inspection and the game-based training. Such a picture is needed to get a full understanding of what happened during the training—why it worked or did not.

I asked participants to respond to various questions and statements on the pre- and post-questionnaires, which were filled out during the start- and end-meeting, respectively, and which were used to retrieve baseline characteristics and perceptions regarding games in general, *Levee Patroller* and the training in particular, and levee inspection above all. Before and after the training I also scheduled a number of interviews (Level 10). In between, I had often an informal talk with the participants about their work and motivation too, giving me additional insights into the lives of these levee patrollers. This level will, however, focus on the questionnaire results.

It is structured by focusing first on participants' responses on these questionnaires. The next step in this level is to reveal how the responses relate to each other. This allows us to see what variables are influential in this training. The third and final step is to consider a number of characteristics that may have influenced the outcomes of the questionnaires.

The goals of this level are to describe

- Participants' characteristics and responses on the pre- and post-questionnaires;
- How responses relate to each other; and
- How responses differ among a number of sample characteristics.

## Retrieving the Questionnaire Responses

One of the purposes of the pre-questionnaire was to garner information on a number of basic characteristics. This should answer the question “Who are these levee patrollers?” The pre-questionnaire then moved on to their preferences and expectations: What do they like or do not? How motivated are they? Do they expect to learn from the game? The purpose of the post-questionnaire was to see if some of the expectations were met and to retrieve participants’ reactions about the training and the game.

Both questionnaires consisted of two identical parts. One is about inspection perception; the other about knowledge perceptions. By measuring the participants’ rating before and after the training I could determine if the game influenced these perceptions. It is important to note that I did not intentionally built any scales or test the questionnaire up front. Consistent with the exploratory character of this research (and the evaluation principle “See the Big Picture”) I decided that breadth was more important than depth. By creating items of many variables I would achieve a much broader image of the training.

Another rule of thumb that I applied in developing the questionnaire is that if an item measured two directions, I constructed a 7-level item, that is an item with seven rating choices.<sup>1</sup> A typical disagree/agree is an item that has two directions: one can disagree or agree (to a lesser or greater extent).<sup>2</sup> I opted for a 7-level item in such instances to be able to fine tune to what lesser or greater extent a participant rated an item within one of the two directions. With fewer levels the response becomes more black and white, either the participants agree or disagree. More than seven levels seemed unnecessary and would have been a greater burden on behalf of the participants in filling out the questionnaires.

Whenever an item has only one direction I deemed it a 5-level item, that is an item with five rating choices, to be sufficient, especially because with one direction it is unnecessary to include a “neutral” or “neither disagree or agree” option, an option which I did include with the 7-level items. With five rating choices enough variation exist to determine to what lesser or greater extent participants position themselves. An example of an one direction item concerns the commitment item. Participants could opt for being not committed, somewhat committed, fairly committed, committed, and strongly committed.

---

<sup>1</sup> I could not apply this rule of thumb with the game questionnaire for usability reasons and had to use 5-Level items (Level 5).

<sup>2</sup> Direction does not equal dimension. The typical disagree/agree statement has only one dimension, that of agreeableness, but it is defined with two different directions (or two different labels that each represent one end of an imaginary continuum).

## ***The Patroller Basics***

To get an understanding of participants' background let us start with the basics or the "baseline characteristics." I asked the participants to provide information about their age and occupation and asked them to indicate their involvement and experience among others. All variables are described in Table 7.1. This table also shows to what extent responses are similar or different between the three organizations.

The baseline characteristics are based on all (pre-)questionnaires I received—from those who I identified as participants as well as non-participants (Level 4). For (un)known reasons the questionnaire is missing for some participants, explaining why the total number of participants listed is 145.

### **They are relatively old**

Games are often (no so accurately) associated with younger people—children or adolescents. With an average age of 47.6 years ( $SD = 12.1$ ) this target group is much older than the conventional image we have of gamers. I divided the participants into four groups.<sup>3</sup> I noticed that among participants younger than 40 years old, the ages differed widely. This concerns the first group and is almost one-fourth of the total participants. Many of the younger participants were employees and some were children of other volunteers. The youngest participant was 18 years old. He was still a student at school.

I divided the majority of the participants into two age groups, one of between 40–50 years and the other 51–60. Everyone above 60 years comprises the fourth group. The oldest participant was 67 years old.

It is reasonable to conclude that this target group is relatively old. Although no difference is to be noted between the organizations, the employees are younger,  $t(143) = 3.25$ ,  $p = .001$ ,  $r = .26$ .

### **It is a male world**

If it was not clear by now I want to confirm it here again: it is a male world. Only four females participated, and their numbers were too low to make a gender-based comparison possible.

---

<sup>3</sup> I divided the participants in four age groups to prevent outliers from influencing the results and to make them easier to interpret. Using a three-way factorial ANOVA I checked if this categorization had an influence. It did not.

## Education is mixed

Levee patrollers are generally considered relatively less educated. It is typically seen as a job performed by vocationally trained (or self-educated) people.

Although a good amount (22%) did not finish anything higher than their secondary education, I was surprised to see that over one-third achieved at least the equivalent of a Bachelor's degree. About 4% even obtained a Master's degree. Assuming this group is representative, we should therefore correct the image. The results show no differences between the participating organizations and between volunteers and employees.

In a random discussion during one of the meetings Participant #4 pointed out that patrollers are a mixed bunch of people—also educationally. He told me that at another water authority even a full professor was inspecting the levees. I am sure that this professor is not one of many, but he was most certainly right in telling me that the level of education is mixed. This was confirmed during one of the interviews:

*IPpre-#79:* When we talk about my levee segment, we talk about a mixed company. We have people who know nothing at all. Many senior citizens. Those just have a fun outing. They are enthusiastic, but they do not have much baggage and you cannot teach them much. You also have some young academics. Those are super interested in all kinds of techniques and they want to know everything...Others come as a sort of therapy...

*Casper:* A sort of therapy?

*IPpre-#79:* Yes, we had one who was sick for a while. He felt better and he had to get among people. So he joined the levee inspection.

## Occupations are diverse too

Although formerly a job done by above all farmers and horticulturists, today's volunteers have much more diverse occupations. Urbanization may have played a role here and in case of for example Organization B and C, the need for fresh and especially more volunteering blood too.

Besides working in agriculture, horticulture, or stocking cattle, many patrollers also come from construction and engineering. Most do so with a focus on water and the environment. One of the participants (#74) is, for example, a mechanic for road and water construction machines. Another (#47) is a manager at a dredging company.

The remaining volunteers I could not fit into a large enough category. I put them together into a category named "other." The jobs within this category are varied, from municipality clerks, teachers, and journalists to architects, mailmen, and security officers. Although the group is varied, two subcategories became clear. The first concerns government and education. The second is unemployment or retirement. Both are similar in size: about 8% of the total sample population falls in either category.

Unlike the previous variables, this one does differ among the three organizations. A difference with Organization C could have been expected, because there 64% are employees and that is much more compared to the 16% and 17% of Organization A

and B, respectively. Between these latter two a difference exists too. At Organization A the more traditional occupations appeared. Many participants are horticulturists. At Organization B many more jobs in the “other” category appeared.

This difference is a result of the regions the organizations are located in. Organization A is located in a region with many greenhouses which are more at risk for flooding, so historically horticulturists have been involved with inspection. It is a result of the recruiting at Organization B as well. Because they needed many volunteers, they recruited on a large scale from all layers of society.

### **Commitment is low at Organization A**

In game-based training, commitment to the subject of interest may influence the end result. Commitment tends to correlate with motivation, which in turn correlates with time and energy invested in the game. It seems that the commitment of the participants was adequate, because most answered they felt strongly committed. But one must be wary of a possible social desirability bias here. Participants will be inclined to answer this item positively, even if their responses are processed anonymously.

A difference between organizations is noticeable again. Organization A's members felt less committed than those of the other two.<sup>4</sup> Knowing the situation at Organization A, this is not surprising. Little was done with the volunteers in the past years. This part of the levee inspection organization was even nominated to be abolished (Level 4). If the organization is not committed to its members, its members will feel less committed too.

### **Only about half have failure experience**

I also asked about what experience participants had with failures. If so, they had to write down what they encountered. From this, it turns out that approximately half of the participants (48%) had no experience at the start of the training. The number of inexperienced participants was fairly higher for Organization B and that is because many of their participants recently joined.<sup>5</sup> Some were not even a member for a year.

This becomes clearer when we consider the experience in years. Compared to the average at Organization A ( $M = 17.7$ ,  $SD = 14.6$ ), the average experience at Organization B is much lower ( $M = 1.89$ ,  $SD = 1.31$ ). Organization C sits in between, because although it has many experienced volunteers, it likewise recruited

<sup>4</sup> Post hoc tests reveal that a statistically significant difference exists between Organization A and B,  $U = 679$ ,  $p < .001$ ,  $r = .40$ , and between Organization A and C,  $U = 192$ ,  $p < .001$ ,  $r = .60$ .

<sup>5</sup> A difference in failure experience was noted,  $\chi^2(2, N = 145) = 7.43$ ,  $p = .024$ , Cramer's  $V = .23$ . Post hoc tests reveal that this difference is due to Organization B. No difference exists between Organization A and C.



many volunteers in the past years.<sup>6</sup> This experience in years does not say much. One of the participants (#14) has been a member for over 47 years and has never encountered a failure.

Those who did encounter one mentioned especially having seen cracks, settlements, water outflow, or erosion. Cracks are seen more often at Organization A, settlement at Organization B, and water outflow at Organization C. Erosion is encountered about equally at all three organizations.

What struck me in analyzing the results is that participants from Organization B and C used already some of the more difficult inspection terms, such as macro-instability and micro-instability. None at Organization A did.

### **They do not have much game experience too**

About half of the participants rarely play games, whether analog or digital. Differences between playing analog and digital games are evident. Most remaining participants play only a couple of times a year an analog game, whereas one-fourth of the participants play digital games weekly or even daily. So those who do play digital games, do so often. It also turns out that participants at Organization A play more analog games compared to those of Organization C,  $U = 404$ ,  $p = .015$ ,  $r = .29$ . With digital games no differences exist.

Playing games often could make a difference. They may enjoy it more and become more accustomed to it, making it easier for them to pick up a new game. For the same reason I specifically asked if participants played a First-Person Shooter (FPS). This is a genre that has much in common with the *Levee Patroller* game. The controls and viewpoint are for example completely identical. Experience with this genre will give an extra benefit. About three-fourths indicated that they never played a FPS in their lives.

Of course, having previous experience with playing *Levee Patroller* could make a difference too. Here too about three-fourths did not play the game before onset of the training. Only this time we see that the participants at Organization A have much more experience compared to the other two: more than half already played the game! Just after the release of the game Organization A organized on their own behalf sessions with the game. After that they hardly organized anything.

### **And they have little computer skills**

In playing a game on a computer it helps to have played such games before. This is what is referred to as game literacy (or ludoliteracy, see Level 3). Similar to reading and writing, playing games requires skills. Computer literacy—having the skills to use a computer—might very likely make a difference too. Only about 30% consid-

---

<sup>6</sup> Because of an unfortunate error in the data collection procedure I was unable to retrieve the experience in years numbers for Organization C. Based on the interviews and talks with the participants I am confident that the average sits in between Organization A and C.

ered themselves skilled or very skilled and this number does not surprise me with my experiences in helping the participants during the meetings (Level 4).

**Table 7.1** Baseline characteristics of the participants of the three participating organizations

Characteristics	Participants, <i>n</i> (%)				<i>p</i> <sup>a</sup>
	A ( <i>n</i> = 36)	B ( <i>n</i> = 76)	C ( <i>n</i> = 33)	Total ( <i>N</i> = 145)	
Age, <i>y</i>					
< 40	6(17)	20(26)	8(24)	34(23)	.39
40 – 50	10(28)	19(25)	14(42)	43(30)	
51 – 60	14(39)	22(29)	8(24)	44(30)	
> 60	6(17)	15(19)	3(9)	24(17)	
Gender <sup>b</sup>					
Male	35(97)	74(97)	32(97)	141(97)	—
Female	1(3)	2(3)	1(3)	4(3)	
Education					
Secondary education	12(33)	11(14)	9(27)	32(22)	.32
Vocational education	13(36)	34(45)	12(36)	59(41)	
Higher education	11(31)	29(38)	11(31)	51(35)	
Not provided	0(0)	2(3)	1(3)	3(2)	
Occupation sector					
Water authority	6(17)	12(16)	21(64)	39(27)	< .001***
Agriculture	10(28)	5(7)	3(9)	18(12)	
Construction	13(36)	26(34)	3(9)	42(29)	
Other	7(19)	31(41)	6(18)	44(30)	
Not provided	0(0)	2(3)	0(0)	2(1)	
Commitment					
Not committed	1(3)	0(0)	0(0)	1(1)	< .001***
Somewhat committed	5(14)	7(9)	0(6)	12(8)	
Fairly committed	21(58)	18(24)	8(24)	47(32)	
Committed	7(19)	43(57)	15(46)	65(45)	
Strongly committed	0(0)	7(9)	10(30)	17(12)	
Missing	2(6)	1(1)	0(0)	3(2)	
Failure experience					
Cracks	9(28)	8(18)	10(21)	27(22)	—
Settlement	7(22)	13(29)	5(10)	25(20)	
Water outflow	7(22)	6(13)	14(29)	27(22)	
Erosion	6(19)	8(18)	9(19)	23(18)	
Biological activity	3(9)	4(9)	4(8)	11(9)	
Other	0(0)	6(13)	6(13)	12(10)	
Playing analog games					
Rarely	9(25)	33(43)	16(49)	58(40)	.038*
Couple times a year	15(41)	27(36)	13(39)	55(38)	
Monthly	8(22)	13(17)	3(9)	24(17)	
Weekly	4(11)	3(4)	1(3)	8(6)	
Playing digital games					

*Continued on next page*

*Continued from previous page*

Characteristics	Participants, <i>n</i> (%)				<i>p</i> <sup>a</sup>
	A ( <i>n</i> = 36)	B ( <i>n</i> = 76)	C ( <i>n</i> = 33)	Total ( <i>N</i> = 145)	
Rarely	17(47)	39(51)	21(64)	77(53)	.11
Couple times a year	5(14)	10(13)	8(24)	23(16)	
Monthly	7(19)	4(5)	2(6)	13(9)	
Weekly	5(14)	18(24)	2(6)	25(17)	
Daily	2(6)	5(7)	0(0)	7(5)	
Played FPS					.97
No	28(78)	59(78)	25(76)	112(77)	
Yes	8(22)	17(22)	8(24)	33(23)	
Played <i>Levee Patroller</i>					< .001***
No	14(39)	66(87)	26(79)	106(73)	
Yes	21(58)	10(13)	7(21)	38(26)	
Missing	1(3)	0(0)	0(0)	1(1)	
Computer skills					.062
Not skilled	6(16)	7(9)	3(9)	16(11)	
Somewhat skilled	12(32)	19(25)	8(24)	39(27)	
Fairly skilled	12(32)	22(29)	12(36)	46(32)	
Skilled	6(16)	21(28)	5(15)	32(22)	
Very skilled	0(0)	7(9)	5(15)	12(8)	

\*\*\*  $p < .001$ ; \*\*  $p < .01$ ; \*  $p < .05$  (two-sided).

<sup>a</sup> Significances are based on chi-square tests of association for the nominal and dichotomous variables; for ordinal variables a Kruskal-Wallis test was applied. Missing data were excluded. Post hoc tests with chi square test results were executed by excluding one of the organizations and running another chi-square test. With the ordinal variables Mann-Whitney tests were used with a Bonferroni correction.

## *Attitudes, Expectations and Reactions*

Preferences and expectations may also impact a training. Here I will highlight those of the patrollers. I will also discuss whether those expectations were met and how participants reacted about the design of the game. To assess the attitudes, expectations, and reactions the same typical 7-level (or 7-points) items were used:

1. Strongly disagree
2. Disagree
3. Somewhat disagree
4. Neither disagree or agree (neutral)
5. Somewhat agree
6. Agree
7. Strongly agree

Table 7.2 provides an overview of the statement items used and their results. Like with the baseline characteristics I will not go into all of the details. I will highlight the most relevant findings from this.

### Favorable toward use of games

From the baseline characteristics we know that patrollers do not play many games. One reason could be that they do not enjoy it as much. Games are supposed to be fun, but the patrollers did not overwhelmingly agree on this (S1;  $Mdn = 4$ ,  $IQR = 2-5$ ). Despite this, they did have a favorable attitude toward games. They may not enjoy games as much, yet they indicated that they think one learns from playing (S2;  $Mdn = 5$ ,  $IQR = 4-5$ ) and even stronger agreed on that games are valuable for training and education (S3;  $Mdn = 6$ ,  $IQR = 6-6$ ). Taken together, these statements indicate that participants had a positive attitude toward games.

This positive attitude is much different than what we would expect based on some of the (mis)perceptions of games in society, such as that playing games is a waste of time, trivial, and childish. Up front the players knew about *Levee Patroller*. Some even played it already. The prospect of playing this game may have influenced their choices. Then it still shows at the very least that they are in favor of using games like *Levee Patroller*.

Knowing that these people voluntarily decided to participate, this may have been expected. However, although the participants at Organization C did not participate voluntarily, they did not respond differently. They only found playing games less enjoyable than the participants at Organization B,  $U = 879$ ,  $p = .012$ ,  $r = .24$ . We already know they also play less analog games than the participants at Organization A, so the participants at Organization C seem to have the least affiliation with games.

### Big expectations and motivation

Participants also seemed to have sizable expectations and were very motivated to learn. Although in general they did not find playing games fun, they expected that playing *Levee Patroller* would be fun (S4;  $Mdn = 6$ ,  $IQR = 5-6$ ).<sup>7</sup> They had similar expectations about how much they would learn from it (S6;  $Mdn = 6$ ,  $IQR = 5-6$ ). Maybe because of their awareness of not having played too many games and having little computer skills, they were reserved about how well they would perform (S5;  $Mdn = 4$ ,  $IQR = 4-5$ ).

Clearest of all is that participants were eager to learn more about levee inspection (S7;  $Mdn = 6$ ,  $IQR = 6-7$ ). Motivation to learn is important for education and training and it seems the needed motivation was there. But did the game fulfill its promises?

---

<sup>7</sup> If players indicated that they already played the game, the statement was (automatically) formulated differently. It would say "I enjoy playing the game" instead of "I expect I will enjoy the game."

**Table 7.2** The results in percentages on the statement items before ( $N = 145$ ) and after ( $N = 136$ ) the training. Each item has the typical 7-level item format, ranging from strongly disagree (1) to strongly agree (7)

Statement item (S)		Disagree			Neutral	Agree		
		1	2	3		4	5	6
<i>Pre-questionnaire, N = 145</i>								
Game attitudes								
1.	I enjoy playing digital games.	11	17	8	22	20	16	6
2.	I learn from digital games.	3	8	9	28	38	12	1
3.	It is a good development that games are being used for training and education.	1	2	1	6	10	63	17
Expectations								
4.	I expect I will enjoy the game.	1	1	2	13	21	58	5
5.	I expect to achieve high scores with the game.	1	5	4	55	25	10	1
6.	I expect to learn much from the game.	1	0	0	12	23	55	10
Motivation								
7.	I would like to learn more about levee inspection.	0	0	0	2	7	59	32
<i>Post-questionnaire, N = 132</i>								
Evaluation								
8.	I enjoyed playing the game.	3	2	4	6	21	46	19
9.	In general I experienced the game as realistic.	2	3	3	12	24	52	5
10.	I learned much from the game.	2	2	2	8	28	44	15
11.	What I learned in this game, I could use in practice.	2	2	2	11	27	46	9
12.	The use of this game for levee inspection is valuable.	2	3	3	4	15	49	24
Design								
13.	The controls are easy to learn.	2	8	8	8	22	28	24
14.	I missed sound in the game.	7	24	4	17	18	18	11
15.	I received sufficient feedback in the game.	5	7	18	12	21	32	5

### The game fulfilled its expectations

On the surface, all participants seem to be satisfied with the game and training afterward. They agreed about equally that the game was fun, realistic, educational, relevant for practice, as well as useful as training tool (S8–S12;  $Mdn = 6$ ,  $IQR = 5–6$ ). No differences were observed regarding the expectations at the start.<sup>8</sup> The game fulfilled its promises.

<sup>8</sup> To compare if the expectations were met I compared S4 with S8, S6 with S10, and S3 with S12. I used a Wilcoxon Signed Ranks test to make the comparisons.

However, whereas beforehand hardly any differences were found between the three participating organizations, afterward these are found on each statement item.<sup>9</sup> In particular, participants at Organization B were far more positive than those at Organization C and somewhat more than those at Organization A.<sup>10</sup> Nevertheless, if we look at either Organization A or C separately their expectations are still fulfilled. It is rather that the expectations for Organization B were exceeded, at least when it comes to being fun,  $z = 2.57$ ,  $p = .010$ ,  $r = .31$ .

### **The game was more than sufficient**

Consistent with this fulfillment, the players gave the game an average rating of (on a scale of one to ten) seven ( $M = 7.00$ ,  $SD = 1.32$ ). A clear majority (67%) chose this rating. Only 11 people (8%) chose a number below six, an “insufficient” rating.

The rebel and co-rebel (who rated it a one and two, respectively) are outliers that influence the results. If we decide to exclude them, no differences are noticeable between the organizations on rating the game. If we include them, the participants at Organization B rate higher than Organization C, but not compared to Organization A. What we can take from this is that no huge differences exist. We can only observe again a tendency that the game and training are received more positively at Organization B.

This rating tells us that the participants were satisfied about the game. In general it neither exceeds expectations nor fails to meet expectations. From the design-related statements about which we had some doubts, it became clear that in retrospect participants found it easy to get used to the controls (S13;  $Mdn = 6$ ,  $IQR = 4-6$ ). About the lack of sound (S14;  $Mdn = 4$ ,  $IQR = 2-6$ ) and receiving feedback (S15;  $Mdn = 5$ ,  $IQR = 3-5$ ) participants felt more mixed. Some missed sound and/or sufficient feedback.

### ***The Inspection Perceptions***

Another interest was to see to what extent the game influenced participants’ perceptions of the inspection itself. Hereto I constructed three 7-points semantic differential items—for knowledge (I1), stress (I2), and complexity (I3)—and two 7-points Likert items—for routine (I4) and impact (I5). The quick and dirty answer is “Yes!” At all except for the stress inspection item a difference is to be seen (Table 7.3).

<sup>9</sup> Kruskal-Wallis tests reveal that differences exist between the three organizations on fun,  $\chi^2(2, N = 131) = 18.5$ ,  $p < .001$ ; on realism,  $\chi^2(2, N = 132) = 6.48$ ,  $p = .039$ ; on learning,  $\chi^2(2, N = 131) = 8.89$ ,  $p = .012$ ; on relevance,  $\chi^2(2, N = 131) = 8.67$ ,  $p = .013$ ; on usefulness,  $\chi^2(2, N = 131) = 11.6$ ,  $p = .003$ .

<sup>10</sup> Mann-Whitney post hoc tests show that concerning fun and usefulness Organization B scored higher than the other two. Regarding learning and relevance it also scored higher than Organization C.

### The knowledge perception gap closed

One of the more obvious purposes of this training was to increase the knowledge of inspecting levees. Although most participants were still somewhat reserved about their levee inspection knowledge, the participants indicated that they know more afterward ( $Mdn_{post} = 4$ ,  $IQR_{post} = 4-5$ ) than at the start ( $Mdn_{pre} = 4$ ,  $IQR_{pre} = 3-5$ ). It needs to be noted that more than one-third (42%) expressed to have the same knowledge. In addition, about 17% thought they were worse off at the end. They may have realized—because of the training—that they know less or should learn more.

What is rather interesting is that the perceived knowledge differed among the organizations,  $\chi^2(2, N = 145) = 8.78$ ,  $p = .012$ . It appeared that the participants at Organization A considered themselves less knowledgeable, especially compared to Organization C,  $U = 347$ ,  $p = .007$ ,  $r = .33$ . This difference is not noticeable after the training.

What most likely explains this is that Organization C has many employees and for some of them levee inspection is even their daily work. Only three participants (#118, #138, and #139) indicated that they had “very much” knowledge and they were all from Organization C. The employees at Organization A, on the other hand, were “starters.” Except for one, all of them had little experience with levee inspection.

The expert employees (probably rightfully) perceived themselves to have more knowledge than the volunteers and employees, before  $\chi^2(2, N = 145) = 24.6$ ,  $p < .001$ , and after,  $\chi^2(2, N = 134) = 25.6$ ,  $p < .001$ . These expert employees did not change their knowledge perception. The others did, explaining why the knowledge gap closed and no difference is noticeable anymore between organizations.

Besides the fact that Organization B had some expert employees as well, the participants at Organization B received various lectures and training in the past years, making them feel more knowledgeable compared to the participants at Organization A who did not do much for the past years. This explains why Organization B does not suffer from a knowledge gap. It further shows that regular training may make an impact on knowledge perception.

### Stress is hard to imagine

As mentioned, the perception of stress when inspecting levees was not impacted ( $Mdn_{pre} = 4$ ,  $IQR_{pre} = 4-5$ ;  $Mdn_{post} = 4$ ,  $IQR_{post} = 3-5$ ). This was a tough question to answer, because it requires to imagine inspecting levees in real life, which most have never done so. Still, one could imagine that playing situations in virtual environments may reassure people and give them more confidence in how to handle unfamiliar things. It is also possible to think the other way around: knowing what it means to inspect levees, this may increase stress. The results show, however, that the participants did not perceive this.

### More difficult than first assumed

On many occasions I talked to patrollers who more or less told me inspecting levees is not hard: “You just see it.” You see it and report it. That is everything but hard. Not that everyone found it easy. Initially, people were somewhat mixed about it: it was neither difficult nor easy ( $Mdn_{pre} = 4$ ,  $IQR_{pre} = 3-5$ ). The participants at Organization C who have the most experienced patrollers seem to consider it less difficult compared to those at Organizations A and B,  $\chi^2(2, N = 144) = 6.72$ ,  $p = .035$ .<sup>11</sup> From this one would think that with more experience, one would find a task easier. That sounds pretty logical and is further confirmed with the mere fact that employees find it easier than volunteers,  $U = 1583$ ,  $p = .029$ ,  $r = .18$ .

This was all before the training. After the training, the participants found it more difficult rather than easier ( $Mdn_{post} = 5$ ,  $IQR_{post} = 4-5$ ). Although finding it more difficult is a clear, general trend, differences between the organizations seem to have become stronger,  $\chi^2(2, N = 135) = 9.53$ ,  $p = .009$ . Further investigation shows that playing the game has had the strongest effect on Organization B and the least on Organization C,  $U = 650$ ,  $p = .002$ ,  $r = .31$ .

What happened is that playing the game made participants realize that levee inspection involves much more than they thought. The game urged them to think about failures and answers are not always clear-cut. The reason why its effect seems to have been most strong on the Organization B and least on Organization C is that Organization B had the most inexperienced patrollers and Organization C the most experienced ones. The experienced ones know what levee inspection involves and because of this those with the most expertise kept finding it relatively easy.

What confirms this idea is that no difference is evident between volunteers and employees after the training. The expert employees did not change their minds; the inexperienced employees did. The latter group also had the aha-effect of realizing levee inspection is not as easy. It is more difficult than they first assumed.

### Virtual experience gives routine

Exactly half of the participants agreed that they considered inspecting levees more of a routine after the game ( $Mdn_{post} = 4$ ,  $IQR_{post} = 3-5$ ). Initially they tended to disagree with this ( $Mdn_{pre} = 3$ ,  $IQR_{pre} = 2-5$ ). Very striking is that the mode of responses changed from “disagree” to “somewhat agree.”

The disagreement at first was expected. With no or little experience to draw upon, it is hard for something to become a routine. The tendency to agree with the statement shows that the practicing inside a virtual environment gave participants the feeling of getting a routine. Consistent with this, many participants told me at the end-meetings that at a point, the reporting of failures became a routine. No game

<sup>11</sup> I wrote “seem to consider” because post-hoc analyses with Mann-Whitney tests with a Bonferroni correction did not show—strictly speaking—significant results between Organization C and A,  $U = 412$ ,  $p = .022$ ,  $r = .28$ , and Organization C and B,  $U = 899$ ,  $p = .019$ ,  $r = .23$ .



sensemaking needed to be performed anymore. It became a routine and this routine seems to have become a substitute for inspecting in the real world.

This effect takes place equally among organizations and types of participants. It is a general effect and makes clear that virtual experience and practice gives routine independent of the background of the person playing.

### **Knowing impacts perception of consequences**

Similar to stress, I incorrectly thought that the training would not influence people's perceptions of the impact of levee failures. In fact, the training had a clear effect. Although participants somewhat agreed already that they know about the consequences ( $Mdn_{pre} = 5$ ,  $IQR_{pre} = 5-6$ ), they agreed more with it afterward ( $Mdn_{post} = 6$ ,  $IQR_{post} = 5-6$ ).

I believe that the crux of this change is reflected in the word "know." Participants may have been aware of the importance and possible consequences, yet they do not *know* the exact consequences. These are clearly visualized in the game. The game shows how an seemingly innocent crack turns itself into a huge settlement that could sooner or later breach and flood a region.

It turns out that beforehand the expert employees agreed more than the volunteers,  $U = 271$ ,  $p < .001$ ,  $r = .34$ , or regular employees,  $U = 66$ ,  $p = .002$ ,  $r = .51$ . After the training we cannot speak of a difference. This confirms that knowing is a cause of this change. It impacted players' perception of the consequences of levee failures.

### ***The Knowledge Perception***

In yet another part of the questionnaire, we delved into the specific knowledge the participants had. The items were based on the five learning objectives: observing, reporting, assessing, diagnosing, and taking action. Each of these five relate to important aspects in dealing with risks. For observing it is necessary to know what kind of failures occur (K1); where they occur (K2); and, of course, to recognize them (K3). Reporting (K4), assessing (K5), and taking action (K8) have each one aspect that relates to them. Diagnosing has two: predict how a failure might develop (K6) and determine the failure mechanism (K7).

The question was what knowledge perception participants had before and after the training. To measure this, a 5-level Likert items were used. On these items the participants had to indicate how "well" their knowledge was with regards to one of the eight knowledge perception items. Concerning knowledge, I preferred asking this directly over an indirect disagree/agree item construction. I included one direction, because it seemed questionable to include "bad" or any other negative alternative to the item. Each item had this level-structure:

1. Not well

2. Somewhat well
3. Fairly well
4. Well
5. Very well

The overall results show that the training had a strong influence on participants' perception of their knowledge (Table 7.4). This indicates that the game had an effect on all relevant learning objectives and that perceptually it made a positive impact. Let us now consider the details pertaining to these items.

### Observing improves

The three statements related to observing (K1–K3) show a similar pattern. Whether it is about knowing what kind of failures occur, where they occur, or how to recognize them, a shift occurred from the start ( $Mdn_{pre} = 3$ ,  $IQR_{pre} = 2-3$ ) to the end of the training ( $Mdn_{post} = 3$ ,  $IQR_{post} = 3-4$ ). This gives us a strong reason to believe that in terms of improving (the perception of) observing the game was successful.

The patterns are similar, yet not if we consider the three organizations. With the location of failures (K2) no differences exist among them, but if we look at the kind of failures (K1), Organization C outshines Organization A,  $U = 421$ ,  $p = .016$ ,  $r = .29$ , and with recognizing failures (K3) Organization A again,  $U = 402$ ,  $p = .019$ ,  $r = .28$ , and this time also Organization B,  $U = 874$ ,  $p = .006$ ,  $r = .26$ . After the training these difference diminished.

Most participants at Organization B recently had a lecture in which the different failure types were explained, so that is why with observing failure types they do not differ from Organization C, whereas Organization A does. Listening to a lecture does not impact recognition skills and this may explain why Organization B joins A in perceiving to do less well compared to Organization C.

### Reporting follows observing

Knowing what to pay attention to when reporting is somewhat identical to knowing the failure types. Beforehand most participants thought they knew this somewhat well or fairly well and after the training this became fairly well to well ( $Mdn_{pre} = 3$ ,  $IQR_{pre} = 2-3$ ;  $Mdn_{post} = 3$ ,  $IQR_{post} = 3-4$ ). In addition, here too Organization C indicates they know relatively more than Organization A at the start,  $U = 361$ ,  $p = .002$ ,  $r = .37$ . The same reasoning applies. Organizations B and C have been regularly instructed on how to deal with failures; those at Organization A have not.

### Assessing and diagnosing somewhat different

The three items associated with assessing and diagnosing have somewhat different outcomes than those associated with observing and reporting. If we take the assess-

**Table 7.3** The results in percentages on the inspection items before ( $N = 145$ ) and after ( $N = 136$ ) the training. Each inspection item has its own scale from one till seven

Inspection item		Disagree			Neutral	Agree			Wilcoxon			
		1	2	3		4	5	6	7	<i>z</i>	<i>p</i>	<i>r</i>
1.	I have much knowledge about inspecting levees.	Pre	8	10	14	38	21	8	2	4.23	<.001	.37
		Post	3	5	8	37	38	8	2			
2.	I experience inspecting levees as relaxing.	Pre	0	1	16	38	27	17	0	1.08	.28	.09
		Post	0	2	26	33	23	16	1			
3.	I find reporting failures difficult.	Pre	0	15	20	39	23	4	0	4.37	<.001	.38
		Post	0	8	15	27	41	9	0			
4.	I experience inspecting levees as a routine.	Pre	3	27	23	18	17	10	1	3.30	.001	.29
		Post	3	15	17	26	27	11	2			
5.	I know what the consequences could be of a failure	Pre	0	5	8	7	38	33	10	3.62	<.001	.32
		Post	0	1	2	8	30	50	9			

ment of the severity of failures we see a less strong improvement ( $Mdn_{pre} = 3$ ,  $IQR_{pre} = 2-3$ ;  $Mdn_{post} = 3$ ,  $IQR_{post} = 3-3$ ). That is not an unexpected result. Assessing is a difficult task, because it is not always clear-cut what the severity is. The game has given the players a feeling about the severity of failures and made them think about it, but also highlighted the difficulty determining the severity.

Showing how a failure develops over time is one of the stronger points of the game and luckily this seems to pay off ( $Mdn_{pre} = 2$ ,  $IQR_{pre} = 2-3$ ;  $Mdn_{post} = 3$ ,  $IQR_{post} = 3-3$ ). In general participants have been more reserved in answering this question and this has to do with being able to determine this. Similar to the location of a failure, uncertainty exists as to how a failure develops. Absolute certainty does not exist and with this in mind participants might be less likely inclined to state here that they know this “very well.”

The strongest effect is to be observed with determining the failure mechanism ( $Mdn_{pre} = 2$ ,  $IQR_{pre} = 1-3$ ;  $Mdn_{post} = 3$ ,  $IQR_{post} = 2-3$ ). Some had never heard of it, so the learning curve involved with it must have been very high. However, not everyone experienced the same learning curve. Especially the participants at Organization A had issues with the failure mechanisms and this became more evident after the training,  $\chi^2(2, N = 136) = 11.0, p = .004$ .<sup>12</sup>

<sup>12</sup> Post-hoc analyses show that Organization A differs in determining failure mechanisms from Organization B,  $U = 831, p = .002, r = .31$ , and Organization C,  $U = 331, p = .002, r = .32$ .

### Clear effect for taking measures

With respect to taking measures we see a similar strong effect as with diagnosing the failure mechanism ( $Mdn_{pre} = 2$ ,  $IQR_{pre} = 2-3$ ;  $Mdn_{post} = 3$ ,  $IQR_{post} = 3-4$ ). In contrast to failure mechanisms, participants may have had some ideas of what measures to take. The game seemed to have helped inculcate a better understanding.

**Table 7.4** The results in percentages on the knowledge perception items before ( $N = 145$ ) and after ( $N = 136$ ) the training

Knowledge item			Not well to Very well					Wilcoxon		
			1	2	3	4	5	$z$	$p$	$r$
1. I know what kind of failures could appear.	Pre	2	24	58	15	2		5.43	<.001	.47
	Post	0	7	56	35	2				
2. I know where a failure could occur.	Pre	2	24	55	18	1		5.65	<.001	.49
	Post	0	8	53	37	2				
3. I can recognize failures.	Pre	3	30	51	14	1		6.29	<.001	.55
	Post	0	8	55	36	2				
4. I know what to pay attention to when reporting a failure.	Pre	3	30	50	15	2		5.24	<.001	.46
	Post	0	15	54	28	3				
5. I am able to assess the severity of a failure.	Pre	1	37	48	14	1		3.31	<.001	.30
	Post	0	21	60	18	1				
6. I am able to determine how a failure will develop.	Pre	4	44	38	13	1		5.08	<.001	.45
	Post	0	20	60	18	2				
7. I am able to determine the failure mechanism of a failure.	Pre	27	33	27	11	2		6.76	<.001	.59
	Post	2	25	53	16	5				
8. I know what measures prevent a failure from becoming worse.	Pre	8	46	34	11	2		6.73	<.001	.59
	Post	0	20	49	26	5				

### Reducing the Data

Reducing the acquired data will help facilitate further analysis. Hereby multiple items are put together that measure a *latent variable*. A latent variable is not directly observed; it is inferred from others. Reducing data and thereby finding latent variables is done to achieve parsimony, that is, the use of fewer variables to explain phenomena. Parsimonious models are easier to comprehend and are therefore more persuasive.

However, simplicity of models does not entail accuracy. Another, more important reason to reduce data—especially when using Likert items—is to improve validity. One single item is unlikely to accurately represent a variable, especially if it concerns “difficult” constructs such as knowledge or attitudes.

Each item in the questionnaire represents a variable I considered relevant for understanding the input, use, and outcomes of a game-based training. Each may not be very accurate but it still gives us some idea of how participants perceived.

These perceptions I have just elaborated upon. We have seen that players found inspection more complex, thought to have gained routine in inspecting, and had a better idea of taking measures among many other very specific variables they were asked to give their opinion on—before and/or after playing the game. It might be that some of these separate item variables do measure together a deeper latent variable, one that resides on a higher aggregation level. My purpose was to explore this possibility with the items described in this level.

I made use of Principal Components Analysis (PCA) because I did not aim to confirm a model (making Confirmatory Factor Analysis inappropriate) nor did I aim for developing one (making Exploratory Factor Analysis inappropriate too).<sup>13</sup> I only needed to reduce my data. PCA identifies clusters of inter-correlated variables by considering *all* the variance in each variable. These clusters are called “components.” I ran the analysis separately for the Likert items in the pre- and post-questionnaire (with varimax and controlling for oblique rotations).

### ***The Pre-Questionnaire Components***

I inserted all variables (S1 to S7; I1 to I5; and K1 to K8) into a PCA and this immediately gave good, interpretable results (Table 7.5). Four components (C1 to C4) were extracted, explaining in total 62% of the variance. No complex variable seen; that is, no variable loads highly on more than one component. In addition, other important indicators for a successful PCA are fulfilled too (Field, 2005).

The first component, which always explains most of the variance, includes all knowledge perception items and two of the inspection items. Up front I expected to find more than one component among the knowledge perception items, but after inspecting their correlations I started to think differently. Despite some differences in responses, especially regarding assessment (K5) and diagnosing failures (K7), the knowledge perception items correlate strongly ( $> .40$ ) with one another. Not too strongly ( $> .90$ ), because otherwise this would be a sign of multi-collinearity and a reason of concern for interpreting the PCA results. With such strong correlations it is not surprising to find all of them loading highly onto one component.

The component interpretation is given by the I1 inspection item. This asks about how much knowledge the person thinks he or she has regarding inspecting levees. This could be seen as the general knowledge question; the knowledge perception

---

<sup>13</sup> Some researchers equate Principal Components Analysis (PCA) and Exploratory Factor Analysis (EFA). This is incorrect (Field, 2005).

items are items related to a particular subject of inspection a person should have knowledge of. But all these statements, as implied by their name, are about knowledge. The second inspection item (I5) is too, because it asks to what extent a person has *knowledge* of “what the consequences could be of a failure.” Therefore, the first component (C1) represents *knowledge perception*. Apparently, we do not need to make any further distinction into any types of knowledge.

**Table 7.5** Principal components analysis (with varimax rotation) of the pre-questionnaire Likert items ( $N = 133$ ). The first analysis is shown and the second after excluding two variables. Variables are ordered according to their loading on the first analysis

Item	1st analysis				2nd analysis			
	C1	C2	C3	C4	C1	C2	C3	C4
K1 “what kind of failures”	.871				.875			
K8 “what measures prevent a failure”	.846				.852			
K2 “where a failure could occur”	.841				.848			
K3 “recognize failures”	.841				.832			
K4 “what to pay attention to”	.812				.808			
K6 “how a failure will develop”	.812				.816			
I1 “know much about inspecting”	.768				.763			
K7 “determine the failure mechanism”	.766				.774			
K5 “assess the severity”	.744				.747			
I5 “what the consequences could be”	.588							
S1 “enjoy playing digital games”		.779				.791		
S2 “learn from digital games”		.765				.780		
S3 “games...used for training”		.602				.621		
S5 “expect...high scores”		.439						
S7 “would like to learn more”			.771				.778	
S6 “learn much from the game”			.725				.733	
S4 “I will enjoy the game”			.634				.632	
I4 “experience...as a routine”				.745				.759
I2 “experience...as relaxing”				.719				.737
I3 “find reporting...difficult”				.540				.549
Explained variance, %	34	14	8	6	36	14	9	6
Cronbach’s alpha	.922	.644	.645	.565	.928	.662	.645	.565
Kaiser-Meyer-Olkin			.885				.883	
Bartlett’s Test			<.001				<.001	

*Note.* To increase legibility only a fragment of the statements are included.

The second extracted component (C2) includes the three statements that I grouped as *game attitudes*. A fourth variable relates to this too: the expectancy of achieving high scores. It seems that participants who think they will achieve high scores are more likely to score high on the three game attitude statements. This is plausible. Someone with a positive game attitude are people who either play these games or are fond of these new types of information technology. They are at least not “afraid” of it. Then it follows that such people are more confident about being able to do well.

Of course, this is also a personality issue. Some people are more reserved about their performance than others. That is probably why the score expectancy variable loads much lower on the component compared to the other ones.

The two remaining expectancy variables—about expecting to have fun with and learning from the game—combine together with the motivation variable the third component (C3). Conceptually motivation is different from having expectancies. Motivation is about a willingness and need to do something; expectancies are beliefs about what something will achieve. Despite the conceptual difference, both relate to each other into something I refer to as *success potential*. If a person has low expectancies, that person may already think that this training is not really for him or her. The potential for success is low. Similarly, as being motivated to learn is crucial for a successful training, the success potential is low too if one is not so motivated.

I never expected the remaining three variables to be so closely related, because they measure quite distinct concepts. These variables concern finding inspecting levees a routine (I4), relaxing (I2), and difficult (I3). The latter variable was reversed, so it actually indicates that participants who found inspecting a routine are more likely to also find it relaxing and easy. To me this says something about having *confidence* (C4). Participants who scored high on these variables are probably more confident about their abilities than those who scored low. The poor reliability score ( $\alpha = .565$ ) is a reason of concern and indicates that these variables may indeed measure distinct concepts.

Although the initial analysis provided solid results, I decided that it was necessary to exclude two variables: knowing the consequences (I5) and expecting high scores (S5). For both less than half of their variance was explained for by the components; deleting them improved the scale's reliability; and a gap is seen between their loading on the component and the other variables. The latter is also true for finding reporting difficult (I3), but excluding this variable leads to an even worse reliability score for the confidence component (C4) and it has good scores elsewhere. The exclusion of these variables does not lead to any rigorous changes among the components (Table 7.6).

This means that we can reduce the pre-questionnaire to four components: knowledge perception, game attitude, success potential, and confidence.

### ***The Post-Questionnaire Components***

Similar to the pre-questionnaire, the initial PCA of the post-questionnaire reveals four components and explains about 62% (Table 7.6). Some similarity to the pre-questionnaire should have been expected, because 13 out of 22 variables (S8 to S15; I1 to I5; K1 to K8; and R) are identical. If the results are somewhat identical, this gives a stronger case to assume that these variables belong to one component. It validates our findings.

The first component is identical to the first on the pre-questionnaire—the one about knowledge perception—and contains pretty much the same variables. It only

includes one more variable: the variable about finding reporting difficult (I3). Now at the pre-questionnaire we already observed that this is somewhat of a troubled variable, because a gap is seen between its loading on the confidence component and the others (C4; see Table 7.6). Here it does not load on the confidence component anymore, but on the knowledge one. To some extent this is explainable. Before participants decided this on the amount of confidence they had in reporting; after it is based on playing the game. This is arguably a better assessment, because before it was largely based on assumptions and after it was based on an actual experience—albeit a virtual one.

Despite this possible explanation, the variable's loading is markedly lower compared to the others and less than half of its variance is explained for by the components. Because this variable also did not appear during the pre-questionnaire analysis, this was reason enough for me to exclude it. Along with this line of consistency I excluded the variable about knowing the consequences (I5) as well. Its variance was again not much explained for ( $< .40$ ) and it also has a comparatively lower loading on the knowledge perception component.

Unlike the pre-questionnaire, the post-questionnaire reveals a strong second component. This component involves almost all evaluation and design statements. It further includes the rating participants gave. On an aggregate level all these measurements are about judging the game (and/or training) and so it seems reasonable to coin this second component *judgment*.

The only statement not included with the judgment component concerns the statement about missing sound (S14). This statement makes up the fourth component together with receiving feedback (S15), which loads highly onto the judgment component too. This makes the feedback variable complex, which is a valid reason for exclusion: less than half of its variance was explained. The cause of this complexity can be traced back to the spread of the answers (Table 7.2). Here we see that participants are split. Some agreed with the statement and others disagreed. It could very well be that a majority of those who judged the game positively, were generally positive about everything and therefore concluded that the game gave them sufficient feedback. That at least explains the loading of the feedback variable onto the judgment component. Another part may have been positive, but was critical about some of the design and then in particular about receiving feedback.

Sound is most certainly a very specific design issue, on which participants were ambivalent. Although it loads onto the fourth component with the feedback variable, no correlation exists between the two variables and the reliability of this possible scale is extremely poor ( $\alpha = .205$ ). After deleting the feedback variable, the sound variable still remained on its own, making up a component that actually is not one. I decided to neglect this fourth component for any further analysis, especially because I did not have any other plans with the feedback as well as sound variable. I simply wanted to know how participants felt about these design issues. The results of the individual items were more than sufficient for my purposes.

Being left alone by the variable about finding reporting difficult (I3), the remaining original confidence variables continue to be a separate component, confirming that this is indeed another component. However, a scale with only two items is in-



**Table 7.6** Principal components analysis (with varimax rotation) of the post-questionnaire Likert items ( $N = 114$ ). The first analysis is shown and the second after excluding two variables. Variables are ordered according to their loading on the first analysis

Item	1st analysis				2nd analysis			
	C1	C2	C3	C4	C1	C2	C3	C4
K3 “recognize failures”	.826				.839			
K1 “what kind of failures”	.798				.796			
K8 “what measures prevent a failure”	.793				.788			
K2 “where a failure could occur”	.788				.848			
K6 “how a failure will develop”	.785				.801			
K4 “what to pay attention to”	.768				.768			
K7 “determine the failure mechanism”	.710				.724			
I1 “know much about inspecting”	.684				.686			
K5 “assess the severity”	.660				.666			
I3 “find reporting...difficult”	.571							
I5 “what the consequences could be”	.492							
S10 “learned much from the game”		.905				.902		
S12 “The use...is valuable”		.880				.887		
S8 “enjoyed playing”		.857				.859		
R Rating (1 to 10)		.814				.803		
S11 “could use in practice”		.771				.779		
S9 “experienced...as realistic”		.760				.769		
S13 “controls are easy”		.717				.717		
I4 “experience...as a routine”			.806				.807	
I2 “experience...as relaxing”			.703				.761	
S14 “missed sound”				.785				.939
S15 “received sufficient feedback”		.465		.783				
Explained variance, %	32	20	7	5	35	21	7	5
Cronbach's alpha	.898	.896	.525	.205	.909	.914	.525	—
Kaiser-Meyer-Olkin			.859				.861	
Bartlett's Test			<.001				<.001	

*Note.* To increase legibility only a fragment of the statements are included.

sufficient and it has a poor reliability score ( $\alpha = .525$ ). What it, therefore, especially confirms is that the confidence component is something we need to be wary of.

In sum, the pre-questionnaire shows two strong components, another knowledge perception component and a judgment component. The remaining two I refer to as miscellaneous, because they do not really make up anything that we should further consider.

## *Validating and Calculating the Components*

To some extent the acquired results from reducing the data are already validated. The pre- and post-questionnaire yield somewhat similar results. Despite this, it is appropriate to perform a split-half validation. Hereby the sample is split in half and the same analyses are calculated. The reasoning behind this is that the larger the sample size, the greater the opportunity to obtain significant findings. By performing the same analyses with both halves of the original sample, it is possible to see if the components remain the same. If they do, we can be more confident that these patterns really exist.

### **In-depth explanation: running the split-half validation**

I ran the split-half a number of times, every time with a random half of participants. The results show that on all occasions the extracted components explain for enough variance. They also show that with the pre-questionnaire, as expected, the knowledge perception remains always the same. The other components are less stable. Although on a few occasions both halves were exactly the same, mostly one of them was similar and the other had some differences. The variables of the game attitude and potential component interchanged or had complex loadings. The infamous variable about finding reporting difficult (I3) sometimes showed a pattern I recalled from the post-questionnaire: to load on the knowledge component and not on the confidence component.

With the post-questionnaire variance was explained for as well and here too instability was to be seen in half of the split-half. Remarkably this instability was attributed to the knowledge perception component and not to the judgment one. The latter was quite stable. What above all happened is that exactly the second half of the knowledge perception items (K5 to K8) formed another component. Up front I expected something like this to happen, because all the knowledge perception items refer to different types of knowing. The first half of knowledge perception items (K1 to K4) refer more to declarative types of knowledge—it is more factual. The second half relates more to conceptual knowledge. It is about understanding what they see. Important to mention is that the first four knowledge perception are what is required of levee patrollers; the second half is what is not required.

The split-half validation results reveal clearly visible pattern. Yet certain cracks in the data are apparent too. Knowing the large variety of participants and the unbalance in the number of participants per organization, some instability is not surprising. It is certainly not caused by any outliers, because on both questionnaires only two outliers were noticeable. On the pre-questionnaire one participant (#139) concerned an outlier on the knowledge perception component; the well-known rebel from Organization C (#121) was an outlier on the success potential component. At the post-questionnaire two participants (#8 and #150) scored very low on the judgment component. Nevertheless, these outliers did not have any effects on the results as presented here.

Aware of the instability I continued by creating summated scales as defined earlier by the PCA (but without considering the component loadings). Because the items differ in points (five, seven, or even ten), I had to be somewhat creative. Only in three situations this was really an issue: with the two knowledge perception com-

ponents (I1 is a 7-points Likert item) and with the judgment component (R is a 10-points item). To prevent these variables from contributing more to the summative scale than others, I weighted their contribution (with 5/7 and 5/10 for the 7-points and 10-points items, respectively).

I further decided to ignore the confidence components. The variable about the difficulty of reporting failures (I3) is rather unstable, as it loads variably on the knowledge perception component and the confidence components; deleting this variable leads to an undesired number of items per scale, and the scale has a somewhat debatable reliability (with and without variable I3). This was sufficient motive to only acknowledge some of the inspection items relate to each other, which may have to do with confidence.

Inspection of remaining summated scales shows that in terms of distribution the four earlier mentioned outliers are problematic. Upon removing them both knowledge perception scales fulfill the criteria of the normal distribution perfectly. The others—especially the judgment scale—had a negative skew, making it necessary to transform them.<sup>14</sup> After this, they approximated a normal distribution.

## Investigating the Relationships

With these new, parsimonious scales established it is time to look into how all of this relates to each other! One important reason for reducing the variables is to prevent a complex scheme of relationships. To look into the relationships of variables is to look into “correlations,” that is, if and how one variable changes with another. Such relationships are not directional, so we do not know what causes what to change, but we do know if a dependency exists.

I first set out to look into the newly established components and from there I added the background variables described at the start. The variables (and components) who have been neglected based on the PCA analysis have been considered here again (just in case), but as expected, they did not lead to anything that needs further attention. It would only unnecessarily complicate the picture I am about to describe.

One word of caution is needed. For consistency I applied everywhere Pearson correlations which strictly speaking are only suitable for variables on an interval level. I have done this because with the statistical software package I used it is only possible to calculate partial correlations—the calculation of relationship between two variables by controlling for a third variable—with Pearson correlation. In addition, some variables, as I just explained, are on an interval level and so their relationship would have been underestimated if other types of correlations (for example, Spearman’s correlation and Kendall’s tau-b) would have been considered.

---

<sup>14</sup> I transformed the attitude, success potential, and judgment scales with square root. Because all three had a negative skew it was first necessary to reflect the distribution (by taking the maximum value and subtracting the variable from this). After calculation I reflected the distribution again to make interpretation easier.

As a consequence, for those variables that are not on an interval level, the relationship are overestimated. I always checked the other correlation possibilities and the results in terms of “significant correlations,” the relationships that matter, the results are very similar.

### *Effect of Self-Fulfilling Prophecy*

We immediately perceive a strong relationship between knowledge perception before and after the training,  $r = .55$ ,  $p < .001$ . This means that to a large extent people with low perceptions remained having low perceptions and those with high ones kept on having high perceptions. However, this does not mean that people's perceptions did not change. On the individual items we have seen already medium to strong improvements in perception. With the summated score this pattern is the same: it shows a strong difference in perception,  $t(111) = -8.49$ ,  $p < .001$ ,  $r = .63$ .

Initially, the other three scales—game attitude, success potential, and judgment—relate to each other (all  $p < .001$ ). Based on the split-half analysis, where I observed that certain variables shifted from one scale to the other, I expected game attitude and success potential to be somehow related and indeed they are related,  $r = .30$ . Theoretically, this is not a revelation. If one has a positive attitude toward games, then the success potential of this training is most likely regarded much higher. In fact, the relationship could have been expected to be stronger. Reason why it is not is that some participants may not have had so much with games, but were open-minded about this training.

Because all three relate to each other I proceeded by calculating the partial correlations. It turns out that now only attitude relates to judgment,  $r = .36$ ,  $p < .001$ . Apparently, the relationship of success potential is mediated by others.<sup>15</sup> Despite that success potential does not relate to it anymore, the relationship with attitude does make clear that people's initial position on the training makes a difference. When I saw that rebel during the start-meeting I also did not expect he would judge the game any favorably at the complete end. Although people could change their minds (and some of them may very well did), people's initial take on an activity has an influence on how they perceive it. To an extent it remains a self-fulfilling prophecy.

Whereas the pre-knowledge perception only relates to the knowledge perception after, the latter does have relationships with the other components. It has medium relationships with success potential as well as judgment, both  $r = .33$ ,  $p < .001$ , and a small, almost negligible one with attitude,  $r = .19$ ,  $p = .043$ .<sup>16</sup> After controlling for possible mediating variables, it turns out that attitude does not relate to it anymore; success potential still does,  $pr = .31$ ,  $p = .001$ . This highlights again that an effect

<sup>15</sup> To look into the partial correlation of success potential with judgment I controlled for attitude as well as post-knowledge perception. Both relate to judgment and to success potential.

<sup>16</sup> If instead of Pearson correlation we would consider the Spearman correlation, the relationship between the post-knowledge perception and attitude is non-existent.

of self-fulfilling prophecy might have taken place. It shows that those participants with high expectations and motivation perceived to have learned more.

### *Adding the Background Variables*

From here I added the background variables. I did not consider all variables, because some of them are categorical, such as gender and occupation, and not suitable for correlation analysis. Therefore, of the 11 background variables, only five were considered: age, playing analog games, playing digital games, computer skills, and commitment. To start with the first, this not so surprisingly related to computer skills. The older participants are, the less their (perceived) computer skills.

Age plays a role—although not a big one—in how the participants eventually judged the game. This is not caused by computer skills because I made sure this variable was controlled for. Popular views may explain this as that older people tend to be more conservative or critical, but various research, especially in the realms of political choice, contest this belief (e.g., Pillemer, 2011; Pollak, 1943). Older people may be very open-minded and flexible.

I even encountered this during the training. At the start-meeting I met Jan (#34), a 63-year old and one of the few people who asked me if he could fill out the questionnaire with pen and paper. When he started playing he was thrilled about what he was doing. He did not have much computer skills, but he was eager to learn. At the end of the start-meeting he was able to get his way around in the virtual environment.

Later he emailed me to say that he could not play the game on his computer. I came to visit him and discovered he had an integrated graphics card on his computer, which tends to cause technical problems. I loaned him a laptop so he could play. When I was about to leave him, he told me he was trying to get a license to become a security guard. I am not so sure what kind of market exists for 63 year old security guards, but he told me “You are never too old to learn.”

On the other hand, I have also encountered participants (e.g., #20 and #21, both 49 years old) who thought this was something for their kids. So it kind of depends on people’s attitudes and age seems to play a (small) role in this. Therefore, conservatism could play a role but it certainly does not have one as we would expect.

I thought beforehand that if participants play games regularly, they are more likely to judge the game positively. I made a distinction between analog and digital games just in case, because liking analog games could very much lead to having such positive attitude too. It turns out that people who play analog games are more likely to play digital games, but that this does not relate to the game attitude component, which should rather be called “digital game attitude.” It measures what attitude participants have regarding digital games, not regarding gaming in general.

Playing digital game relates very strongly to this game attitude, indicating that people who play games more frequently will have more positive affiliations and expectations about games. Much less obvious is how having computer skills relates to being committed to the inspection. It turns out that people working for the water

authorities are more committed, and reasonably so, than volunteers,  $U = 1424$ ,  $p < .006$ ,  $r = .23$ . The employees tend also to be younger and this is related to having computer skills too.

### *Creating a Structural Model*

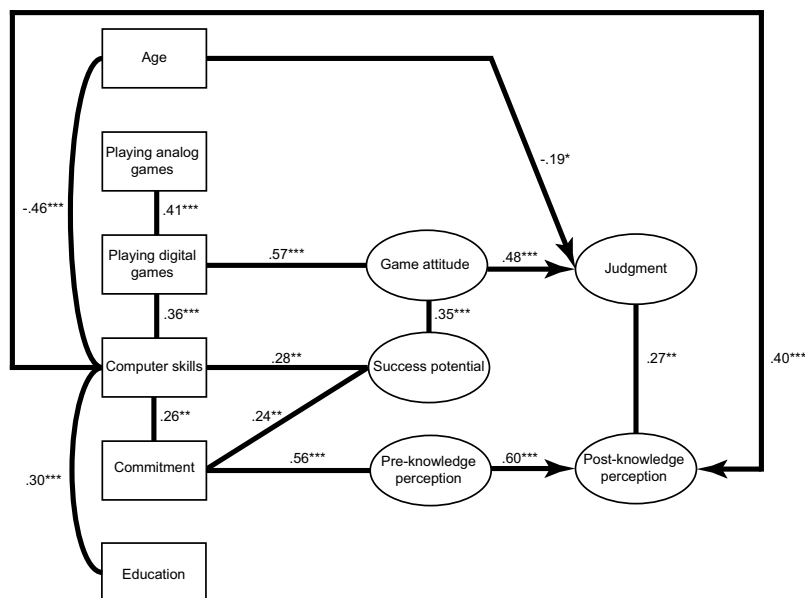
Although I already talked about how the background variables connect to some of the components, let us now put it all together. To make this possible what I was after is to construct a *clean structural model*. A structural model shows causal dependencies between variables. With “clean” I refer to a model that is simple and only shows the most relevant relationships, that is, relationships that are not mediated by others. For creating this model, of which the final result is visualized in Figure 7.1, I followed the following procedure:

1. I calculated the correlations between all components and non-categorical background variables. I first looked at a core set of variables and then extended the model by adding more variables to the model.
2. I calculated partial correlations if relationships existed between multiple variables, which indicates that possibly one of the variables mediated the relationship with another. If a variable turned out to mediate the relationship I deleted the mediated relationship.
3. If a variable had more than one indirect relationship with another variable, each indirect relationship was corrected for in calculating the partial correlations. If it turned out that the relationship remained significant after these corrections, the relationship was maintained. Otherwise it was deleted.

The model demonstrates that only playing digital games relates to having game attitudes. A relationship with computer skills was initially present too. This was, however, mediated by playing digital games. This variable is the only one that affects the game attitude component. Similar to playing analog games, it does not relate to any of the other variables (directly). As the game attitude affects people's judgment, playing digital games as well as age seem to have a role here.

The remaining two background variables, computer skills and commitment, do relate to more than one component. Having computer skills relates to the success potential. It also relates to post-knowledge perception, even after correcting for possible mediating variables. Those with higher computer skills perceived to have learned more. This confirms the observation in Level 5, that players first have to learn how to play the game, before they are able to learn from the game. If due to having few computer skills participants get stuck at learning how to play the game, they will hardly learn from the game.

This relationship necessitated an examination of the data, whereby I found one person who severely decreased his knowledge perception—about a 26% decrease. This was no one other than the criminal investigator Henk (#42). Before the training Henk's perceived knowledge was way beyond the average ( $M_{\text{pre}} = 24.4$  and Henk's



**Fig. 7.1** Correlations (Pearson) between background variables and the components. Adjustments are made for partial correlations. The ovals represent the components and the rectangles the background variables. \*  $p < .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$  (two-sided)

score = 35.6); after it was below it ( $M_{\text{post}} = 28.3$  and Henk's score = 23.9). The other participants with a decrease in knowledge perception seem to be either people with few computer skills or participants who hardly participated. These participants did not have such an extreme difference like Henk, so he was a real exception in this regard.

It goes the other way around too. If someone has computer skills, he is more likely to get more out of it. Such people can devote their cognitive resources completely to the content of the game. In this light, we find that some of the participants with the highest gains were (relatively) young participants who indicated to have high computer skills.

Like computer skills, commitment seems influential. In addition to computer skills it relates to the success potential and pre-knowledge perception. In other words, these relationships mean that committed participants are those who have a higher perceived knowledge up front. That is something we could have expected, because if one is committed, one will devote time and energy into knowing the material.

Committed people could further generally be considered to be more motivated, exactly why it is related to success potential. In looking at the correlations between commitment and the variables that make up the component, it only relates to the motivation variable (S7),  $r_s = .40$ ,  $p < .001$ . Commitment is more than just motivation.

Being committed may have led people to feel an important need to do well and learn from this training. I noticed this with Participant #70. He could not do anything wrong and that is why he called me many times when he ran into a problem. He was one of the few voluntary levee segment leaders and felt responsible. He had to know his business, because he was supposed to make sure his crew would, on their turn, know their business too. Such pride and need goes beyond the “simple” willingness to learn something: they *have to*.

Thus, the relationships become clear and are illustrated in Figure 7.1. It is not as complex as it would have been if we would have used all 39 variables discussed in this level, but it is still daunting. The relationships make specifically clear that post-knowledge perception is directly influenced by participants’ *a*) knowledge perception before the training, *b*) computer skills, and *c*) how they judged the training. And that judgment of the training is directly influenced by participants’ *a*) knowledge perception after the training, *b*) attitude toward games, and *c*) age.<sup>17</sup>

### *Considering the Categorical Variables*

Although I did not consider the categorical variables in creating the structural model, I was still curious to see what these variables tell us about the target group and the results. The first possible differentiation I looked at was by considering a number of categorical variables from Table 7.1: gender, education, occupation, failure experience, and experience in playing FPS games and *Levee Patroller*.

I considered education as a categorical variable, but if we would consider it as an ordinal one, which is not completely awkward because it is about the level of education, then we find that it correlates with computer skills,  $r_s = .36$ ,  $p < .001$ , and judgment,  $r_s = .25$ ,  $p = .004$ . However, the latter correlation is mediated by the first and so education only really relates to having computer skills.

This also applies when considering ANOVA analyses: from all one-way ANOVAs education only makes a difference on judgment, but if we consider a two-way ANOVA with education and computer skills as factors, education is not relevant anymore. Education seems, therefore, not much more than an explanatory variable for why people vary in computer skills. Less educated jobs involve less (complex) computer work, so this could be a reasonable explanation for why education relates to having computer skills. Because this relationship does exist, the education variables was added to Figure 7.1.

Then we go on to occupation. Here we have all reason to believe that people working for the actual organizations have more knowledge, especially if we contrast this to participants who seemingly have no affiliation to levee inspection with their jobs, such as bus drivers or school teachers. On pre-knowledge perception,  $F(3,130) = 7.52$ ,  $p < .001$ ,  $\omega = .36$ , and post-knowledge perception,  $F(3,116) = 4.01$ ,  $p = .009$ ,  $\omega = .26$ , differences are to be seen and the Games-Howell post hoc procedure

<sup>17</sup> These relationships can be validated, such as with multiple regression analyses. These analyses largely confirm the findings.



shows that the earlier expectation is indeed correct ( $p < .001$  for pre-knowledge perception;  $p = .017$  for post-knowledge perception). The differences are especially noticeable between the participants working at the organizations and participants who have a job categorized as “other.”

This other category had the highest means on the remaining components: game attitude, success potential, and judgment. But only in the latter situation significant differences are to be noted,  $F(3,130) = 3.59$ ,  $p = .016$ ,  $\omega = .24$ . Because here the variances were equal, I used the Hochberg post hoc procedure and found that it is in particular the participants who have a job in the agriculture sector that judged the game less positively, especially compared to people of the construction ( $p = .029$ ) and other ( $p = .013$ ) category.

Age could not explain this, because most older participants were as a matter of fact part of the “other” category and the participants of the agriculture category were pretty much spread out over all age categories more likely that it is due to computer skills—or better, a lack thereof. Upon recoding the computer skills into a dichotomous variable (Table 7.1), whereby the original first two levels are considered low skills and the remaining three levels are considered high skills, and making a crosstabulation with occupation, it turns out that 72% of the participants in the agriculture sector have low computer skills. The reverse is more or less true with the other categories, explaining why we find a significant difference among the occupation categories,  $\chi^2(3, N = 143) = 11.8$ ,  $p = .008$ , Cramer’s  $V = .29$ .

Education is not involved in this. Education and occupation reveal no specific peculiarities,  $\chi^2(6, N = 139) = 11.8$ ,  $p = ns$ , Cramer’s  $V = .14$ . Therefore, it seems that besides education, a specific type of occupation, that of working in agriculture (and more specifically, agriculture, horticulture, and/or stocking cattle) relates to having computer skills.

The final categorical variables are failure, FPS, and *Levee Patroller* experience. I recoded the first into a dichotomous variable, with experience or no failure experience, similar to the latter two. Over half of the participants with *Levee Patroller* experience were from Organization A, because during the time they did organize events this was one of them. From the other two organizations only the more expert employees played it, to see if it is suitable as a training tool. The latter explains why having this previous experience just and only affects pre-knowledge perception,  $t(128) = -2.52$ ,  $p < .001$ ,  $r = .30$ .

A known relationship based on Figure 7.1 is that playing digital games positively relates to game attitude. In considering what effect playing FPS games has, we should not be surprised then to see having played such games—even once<sup>18</sup>—brings forth a more positive attitude,  $t(142) = -5.53$ ,  $p < .001$ ,  $r = .42$ . Because game attitude relates to judgment on its turn, we see that furthermore that people who played FPS games judged the game more positively,  $t(128) = -3.59$ ,  $p < .001$ ,  $r = .30$ .

As for failure experience, when we consider the consequences of having failure experience on the five components we come to find that people with failure experi-

<sup>18</sup> I recoded the original FPS variable. Initially I asked participants to what extent—that is, never, once, sometimes, often—they played FPS games.

ence perceive to have more pre-knowledge,  $t(134) = -6.09$ ,  $p < .001$ ,  $r = .47$ , and more post-knowledge,  $t(118) = -2.24$ ,  $p = .028$ ,  $r = .20$ , but that those with *no* failure experience have a larger knowledge perception gain,  $t(114) = 3.73$ ,  $p < .001$ ,  $r = .20$ . In addition, participants with experience have much less of a game attitude,  $t(142) = 2.38$ ,  $p = .019$ ,  $r = .33$ .

Two important categorical variables we have not considered here: type and affiliation. With type I refer to what type of patroller: volunteer, regular employee, or expert employee. With affiliation I refer to the type of organization: Organization A, B, or C. Although I considered especially the latter throughout this level to elaborate on some of the results, I will save further consideration for both variables at a much later stage in this book, in Level 11.

## Lessons Learned

We come to find out that this population is not the most ideal target group for playing a complex digital game such as *Levee Patroller*. We had to deal with a relatively older audience with few computer skills, let alone game experience. However, about half of them had no experience with levee failures, so there was much opportunity to learn.

Regarding learning, participants indicated they expected to learn much and they said they did. In fact, on “virtually” every relevant aspect that they could learn, participants seemed to perceive they have learned. When aggregating all the individual results on levee inspection knowledge, we discover that the game had a strong effect on their perception. Participants’ knowledge perception increased with about 9% ( $SD = 11\%$ ),  $t(115) = -8.64$ ,  $p < .001$ ,  $r = .63$ .

Playing the game had an effect on participants’ perceptions of the inspection itself too. The game made them realize that inspection is harder than they initially thought. It further made them realize what the possible consequences are. Therefore, the game seemed to have heightened participants’ awareness about the need for inspection and training. It also seemed to give them confidence, because they indicated that inspecting has become more of a routine to them.

As for the game, although some participants clearly were not satisfied with it, most participants were positive about it, awarding it a firm seven out of ten rating. In the previous levels, I have already highlighted some issues the participants had with the game.

The relationships among the perceptions let us know that using games for training is not everyone’s cup of tea. Beyond the “traditional” variables we expect for a training to be successful, such as motivation and commitment, with using game-based training the more one has a positive affiliation with (digital) games and the necessary skills, the more likely the game will be successful. Although these aspects are most certainly influential, it needs to be stressed that their effect is not determining the outcomes completely. Many participants with no such affiliation and needed skills judged the game positively too and indicated they learned much from it.

The last goal of this level was to consider if any differences exist among the population sample regarding their perceptions. The purpose of this effort was to see how homogeneous the sample was and to explain for some of the results that were attained. Based on the background variables it already turned out that the three organizations are not one and the same: in terms of occupations, level of commitment, and playing games they differ.

As for the type of occupation, participants closer affiliated with the subject of levee inspection perceived (probably rightly so) to be more knowledgeable, yet those being less closer affiliated appreciated it more. They seemed to have learned relatively more from it, something which is confirmed by the mere fact that people with no failure experience had larger knowledge perception gains.

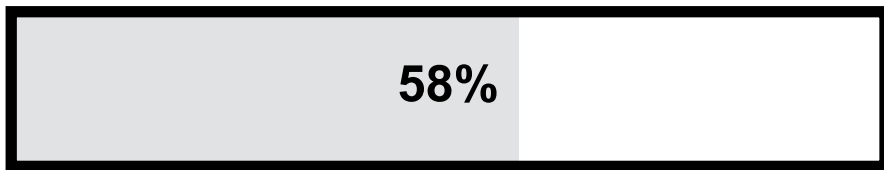
It further became clear that participants working in the agriculture sector were less appreciative of the training, but they also perceived to have less computer skills than others. Having such computer skills is considered clearly a critical factor. But ultimately, a dynamic web of relationships determine the outcomes of a game-based training.

## Level 8

### Picture This!

*A possible damage of the inner slope revetment caused by the tires on the left side of the car—Participant #94 about one of the non-failures*

*To me this does not seem the right place for disaster tourism. The driver might still be around—Participant #113 about the same non-failure*



The previous level dealt with the perceptions the participants had about the game and what they learned from it. Perception or self-assessment, although frequently applied as a proxy in evaluating training programs, is not completely reliable (Sitzmann et al., 2010). People either overestimate or underestimate what they learned and have generally difficulty in knowing what they learned. To get a more accurate indication, another type of measurement is needed. This is where the pictures come into play.

As outlined in Level 3, participants received a set of pictures of virtual and real failures before and after the training. The pictures served a more important role than merely getting to know if the game taught them anything. They were used as a way of exploring how the participants made sense out of real and virtual failures. By looking at how individuals performed on the test before and after (*within-subjects*) and how pictures were made sense of before and after (*between-subjects*), I was able to distill how participants pictured the pictures and what influence the game possibly had on this.

The goals of this level are to describe

- What the sensemaking test is and how it was organized;
- How participants made sense of the pictures before and after the training (*within-subjects*); and
- How the pictures were made sense of before and after the training (*between-subjects*).

## Organizing the Sensemaking Test

To go beyond self-assessment is why the *sensemaking test* was constructed. An additional reason was to be able to capture the phenomenon of sensemaking, as explained and defined in Level 2. The test was inspired by various mental modeling and cognitive mapping techniques who all fell short of the needs of this study, which was to capture how participants made sense, to test their performance, and to be able to measure an impact on communication. I will explain what this test is about and how the results have been analyzed.

### *Showing the Pictures*

The participants had to answer questions for seven pictures at the start-meeting and for seven pictures at the end-meeting, making them judge 14 pictures in total. Of these 14 pictures, exactly seven were pictures of failures that actually occurred. These are the *real pictures*. The other seven were pictures of failures from the game. These are the *virtual pictures*.

Ten pictures related to failures participants encountered during the training; the other four had another role. Two of these four “alternative” pictures were no failure at all. It was desirable to see if people would be able to distinguish failures from non-failures and how this might change after playing the game. The other two concerned failures the participants did not encounter in the game. In this case the idea was to see to what extent the participants were able to stretch what they learned to other areas: to see if they could transfer what they learned in the game to *new failures*, failures they did not encounter yet. With the non-failure and new failure pictures I also showed the participants one real and one virtual picture.

Based on this pool of 14 pictures, two sets were defined, each with seven pictures. I refer to these sets as *Set A* and *Set B*. Each set has one real or virtual non-failure picture, one real or virtual new failure, two or three real pictures, and two or three virtual pictures. To rule out an *order effect* “between” the two sets, approximately half of the participants had to look at Set A (Pictures 1A to 7A) before the training and at Set B (Pictures 1A to 7B) after the training, while the other half looked at Set B before the training and at Set A after. Otherwise, the content of the sets may have confounded the results: if only Set B was received after the training and had been more difficult or easy, the results would have been underestimated or overestimated, respectively. Who received what set first was randomly decided before the start of the training.

To rule out an order effect “within” the two sets, the sequence of types of failures was kept the same for both sets. This effect seemed less problematic than the order effect between the sets, but it could have been possible that what failure participants judged first may have had an influence on their subsequent sensemakings. To determine the sequence within the sets all pictures were coupled. Each couple had a real and virtual picture. So I defined, among others, a stone damage couple and a

**Table 8.1** Overview of the pictures the participants needed to judge

Sequence order	Failure type	Set A (1 to 7)	Set B (8 to 14)
1	Stone damage	Virtual picture 1	Real picture 8
2	Boiling ditch	Real picture 2	Virtual picture 9
3	Small landslide	Virtual picture 3	Real picture 10
4	Illegal driveway	Real picture 4	Virtual picture 11
5	Non-failure	Real picture 5	Virtual picture 12
6	Watery slope	Real picture 6	Virtual picture 13
7	New failure	Virtual picture 7	Real picture 14

new failure couple. If participants received at the start-meeting the real picture of a couple, they would receive the virtual one at the end-meeting. The reverse was the case if they received the virtual picture of a couple. Table 8.1 gives an overview of the exact sequence order and type of failure the participants had to judge.

This coupling of failure pictures was done for another and more important reason too. This enabled comparison at the level of the failures. For example, it made it possible to compare how participants judged the stone damage failure before and after the training. This concerns a within-subjects assessment for an individual picture and I took two possible combinations into account. One combination is that participants received the virtual picture and subsequently the real one (e.g., Combination 1A–1B). The other combination is the reverse (e.g., Combination 1B–1A).

For each picture, the same set of questions were asked. Each question in this set relates to one of the learning objectives of the game. The specific questions were:

1. How do you assess the situation? [Assessing]
  - *No failure*: Nothing is happening. Nothing has to be reported.
  - *Reportable*: A failure can be noticed, but it does not cause a real problem. Monitoring is still required.
  - *Severe*: The failure should be taken seriously and monitored constantly.
  - *Critical*: The situation is out of control. Measures need to be taken immediately.
2. What do you see? [Observing]
3. When reporting this failure, to what should you pay attention to? [Reporting]
4. What failure mechanism is occurring? [Diagnosing]
5. What measures need to be taken? [Taking measures]

The first question, about assessment, was put first and concerned unlike the other questions a closed question. I had a number of reasons for this arrangement. The most important reason had to do with data interpretation. Although this research required open questions, to see how sensemaking takes place, doing so for the assessment question would be problematic. People use different words to describe the severity of a situation and also value those words differently: the word “dangerous” for one person might mean “evacuate immediately” and for somebody else it may

just be an indication that something should be paid attention to. It therefore seemed best to define the situations as narrowly as possible.

Another reason to define the answers up front was to make it possible to eventually see the effect of the game. *Levee Patroller* has three assessment categories—reportable, severe, and critical—and it would be hard to unravel how all of the answers would relate to any of these categories. By making sure the offered options would reflect these categories this difficulty was solved.

The third and most practical reason involved the ability of removing questions. If somebody would answer that the picture did not concern a failure, it would make no sense to fill out the subsequent questions. To prevent this from happening, the questions would need to disappear. By linking the appearance of the questions to the answer on the situation assessment this became possible. Similarly, it became possible to show the question about measures only if participants indicated it concerned a critical failure. Although it could have been reasonable to ask participants what measures would be needed if the situation became worse, this would require them to fill out more questions. As taking measures is not the most important learning objective and the test was already lengthy enough, I decided it would be better to only show this question if participants determined the picture showed a failure.

### *Analyzing the Pictures*

For the purpose of the analysis, I coded the responses. While coding I tried to adhere to the exact wording of the participants as much as possible (except for Question 3, about reporting the failure). For example, if somebody would say “rock” and another one “stone,” I would code both as rock and stone, respectively. My reason to code literally was to catch the diversity of the responses. While rock and stone may be categorized similarly, both may have different connotations in the field or could evoke a different mental image. For me the word rock relates to mountains, to big things, and spiky structures. With stone I think of streets, throwing peddles in the water, and of crafted, well-organized structures.

I abandoned this literal coding if:

- A person made a description in which two or more words occurred that together would comprise a good label. In that case I coded the label and not one of the separate words. For example, as for describing the pitching stone, saying “stone revetment” is much more accurate than saying either stone or revetment. Some participants would, however, not directly say “stone revetment” but rather say “Stones are missing from the revetment.” In those situations I would still code this as “stone revetment.”
- Words had been differently conjugated. For example, aside from “bulge out” you can write bulges out, bulged out, and bulging out. In these circumstances I chose one code and applied these to all conjugations. But I did so with the caveat that if one of the conjugations was part of the exact phrasing from the

game, then I would use two codes, one for the conjugation corresponding to the game and one for all other conjugations.

Although literal coding seems straightforward, I still had to pick the right word(s). Sometimes this was not that clear-cut. Or I had to revise earlier codes, because of further insights. Out of consistency and to minimize any errors on my behalf I applied a three step approach to my coding. During the first step I coded the responses before and after of one picture. The second step involved checking these codes with all possible codes in mind of the picture I was working on. After I coded all the pictures this way, the third step was made. This involved checking all the codes again, with all the coding from all the pictures in mind. My coding continually evolved, in line with Straus and Glaser's (1967) discussion of grounded theory. It only stopped evolving after I had finished all pictures. When that happened it automatically required me to check all previous pictures.

To judge the codes, I looked at three criteria based on the outcomes I intended to measure with the test: accuracy, word count, dispersion, and vocabulary. Each of these criteria shed some light on the meaning of the responses and how the game possibly had an influence. Accuracy concerns the operationalization of sensemaking performance and the other three criteria are operationalizations of communication for this particular test (Level 3). The criteria are further explained below.

*Accuracy* Whether a person uses a few or many words, it is above all indispensable to provide accurate information. For determining the accuracy I used the logic and vocabulary from the game as a benchmark. Although the game was developed closely with experts, making the game a valid benchmarking tool, different categories were used to label the literal codes and one of such categories concerned valid alternatives. By comparing the accurate observations, based on the game or not, against less accurate ones, it can be seen to what extent the game helped to become better in providing accurate information.

*Word count* The number of words a person uses to explain a phenomenon is indicative for two possible conclusions. Either a person is highly knowledgeable about a subject: "I see missing pitching stone. This is a severe situation, because if the water level will rise the levee may quickly erode." Or a person cannot catch the phenomenon with a certain label and has to revert to wordy descriptions: instead of saying "missing pitching stone" a person may say "I see an organized set of stones, used for protection of the levee, and one or two of these stones are missing." The latter is certainly not wrong, yet it can be ineffective in communicating with others. The use of *Levee Patroller* may cause both to happen. Where participants first may not know what to say, they may have an elaborate mental model afterward and are able to describe in detail what they see. In contrary, it may also happen that they let go of wordy descriptions and instead only use the label they have learned to use in the game.

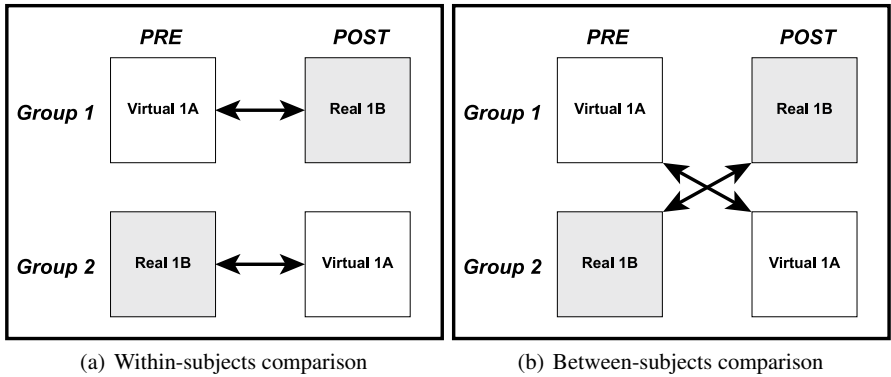
*Dispersion* Another reason to code literally was to see the wide variety in responses and observe how this might alter after playing. It was expected to see a decrease in variation, as participants would probably converge toward using content from the game. To prevent any distortion in the dispersion percentage indicators, I



excluded the codes, “No idea” and “Not filled out.” For certain questions these numbers were relatively high, making it seem that the participants converge very much. In fact they do but in a not so positive way.

*Vocabulary* Vocabulary is quite important in organizational settings, as I have argued elsewhere (Level 2). To see the game’s impact on this particular aspect, I looked for the use of keywords per picture per organization. The keywords were determined by the vocabulary from the game. Any conjugations or derivatives were also considered. Then I simply counted the number of times, out of all the words used, the keywords were used. To put things in perspective, I also counted a number of frequently used words. These were “de, is, er, van, dat, het, een, en, of, op, in.” These are some Dutch articles, prepositions and pronouns. The English equivalents are “the, is, there, from, that, the, a(n), and, of, on, in.” My expectations were that participants’ vocabulary would consist afterward of a higher percentage of keywords.

These four criteria are applied to almost each of the five questions. Almost because word count and vocabulary do not apply to the assessment question for which the participants had to pick their choice. A number of additional aspects and hypotheses were further considered based on new insights while coding, such as whether participants became more situation aware (operationalized as mentioning the need to report the location and a mentioning of the crosscut location of the failure signals) and noticed more detail in the pictures. All of these additional considerations did not lead to any conclusive findings or significant differences and so have not been included in this level. However, this possibility to analyze the data from various perspectives is a great advantage of open questions over closed ones.



**Fig. 8.1** Comparing the picture responses before and after the training

Various analyses have been considered with the four criteria in mind (Fig. 8.1). The first are within-subjects analyses on an overall level. On this overall level totals and subtotals have been summarized and compared with each other. Subtotals concern answers per learning objective and failure classification. One such failure

classification concerns a subtotal of virtual and real failures. Another is a consideration of a *core set* of pictures. This core set excludes the four alternative pictures.

The second are within-subjects analyses on the level of pictures. Here different couple combinations have been considered (Virtual–Real and Real–Virtual). These analyses were made to see for what pictures the game’s effect seemed strongest and to understand why effects may or may not have taken place.

The third are between-subjects analyses. This time the answers of one particular picture have been compared before and after. Because different participants were involved, this was necessarily a between-subjects analysis, creating a bias caused by individual variability. However, with this analysis it became possible to compare answers on the same picture, which is unlike the within-subjects analysis of the couple combinations not biased by variability between the pictures.

Only the most relevant findings are reported in this level. In addition, the level focuses foremost on the accuracy criterion. This criterion single-handedly defines the sensemaking performance, which is one of the main outcomes. At the end of this level I will elaborate on the other criteria, which relate to the secondary outcome of communication.

## Picturing the Learning Objectives

Before presenting the overall results, I will first discuss the findings in terms of accuracy per learning objective. As discussed earlier, for each of the learning objectives a dedicated question was constructed on the sensemaking test. The ideal, “accurate” answers are explained throughout this level per specific failure couple.

## *Assessing Remains Daunting*

Making an assessment *can* be easy. This is the case when clearly defined categories exist and the object of investigation leaves little to no ambiguity as to what category it belongs to. However, even in situations when we think we are dealing with such easy identifications, we encounter problems. Similar observations can be made about other apparently easy identifications (Bowker & Star, 1999).

When no clearly defined categories exist and the object of investigation is rather ambiguous, we enter the world of rather difficult identifications. To this world most risks belong. Risks are difficult to assess and for a number of reasons: many factors have to be taken account; a lot of uncertainty exists; and little information is mostly available. With this in mind, a discussion about the assessment of a certain risk is inevitable.

Therefore, asking the participants to assess levee failures makes for a daunting task. Like most risks, the identification of levee failures presents special challenges. Even the experts continue to disagree among each other (Level 10). Experience

plays an important role in the assessment. If no validated standards are in existence, such as with levee failures, people rely more heavily on other information sources. If experiences differ or other information sources are used, different conclusions may very likely result.

Important to mention is that pictures are already ambiguous by their very nature. They are snapshots of a certain situation in a certain time and this particular information cannot be fully retrieved from the pictures themselves, opening them to multiple interpretations. To minimize interpretive ambiguity, I added the following statement before the start of the test.

In the test you will have to answer some questions regarding a number of pictures of levee failures. To answer these questions, imagine you are asked to go on a patrol in an emergency situation.

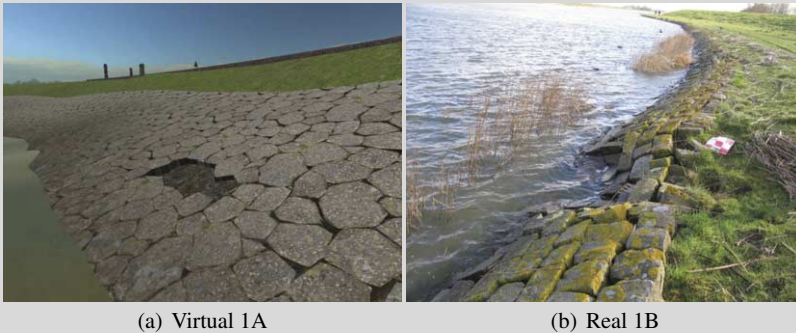
Thus, participants would not assess the pictures as if they encountered them during a regular inspection. For such an inspection the sense of urgency is much less. No river or sea is about to cause any trouble, so observations are logically differently evaluated.

Something else to reiterate is that assessing is “officially” never the responsibility of a levee patroller. Of course, in practice their opinions are heard and the severity of the situation is discussed when a failure is reported. But principally this is a task delegated to others. The purpose of including this in the game is manifold, but the main one has to do with giving the players a better sense of what is severe and what is not. As the purpose was to increase their risk sensitization, the importance of being able to accurately assess failures is much less compared to the other learning objectives. In addition, as even the experts still disagree about the severities attributed to the pictures, it is hard to speak of right and wrong in making an assessment.

Yet, like in the game one has to give scores to say something about the performance of the player, in the evaluation one has to say something about the accuracy of the responses. To determine the severity of pictures the categories from the game were used: non-failure, reportable, severe, and critical. What only needed to be decided on is to determine which of these categories belongs to what picture. For the virtual pictures, this was already done. For the real pictures the classification took place by applying the rules from the game. Validation was done by comparing the choices with those made by experts (seven in total). Little agreement existed among experts, but it turned out that in most cases a small majority was in agreement with the choices based on the game. In case no majority appeared and opinions differed, the choice of the game was maintained. Only for two real pictures (Pictures 6A and 7B) the initial game choice was changed as a consequence.

Then to decide on the accuracy, a binary position was taken: either the participant was accurate (choice = predefined categorization) or was not (choice != predefined categorization). On average 45% of the participants were accurate before the training, while 57% were so afterward (within-subjects). This is only a modest increase but still significant. Consideration of the failure couple combinations reveals subtle improvements for most. Participants excelled especially at the non-failure pictures already (Combinations 5A–B and 5B–A). For both non-failure couple combinations

### Stone damage failure couple



**Fig. 8.2** Pictures of the stone damage failure couple

- *Assessing*: Severe. If stones are missing the levee is directly exposed to erosion. Only a bit of water will already cause damage.
- *Observing*: Pitching stone. For Real 1B it would also be correct to mention a “settlement.”
- *Reporting*: (1) If stones are missing or moved; (2) the length and width of the failure; (3) if soil is flushing and if so how much.
- *Diagnosing*: Erosion outer slope.
- *Taking measures*: If it would get worse, foil with sandbags is appropriate.

more than 70% of the participants noticed this before and after. The non-failures were apparently easy to discover.

We need to be somewhat wary of any improvements by considering the couple combinations, because they might be a result of a difference between the two pictures. In fact, if we consider the pictures individually and observe how they were assessed before and after between participants (between-subjects), we see less significant differences. In this case only five out of 14 pictures were assessed differently afterward. If we assume the groups are equal, these results are striking. They indicate that changes only occur if:

1. Participants learn a clear rule which they can apply to determine the severity, even in situations they have not seen before, such as “If pitching stone is missing, it is always at least severe” (based on Pictures 1A and 1B).
2. Participants actually realize that something is a failure (based on Picture 4A).
3. Participants unlearn their initial intuitive guesses by being confronted with exactly the same situation as they practiced before (based on Pictures 2B and 3A).

The first changemaker has a strong effect and is largely similar for the virtual as well as real pictures. Indications exist that the transfer may even go too far, as signs of *overgeneralization* appear from the data. This is based on the fact that 20% of the

participants perceived a similar failure as also severe, while only 4% thought this before the training (Picture 7A). This failure is, however, less severe. As this failure was not part of the training, the participants did not know any better and a good amount applied a rule they learned. It is as if they only knew the rule of adding -ed to a verb to signify the past tense without knowing about any of the exceptions.

Transfer may also happen if people are confronted with situations they were unaware of; for example, the temporary sand road on top of a levee (Pictures 4A and 4B). This reveals a clear improvement in the number of accurate responses—twice as many participants were correct afterward. This result is quite strong, especially if we take into account that not everybody encountered this situation in the game, as this failure was only included in the last two levels; some ignored it while playing, because they did not perceive it as a failure, and others maintained the opinion that it did not concern a failure. The latter opinion is defensible, as in and of itself, it is not dangerous. Only when other signals come into play, such as a high water level, does the situation become problematic.

One interesting aspect regarding these pictures is that initially a significant change can be observed for the real variant and not for the virtual one. An explanation can be found in the difficulty of participants to interpret the virtual picture beforehand. Some did not see an extra layer of sand on top of the levee, but actually thought they were looking at a big gap in the levee. Consequently, about 25% of the participants thought this was a severe to critical situation. If we correct for this, by excluding those who misinterpreted the picture, the virtual version also becomes highly significant. Nevertheless, this example highlights a possible need for people to learn how to “read” virtual images.

The third and final changemaker is about applying what is exactly learned in the game. At the start, participants had to rely on their intuition, but afterward they could rely on their experiences and answers from the game when confronted with a virtual picture. They would recognize the situation and remember that it should be non-reportable, severe, or critical. With this changemaker, what was learned did not have a clear transfer to the real equivalents.<sup>1</sup>

The assessing results suggests participants moderately improved their accuracy. This only happened when participants learned a clear guideline that they could apply in a situation where they have to make sense. Such a guideline could be a clear rule, a new insight, or a previous experience. These guidelines are important, because the real world—just as the real pictures—are in most cases more ambiguous than the virtual world. From these results we can see that the game aided in this direction. Nevertheless, assessment remains a daunting task.

---

<sup>1</sup> However, for Pictures 2B and 3A it should be noted that one real equivalent was too easy and rather unambiguous to answer (Picture 2A) and the other was a bit too ambiguous (Picture 3B). This may possibly explain why the transfer did not occur.

## ***Observing Improves Conditionally***

Unlike assessment, observing is at the core of what levee patrollers do. To fulfill this task competently, a patroller first must notice what needs to be noticed, and then communicate it properly. This touches on the content and vocabulary of the observation.

*Levee Patroller* was designed to assist in both learning objectives (Level 2). By being confronted with virtual failures, players would recognize what possible failures could occur. Hypothetically players would subsequently know better where to pay attention. And for an effective communication to occur, it helps if the people involved know what they are talking about and have a similar association with what is said. In other words, a *shared mental model* is what is needed. By extensively practicing with a certain framework such as implemented in the game, standards will arise for communication, making it eventually much more effective.

To determine the effects of the game, participants were asked to tell what they saw on the pictures. In their answers I looked for a specific label, consisting of one or more words, that describes the signal(s) on the pictures. Signals are indications of a failure and the types of signals patroller can encounter are classified in the game into specific categories with textual labels. Examples of these labels are pitching stone, settlement, crack, water outflow, and liquefaction. Ideally, from the perspective of the game as intervention, participants would start using these labels to describe what they see and, of course, apply them correctly.

But if patrollers use another well-established term or even simply describe what they see, this does not mean it is incorrect. Therefore, for judging the accuracy of this core aspect of the inspection, the following rules were employed:

- *Very accurate (VA) = perfect fit*: The observation is literally similar to the textual label from the game.
- *Accurate (A) = near perfect or good alternative*: The observation is closely similar to the text from the game. Participants may have used a synonym (pitching rock instead of pitching stone) or used a proper replacement word (sand outflow instead of water outflow). In addition, for many of the pictures good alternative labels can be used. These labels are well-established in the field as well.
- *Slightly accurate (SA) = descriptive, a bit vague or failure mechanism*: This category is applied to observations that are not necessarily wrong but very descriptive (stones are missing), or when vague language is used (bubbling water instead of water outflow). And finally it happens that people directly mention the failure mechanism.<sup>2</sup>
- *Inaccurate (IA) = incorrect or too vague*: The last category corresponds to wrong observations. Wrong sounds harsh, because on many occasions people

<sup>2</sup> For one picture couple, the boiling ditch couple (2A–2B), I made an exception for judging the mentioning of the failure mechanism as slightly accurate. This type of failure is rather specific—only one sort of failure mechanism can be associated with the observed signals. It is quite common that people directly mention its failure mechanism. For this reason I categorized the mentioning of the failure mechanism for these pictures as accurate instead.

simply interpreted the picture differently. But compared to the ideal answers, these different interpretations are incorrect. Then we have observations that are just too vague; they ask for more questions than they answer (dredge water instead of water outflow).

For three pictures—Real 1B, Virtual 3A, and Real 3B—some extra coding was needed, since they show two signals instead of one. This is not something special. In the game many failures have more than one signal. To take this into account a distinction was made between the *main signal* and a *contributing signal*. The main signal was chosen on the basis of what is more typical of the failure situation as depicted in the picture. For example, a horizontal movement says more about what type of failure is occurring than a crack (see Virtual 3A).

### Boiling ditch failure couple



(a) Real 2A



(b) Virtual 2B

**Fig. 8.3** Pictures of the boiling ditch failure couple

- *Assessing*: The real picture is severe, the virtual one critical. The difference is based on the amount of flushing soil.
- *Observing*: Water outflow.
- *Reporting*: (1) The width and length of the damage; (2) the velocity of the water flow; (3) the amount of water; (4) if it is just one or multiple; (5) if soil is flushing; (6) and if so how much.
- *Diagnosing*: Sand boils.
- *Taking measures*: Sandbag containment ring.

In addition, in their descriptions a greater number of participants ignored the contributing signal and not the main one, confirming the choice. Furthermore, few participants—an average of 10% for the three pictures—only described the contributing signal and not the main signal. Quite striking is that this percentage is similar for before and after playing the game.

The number of people who mentioned the contributing signal before and after at all was similar as well. For two pictures (Real 1B and Virtual 3A), about 60 to 70% mentioned it; for the third picture (Real 3B) this was much less, about 30 to 40%. The contributing signal on the latter picture was less visible and lesser known, making this difference understandable.

Thus, it should not be a surprise to find out that in terms of accuracy no difference can be found.<sup>3</sup> But as the game stresses very much that failures consist of one or more signals and they are confronted with this many times, I expected them to be much more aware about this possibility. Consequently, I thought the participants would perform much better afterward. Although the number of pictures with more than one signal is too low to draw rigid conclusions, I think two important reasons exist why the accuracy did not increase:

1. Except for one contributing signal (liquefaction in Real 3B), the signals can be described in laypeople's words (settlement in Real 1B and crack in Virtual 3A). No increase in terms of vocabulary should have been expected; and
2. Participants simply did not need to learn to see and describe multiple signals. This comes naturally, although it may have been fostered by the nature of the task. Much like the well-known "find the X differences"-pictures, participants had to focus on a clearly demarcated area and kept looking until they found all signals. It may well be that in other circumstances a raised awareness about the possibility of multiple signals becomes apparent.

The consequence of this result is that a decision was made to not take the contributing signals further into account. They would already make it difficult to make an equal comparison between the pictures and would make an uneven contribution to the overall scores.

In contrast to the contributing signals, major differences can be observed with the main signals.<sup>4</sup> In the case with the strongest effect 71% of the participants used the exact label from the game whereas only 4% used this before playing (Virtual 1A with "pitching stone"). With the other pictures, about 20 to 50% were using the exact labels compared to mostly 0% before. With only two other pictures (Real 3B and Virtual 7A) some participants already used the game labels as well ("settlement" and "floating waste," respectively).

Considering the significant length of game playing, this may even seem a not so strong effect. Scores much closer to 100% could have been expected. A significant number of reasons exist to account for why this did not happen. But first it is important to highlight that as many participants may not have been very accurate (VA) afterwards, they did become accurate (A). Except for two pictures (Real 2A and Virtual 2B), which I will elaborate on in a moment, more than half of the participants

<sup>3</sup> No difference was found between-subjects in observing the contributing signal, for Real 1B  $U = 2148$ ,  $p = .40$ ,  $r = .072$ ; for Virtual 3A  $U = 2070$ ,  $p = .11$ ,  $r = .13$ ; for Real 3B  $U = 2218$ ,  $p = .46$ ,  $r = .062$ .

<sup>4</sup> This is based on the between-subjects as well as within-subjects results. Only with the latter some results were not significant. This was due to a difference in difficulty of the pictures, not because participants did not improve.



were either slightly accurate (SA) or inaccurate (IA) before. After playing the game this reversed—therefore, a clear shift toward more accuracy can be clearly seen.

One explanation for why not everyone achieved an accurate score is that a significant number of participants used words that resembled the labels. Sometimes they reversed order (stone pitching instead of pitching stone); used a synonym (human action instead of human activity or animal activity instead of biological activity); replaced one of the words (sand outflow instead of water outflow); confused signals (horizontal settlement); or were merely incomplete (movement instead of horizontal movement). At times, participants even made up their own labels based on the game. For example, one participant did not speak of settlement but rather of a “vertical movement.”

One reason why not everybody reached a very accurate score is that a good amount of participants used words that resembled the labels. Sometimes they reversed order (stone pitching instead of pitching stone); used a synonym (human action instead of human activity or animal activity instead of biological activity); replaced one of the words (sand outflow instead of water outflow); confused signals (horizontal settlement); or were merely incomplete (movement instead of horizontal movement). At times it even happened that participants made up their own labels based on the game. For example, one participant did not speak of settlement but rather of a “vertical movement.”

Another reason is that a good alternative label existed and that because of this not everyone switched to the game label. This became particularly evident with the sand boils failure set (Real 2A and Virtual 2B). This set has already four good alternatives<sup>5</sup> and participants were already familiar with this, which is noticeable from the fact that over 60% described the failures accurately from the beginning.

A third reason arises from the intuitive use of labels. Some of the labels are simply not self-explanatory. Horizontal movement and liquefaction beg for some explanation for example. While subtle differences exist between these types of signals and something like a settlement, the end result looks quite similar: a mud or landslide occurs, creating a gap or hole in the levee. Maybe due to the non-intuitive use of labels and due to the subtle differences many participants tended to stick to rather general descriptions—something I call the “commonsense approach.” Nothing is fundamentally wrong with this approach, but it is simply less accurate than using the exact labels.

As the commonsense approach did not occur with all pictures and labels, it raises the question of when and why a label is expected to find a wide application. Based on the previous analysis, some factors can be identified: *a*) the strength of the label itself (is it catchy and self-explanatory?); *b*) the existence of competing labels (is it any better than others?); and *c*) the ease of application (can it be clearly recognized?). The label “pitching stone” fulfills all of this and for this reason its effects are much stronger compared to the others.

These results raise questions about whether a game is the appropriate vehicle to transfer something like textual labels. Games may not be very suitable for transfer-

<sup>5</sup> In Dutch the four good alternatives are (1) piping (the failure mechanism term), (2) kwel, (3) wel, and (4) zand(mee)voerende wel.

ring declarative knowledge, such as textual labels. Based on this study, it is impossible to give a full answer, yet we can at least say that this game—even while much room for improvement exists—gets the job done.

### ***Reporting is Poor***

Observing signals is the first and foremost task delegated to patrollers, yet upon finding a signal it is also the start of a whole new chain of events, including to report the signal. This reporting is not merely mentioning what the patroller sees. Reporting involves specifying what the patroller sees. Not every crack is one and the same. Cracks differ in type, length, width, and many other characteristics. The same can be said about many of the other signals. Reporting is thus more or less a characterization of the failure situation.

Such characterization is necessary, as depending on this description an assessment can be made about the severity of the situation (assessing) and of what is possibly happening (diagnosing). While necessary, reporting procedures are hardly formalized at many organizations. Attempts have been made to organize and structure this, but either this remained paperwork or in practice this found hardly any application. For this reason, at the start of training, the expectations regarding reporting were very low.

It is important to mention that *Levee Patroller* was especially praised for its rigorous reporting procedures by the clients and subject-matter experts. This was confirmed during the discussions at the end of the training (Level 9). Therefore, the expectations about having a strong effect on reporting were very high.

To see whether this was true the coding occurred quite differently than with the other questions. The responses were not literally coded but categorized on the basis of what reporting items were mentioned. Each signal has a number of associated reporting items. For example, if patrollers encounter a settlement, they should pay attention to the following reporting items: (1) length and width; (2) height; (3) type of revetment; and (4) direction. In coding the answers I checked what reporting items were mentioned by the participants and I neglected what wording they used. If a participant spoke of “dimensions” or “size” instead of “length and width,” I applied the same code because all of the labels have the same meaning, which is to measure the signal, and therefore refer to the same reporting item.

The reason I neglected labeling with reporting is that regarding this learning objective it is far more important to mention the reporting items than to use the exact labeling from the game. We rather have a patroller report all four reporting items with different labels than just one with the exact corresponding label. I could still have considered labeling, but because of its lesser importance, the reporting accuracy was determined by the number of mentioned reporting items compared to the ones that should be mentioned according to the game’s reporting procedure.

While coding, I noticed that participants mentioned many *general reporting items*. General items include, among others, mentioning the location, getting to

know the history of a failure situation, or ensuring one's own safety. Such items are true for any failure found and so they are not so interesting from the perspective of seeing whether patrollers know what to focus on for a particular failure situation. Instead, what I was looking for were *signal-specific reporting items*. Although many signal-specific items overlap for various signals, such as the length and width, each has a specific importance in describing a certain signal.

For coding the responses, it became quickly clear that often items formed answers to the other questions; for example, the signal-specific item "if stones are missing or moved." Most participants answered this right away in their observation, stating "stones are missing." Another example concerns the item "type of revetment." On many occasions participants directly mentioned this by saying "the asphalt or grass is damaged." This necessitated considering the answers on other questions as well.

I further grouped a number of signal-specific items for which I could reasonably assume participants *a)* used shorthand descriptions for what was needed (measuring the "size" instead of "length" and "width"); and *b)* implicitly knew what was needed (if one says the amount of soil should be noted it speaks for itself that the flushing of soil should be observed). I also considered grouping items if the items did not determine the severity of the situation and/or did not play a clear role in the game. An example of the latter are the items "type of crack" and if "one or multiple cracks occur." Both items need to be reported, but I considered it sufficient if either one of them was mentioned.

As the number of desired signal-specific items differs per picture, no standard categorization was used at first. The number of correct items were simply counted. To make it eventually possible to compare the pictures between each other, the following categorization was applied after that:<sup>6</sup>

- *Very accurate* =  $> 50\%$  of items correct: Participants mention more than half of the desired items.
- *Accurate* =  $\leq 50\%$  of items correct: Participants mention half or less than the desired items.
- *Inaccurate* =  $0\%$  of items correct: Participants mention none of the ideal items.

This is a very rough yet necessary categorization as some of the pictures only had one or two correct items. The difference between the lower item pictures and the higher item pictures would otherwise be quite skewed.

Despite being very lenient toward the number of ideal items and contrary to the expectations, I found eventually that the responses were quite poor—before as well as after the training. For the core set of pictures an average score of 24% was achieved before playing compared to 34% after playing. Only 15% could be considered "very accurate" before and 29% afterward. While a significant improvement can still be seen, judged on their own and compared to the results of the other learning objectives, they are still rather disappointing. This may have been due to the

---

<sup>6</sup> To verify that this categorization did not have a large influence, the results have been calculated with and without this categorization. No significant differences could be noted.

### Small landslide failure couple



**Fig. 8.4** Pictures of the small landslide failure couple

- *Assessing*: Severe. Clear indications exist that the levee is moving inwards. The real one is even close to critical.
- *Observing*: The main signal for Virtual 3A concerns a horizontal movement; for Real 3B this is a settlement. Both have one extra signal. Virtual 3A has a crack on the crest, while Real 3B has liquefaction.
- *Reporting*: For Virtual 3A (1) the type of revetment; (2) the length and width of the failure; (3) what type of crack and/or if one or multiple cracks can be seen. For Real 3B (1) the height; (2) the length and width of the failure; (3) either the direction of the settlement, if soil is flushing, or the type of revetment.
- *Diagnosing*: Macro-instability.
- *Taking measures*: A sand berm would provide a counterweight to the movement.

format. The question, how it was formulated and its openness, the length of the test, and the difficulty some participants had with typing may not have invited all of them to express what they know.

It could also be a consequence of a lack of respondents to fill out tests such as these. As the question did not mention a specific minimum of items, participants may have acted as “satisficers.” They were not out to maximize their answer; instead, they just minimally answered it, by at the very least mentioning one item. Such minimal answers were sufficient for the other questions, but the reporting question required for all but two pictures more than one answer. It should, however, be noted that some participants acted more like “maximizers,” for instance:

This is not necessarily a failure right away. Only in combination with a settlement or a crack development and an imminent threat of overtopping/overwash a problem exists. If we look at the failure, attention should be paid to: size and type of material to know the ground pressure on the levee; consolidation signals such as crack development and deformation; damaged revetment in the surroundings of the work; and if a permit was given to allow this construction—Participant #138 about Picture 4A<sub>post</sub>

Based on the answers, an alternative and complementary explanation is that in the eyes of the participants too little variation exists in reporting the failures. For example, many of the failures always require the length and width to be measured. In addition, many participants had a tendency to mention general items, items that are always applicable to any failure. This generalizability lessened the need to be specific about a particular failure and may have given the impression that the question was answered sufficiently. Consistent with this idea, a number of participants said “See other answers,” indicating that they found the same set of items relevant for different failures. This is an example of a typical answer:

The exact location, size of the failure, possible consequences—Participant #119 about all pre-test pictures

It further appeared that for more than half of the pictures, little to no differences can be seen in the number of participants reporting size as a relevant item for each picture. Of all items mentioned over all pictures, this non-difference can also be seen. Size was reported 18% at the beginning and 19% at the end, demonstrating that *a*) either the game did not improve the player’s awareness about the need to report failure sizes; or *b*) players were—contrary to the belief of the organizations—already quite aware of this.

To investigate causes for such poor results, I compared the signal-specific items of each picture before and after and found a remarkable pattern:

1. A signal-specific item improves if it is a highly characteristic consequence of a failure. For example, the outflow of soil is not just one of many items to pay attention to—it is actually *the* primary consequence of what can be seen if the boiling ditch failure occurs; or
2. A signal-specific item improves if it can be deduced from the situation. With the illegal driveway couple, it was striking to see how many participants started mentioning the search for other signals and in particular of settlement. The latter can be logically explained for: if an illegal driveway is implemented, heavy traffic will make use of this, and this may damage the levee. Similarly, it can be deduced that if revetment is missing, soil outflow becomes possible; and
3. For all other signal-specific items, ones that are harder to explain for, do not seem immediately logical, or are possibly considered less relevant, they may either decrease or increase. What happens might be dependent on certain cues in a picture, but in all cases the increase or decrease hardly yields any significant results. These results are simply part of the variation in responses.

This is my explanation for this pattern: Not only are the improved ideal items more characteristic or logical, they are also explicitly part of the game. They are more than an item on a virtual checklist. The outflow of soil is for example simulated in the game and can become more over time and, to mention another example, a settlement can be observed with the sand driveway and needs to be separately reported. All the other items are not simulated (accessibility of the area), hardly distinguishable (velocity of water), or are only part of a checklist (type of revetment).

Better results may have been achieved with a different question. As general items were of less interest, it should have been prevented that participants mentioned these, by for example reformulating the question to stress the need for signal-specific items.

Moreover, better results may have been achieved if more items had a more prominent role in the game. From the game it should have become clear why items matter and preferably by seeing how this “plays” out. Only mentioning items as part of a checklist and practicing this repeatedly seems to not work so well. These results stress the importance of connecting content to game mechanics.

### ***Diagnosing Performance Quadrupled***

In contrast to reporting, the results on diagnosing are very promising. In percentages, the accuracy quadrupled (from 13% to 54%). This diagnosing concerns the identification of the *failure mechanism* behind a failure situation. A failure mechanism is a typical way in which a failure develops. We identified in total five failure mechanisms: erosion inner and outer slope, macro-instability, micro-instability, and sand boils (Level 2).

Now the idea of diagnosing is that based on the signals and how a failure develops, a patroller is able to recognize its failure mechanism. Such recognition requires a more elaborate understanding of the behavior of a levee. It requires to integrate all the signals from the failure situation and mentally simulate what is occurring or what might be happening. For example, seeing a water outflow with soil should lead to a mental model of a levee underneath which a pipe is created and sustained and that slowly but steadily becomes longer and longer until the whole levee collapses.

This failure mechanism recognition is highly important. Based on the type of mechanism an appropriate measure could be taken. Otherwise it could very well happen that some measure may actually accelerate the degradation of a levee. While considered important, at most organizations diagnosing is not a responsibility of the patrollers and certainly not for those who are volunteers.

Despite this, during the inspection courses patrollers are informed about the failure mechanisms and during the development of the game we also found it important to include it, although it only comes down to choosing the correct mechanism from a list of the five possible mechanisms. But to choose the correct one, players do need to know what the terms mean. They can look this up in their handbook in the game, on the website, or in the paper game guide.

During the pre-interviews, it became quickly clear that the participants had little to no idea of what a failure mechanism is (Level 10). On the pre-test, 44% of the participants immediately said that they did not have any idea what failure mechanism is occurring. Presumably most others were guessing, considering the low accuracy score of 13%.<sup>7</sup> Not everybody was able to diagnose afterward. Yet, with on average

<sup>7</sup> The averages on participants having no idea what failure mechanism is occurring are based on the core set of failure pictures.

**Illegal driveway failure couple****Fig. 8.5** Pictures of the illegal driveway failure couple

- *Assessing*: Reportable. In itself the driveway is not dangerous, although it does damage the grass underneath it. Combined with other signals, such as overtopping, it can however become quickly severe.
- *Observing*: Human activity.
- *Reporting*: (1) The type of human activity and/or if a permit was granted for this; (2) if the levee is damaged; (3) if other signals can be observed.
- *Diagnosing*: Erosion inner slope.
- *Taking measures*: Foil with sandbags on the inner slope will be sufficient.

12% not knowing what failure mechanism is occurring, the group of not-knowers had become significantly less.

Important to note is that many of the participants were familiar up front with certain of the specific failure mechanism terms—especially with sand boils. They just did not know the exact meaning of these terms, something which is not entirely their fault. In practice and also in the literature, the terms for signals and failure mechanisms are often confused or used interchangeably.

Even after playing the game, some participants kept on mixing both. This mixing is also understandable from a practical perspective. If people have repeatedly made sense of a specific failure, they will immediately recognize the failure mechanism and will not take the steps of observing and reporting before reaching this conclusion. But we should not exaggerate this effect. For only two types of pictures (the sand boils and the sand driveway failure set) the effect was considerable; for the others less than 5% of the participants either mentioned a failure mechanism for what they saw (observing) or a signal for the diagnosis (diagnosing). Therefore, generally speaking, most participants were able to make a difference.

For determining the accuracy of diagnosing, the same line of reasoning was followed as with observing. This means responses were categorized into Very Accurate (VA), Accurate (A), Slightly Accurate (SA), and Inaccurate (IA). Similar to

observing, with diagnosing it is about using the appropriate term to signify what is happening. A response would be considered VA if it used the *exact* term from the game; A if it is closely similar to the text from the game or if they used a valid alternative; SA if they correctly described the mechanism instead of naming it; and IA if their description or term is incorrect or if they simply did not know it.

Except for one of the no failure pictures (Picture 5A), an improvement in accuracy can be observed. In terms of percentages the largest improvement can be noticed for the sand boils failure set. This is exactly one of the failure mechanism terms most people already knew about—now that they have given it a place, they were able to appropriately use it. About 80% of the participants did so. This is much more compared to for example the failure mechanism micro-instability (about 24%), a term and mechanism with which many people had trouble remembering let alone understanding. As a term, the same goes for macro-stability, but as a mechanism it is much easier to understand, and consistent with this, many more people applied this very accurately (about 58%).

All in all, results have dramatically improved for this learning objective. However, their knowledge on this aspect was lacking altogether and it remains to be seen whether this acquired knowledge is sustainable. During the post-interviews I discovered that of all learning objectives, this one seem to have been forgotten most easily (Level 10). Some were able to sort of describe what a failure mechanism is and—eventually—after naming the mechanisms, they were able to describe how they evolve. But still, it was remarkable that some could not even name one mechanism right away.

### ***Taking Measures Was Already Accurate***

Unlike many other professions, less emphasis is put on “action” with levee inspection. The action is, in fact, the observing and reporting and not so much really taking care of the problems. As some patrollers do have this responsibility, it makes for a more interesting game, and players will go through a complete sensemaking cycle, the decision was made to include this feature in the game. For the same reasons, it was also included in the study.

Taking measures is about deciding what needs to be done if a situation becomes critical—that is, a levee breach is about to occur. Although many more possibilities exist, in the game we focused on the seven types of possible measures:

1. Covering the inner slope with foil.
2. Covering the outer slope with foil.
3. Placing a sandberm.
4. Removing objects.
5. Creating a sandbag containment ring.
6. Placing sandbags at the toe.
7. Placing sandbags at the crest.



When authorized to take a measure, players have to click on one of the seven options. If the correct measure is taken, the measure will be visualized and they will be praised by the levee expert. If the wrong measure is taken, the expert will explain why and players can try again.

#### Non-failure couple



**Fig. 8.6** Pictures of the non-failure couple

- *Assessing*: No failure. On both pictures nothing harmful can be noticed. Sheep are actually useful. They eat the dead grass and thereby ensure that the levee is well maintained. And while parking a car randomly along the road is something a police officer may find problematic, it is not a useful signal for patrollers.
- *Observing*: Real 5A could be classified as a biological activity; Virtual 5B as a human activity.
- *Reporting*: For both if damage of the revetment can be observed.
- *Diagnosing*: For Real 5A erosion inner slope seems most likely; for Virtual 5B this is impossible to judge. Any failure mechanism term from the simulator was perceived as “correct.”
- *Taking*: Removal of the objects.

Similar to diagnosing, the implementation of taking measures has also not received much emphasis in the game. And as explained before (see “Showing the Pictures”), in the test not much emphasis was put on it as well, as participants only had to answer the question if they assessed the failure situation as critical. The consequence of this choice is that for most of the pictures ten or less participants answered this question, making it impossible to make appropriate use of statistical analyses.

Despite this disadvantage, still an attempt was made to see how participants performed. To find out, fewer categories were used because—except for the sandbag containment ring—the answers from the game are very descriptive themselves. It would not be expected that the participants will exactly rephrase this. In addition, from the initial analysis it appeared that the answers were not widely varied. This

meant that three categories were sufficient for this question to determine its accuracy. The categories are:

- *Very accurate = near perfect or perfect*: The description of the measures to be taken is exactly or almost similar to that of the game.
- *Accurate = correct*: While the description of measures may diverge from the game, they have to be considered correct to be part of this category.
- *Inaccurate = incorrect or too vague*: If descriptions are wrong or too vague they become part of this category. An example of a vague description is just saying “sandbags.” For a measure to be effective, it matters how sandbags are used. Therefore, this description does not provide enough information.

In general it appeared that the patrollers were already quite knowledgeable about measures. More than 50% of the participants were at least accurate from the onset of the training. While taking measures is not the core task of patrollers, it is often talked about—in practice but also in the media. It is further often part of “strong talk” by the patrollers themselves. When stumbling upon some patrollers it is very likely to hear something like this:

When you see a levee moving, we just have to throw a number of those big bags on it. Those bags are huge! That will definitely stop the levee from moving any further.

Such talk is rather interesting from the perspective that most of them will never have an important role in this process. Yet, it seems that they do think and fantasize much about it. A couple of explanations can be given. First, these people love taking care of things. Of all activities, taking measures exemplifies this the most. Second, it seems that people become patrollers because they find it exciting to think of having to walk in bad weather, in the middle of the night, looking for signals. And the utmost excitement can be found when a levee is about to fail—when measures are needed.

Still, from the results it can be seen that some improvement takes place. For those pictures with more than ten participants, this improvement can also be statistically confirmed. And it can be seen that having a good word or phrase helps in achieving this. The only real measure with a catchy term, in Dutch at least, has an impressive improvement. This concerns the “sandbag containment ring” (“opkisten” in Dutch). About 70% or even more of the participants started using this term compared to 20% before.

## Picturing the Overall Results

In the previous sections each learning objective is explained: what it is, how it is analyzed, and what the results are. Now it is time to turn to the overall results on accuracy and also discuss the results on the other criteria, which concern word count, dispersion, and vocabulary.

### ***Accuracy Improved with 26%***

The third goal of this level, to see how the pictures were made sense of before and after the training, was achieved by calculating the total accuracy scores and comparing the totals before and after the training (Table 8.2). The totals were calculated as follows:

1. For all totals the “taking measures” answers were neglected. Not everybody answered this question because it only had to be answered if the picture was considered critical by the participant. Still including these answers would bias the results.
2. Picture totals were calculated by summarizing the accuracy scores on assessing (1), observing (3), reporting (3), and diagnosing (3) for each picture. This means that the maximum score for each picture is ten. For observing, reporting, and diagnosing participants can get up to 3 points; for assessing this is up to one point.
3. Two totals were calculated for Set A and Set B: a *total core* and a *total*. The latter includes all pictures, whereas the total core excludes the non-failure couple and the failure couple. Both excluded couples were not trained by playing the game. They were used to check whether people were able to distinguish failures from non-failures and to see to what extent their learning transfers to completely new situations.

Except for the non-failure couple, performance improved for every couple. It turns out that people were able to discover the non-failures easily, with average scores reaching 88% before and after the training. The pitching failure couple shows the most improvements, with scores of 34% (Virtual 1A) and 37% (Real 1B); the new failure couple the least, with scores of 12% (Virtual 7A) and 9% (Real 7B). These new failure scores are about half of the other pictures. This is still an improvement and so we can conclude that some transfer occurred.

Improvement turned out to be more or less the same between the two sets, with 19% improvement for Set A and 20% for Set B. With the not trained couples out of the way, the improvement for both core sets come down to 26%. This indifference is rather interesting, because overall the performances on Set B are less than A, suggesting that Set B was more difficult. Yet, the improvements are similar. Improvement seems therefore independent of difficulty. Both groups went through the same training experience and this has led to the same increase in performance. This is an important observation, because this is further proof that the improvement was caused by the game and was not a result of the measuring instrument. It also shows that the effects of the game are consistent: if anybody will take this training, they can expect on average a performance increase of 26%.

These results are promising and show that the training was valuable. But if we take the performance scale and assume that a person performs sufficiently when reaching a score of 55% or higher, then we see that for only six pictures, including the non-failure couple, this cut-off is achieved. From this perspective it seems more improvement is possible and maybe even necessary.

**Table 8.2** The total results per picture (between-subjects)

Picture	Max. <sup>a</sup>	Pre		Post		Improvement, %
		<i>M(SD)</i>	%	<i>M(SD)</i>	%	
Virtual 1A	10	3.31(1.6)	33	6.66(2.4)	67	34***
Real 1B	10	1.75(1.4)	18	5.53(2.5)	55	37***
Real 2A	10	3.79(1.8)	38	6.19(1.9)	62	24***
Virtual 2B	10	3.10(2.0)	31	5.54(2.0)	55	24***
Virtual 3A	10	2.50(1.4)	25	5.04(2.4)	50	25***
Real 3B	10	2.71(1.7)	27	5.15(2.1)	51	24***
Real 4A	10	0.958(1.4)	10	3.88(3.1)	39	29***
Virtual 4B	10	0.859(1.2)	9	3.39(2.7)	34	25***
Real 5A	10	8.76(2.9)	88	8.80(2.4)	88	0
Virtual 5B	10	7.86(3.6)	79	8.40(3.0)	84	5
Real 6A	10	2.21(1.5)	22	3.88(2.2)	39	17***
Virtual 6B	10	2.60(1.5)	26	4.19(1.9)	42	16***
Virtual 7A	10	2.71(1.7)	27	3.89(2.5)	39	12*
Real 7B	10	0.973(1.2)	10	1.9(2.0)	19	9**
Total core <sup>b</sup> set A	50	12.7(5.3)	25	25.5(9.0)	51	26***
Total core set B	50	11.0(5.2)	22	23.8(8.3)	48	26***
Total <sup>c</sup> set A	70	24.2(6.5)	35	37.6(10)	54	19***
Total set B	70	19.6(7.3)	28	33.9(10)	48	20***

\*\*\*  $p < .001$ ; \*\*  $p < .01$ ; \*  $p < .05$  (two-sided).

<sup>a</sup> Max. refers to the maximum number of points participants could achieve per learning objective.

<sup>b</sup> The core set excludes the no and new failure pictures.

<sup>c</sup> The total set includes all pictures. If participants assessed the no failures pictures correctly, they received all possible points.

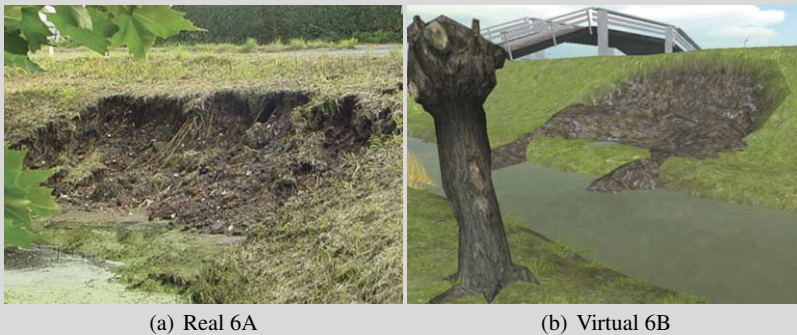
On the other hand, the participants started off bad, with accuracy scores of 25% (Core Set A) and 22% (Core Set B). The training helped them double their accuracy scores and so they made a great leap forward. With the right changes and additions to the training, an even greater leap forward may be achieved.

The findings of the third goal can be elaborated upon with those of the second. This time the accuracy scores for each learning objective over all pictures were summarized and compared (Table 8.3). No distinction was made between the two groups.

Based on these results, we see that diagnosing, with a score of 13%, contributes much to the overall low scores at the start of the training and that the small improvement on reporting—only 10%—may be one of the explaining factors why the results were not any better. If we look at the learning objectives, reporting is in fact the only one who is far below the cut-off of 55%. As explained earlier, reasons beyond the game may exist why this led to a rather disappointing result. But if we consider this as given, this is certainly the area that needs the most improvement.

The total results are similar to the ones of the third goal. Also here we see an overall increase of 26%. This should not be a surprise. For both sets this increase was achieved, so if we look at the performance within-subjects instead of between, a similar result should be expected.

## Watery slope failure couple

**Fig. 8.7** Pictures of the watery slope failure couple

- *Assessing*: Real 6A is severe, while Virtual 6B is critical. Although they more or less portray the same situation, the context is quite different. The first is part of a small levee along a ditch, with no danger of any flooding as no large water can be observed behind the levee. In contrary, with the second this danger is certainly present. The bridge signifies a large water must be behind this levee.
- *Observing*: Liquefaction—due to saturation of the soil, parts of the levee started to move. The end result looks like a settlement occurred and for this reason, this sort of classification was also considered accurate.
- *Reporting*: (1) The length and width of the failure; and (2) if soil is flushing and if so how much.
- *Diagnosing*: Micro-instability.
- *Taking measures*: Either foil with sandbags or a sand bank would help to stop the liquefaction process temporarily.

**Table 8.3** The total results per learning objective (within-subjects)

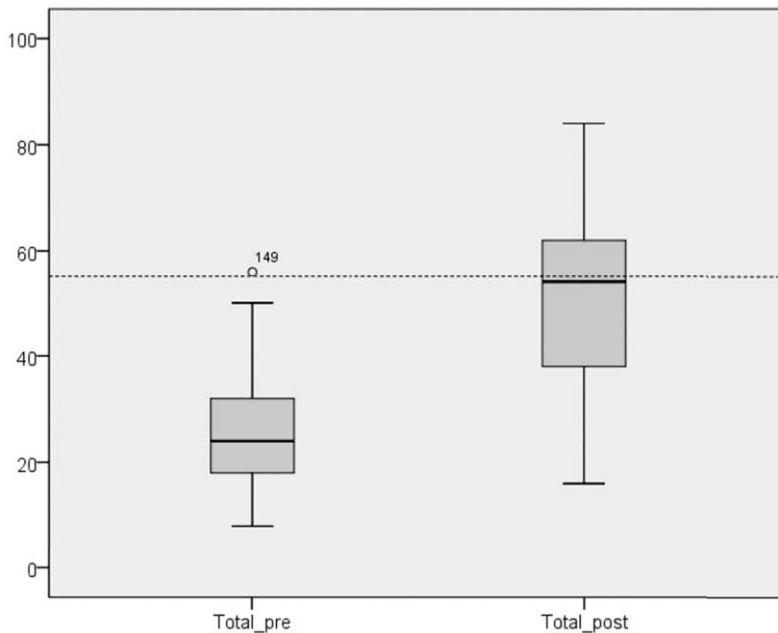
Picture	Max. <sup>a</sup>	Pre		Post		Improvement, %
		<i>M(SD)</i>	%	<i>M(SD)</i>	%	
Assessing	5	1.86(1.0)	37	2.80(1.2)	56	18***
Observing	15	4.64(1.0)	31	8.89(2.7)	59	29***
Reporting	15	3.64(2.0)	24	5.14(2.1)	34	10***
Diagnosing	15	1.97(2.7)	13	8.17(4.4)	54	41***
Total core <sup>b</sup>	50	12.1(5.2)	24	25.1(8.3)	50	26***
Total <sup>c</sup>	70	22.9(7.1)	33	37.5(9.7)	54	21***

\*\*\*  $p < .001$ ; \*\*  $p < .01$ ; \*  $p < .05$  (two-sided).

<sup>a</sup> Max. refers to the maximum number of points participants could achieve per learning objective.

<sup>b</sup> The total core excludes the non- and new failure pictures.

<sup>c</sup> The total includes all pictures. If participants assessed the non-failure pictures correctly, they received all possible points.



**Fig. 8.8** A boxplot of the total accuracy scores in percentages ( $N = 125$ ) of the core set of pictures before (left box) and after the training (right box). Participant #149 is an expert employee and the only one above the cut-off of 55%, which is represented by the dashed line

Figure 8.8 gives a complete overview of the within-subject accuracy results. This gives another look at the data. It shows that the scores are much more varied after than before. Everybody was more or less at the same level of knowledge before. Few people managed to get a sufficient score. The highest score—the outlier (Participant #149)—achieved a score of 60%. After the training, the results are much spread, but more than half of the participants got a sufficient score this time. This shows a more positive look on the scores than if we would look at the pictures separately. The reasons for this variation require further investigation and possible factors are illustrated in Level 11.

### ***Word Count Decreased with 25%***

The number of words participants used for each responses was counted and summed to see if any change was noticeable after playing the game. It was expected that the participants would use fewer words. As people acquire a particular vocabulary and know what to use for what type of situation, they will stop describing what they see and only apply the label. Others—they assume—will understand what they refer

too. For example, the term “pitching stone” says it all. No need exists to describe how the stones look like or what role they fulfill at a levee. As labels require less words than a description, a drop in the number of words could have been expected.

**Table 8.4** The total word count per learning objective (within-subjects)

Learning objective	Pre, M(SD)	Post, M(SD)	Decrease, %
Observing	43.9(32)	31.1(21)	29***
Reporting	50.1(27)	43.5(27)	13**
Diagnosing	18.4(16)	9.83(5.6)	46***
Total core <sup>a</sup>	113(59)	84.6(46)	25***
Total <sup>b</sup>	131(69)	97.6(53)	26***

\*\*\*  $p < .001$ ; \*\*  $p < .01$ ; \*  $p < .05$  (two-sided).

<sup>a</sup> The total core excludes the no and new failure pictures.

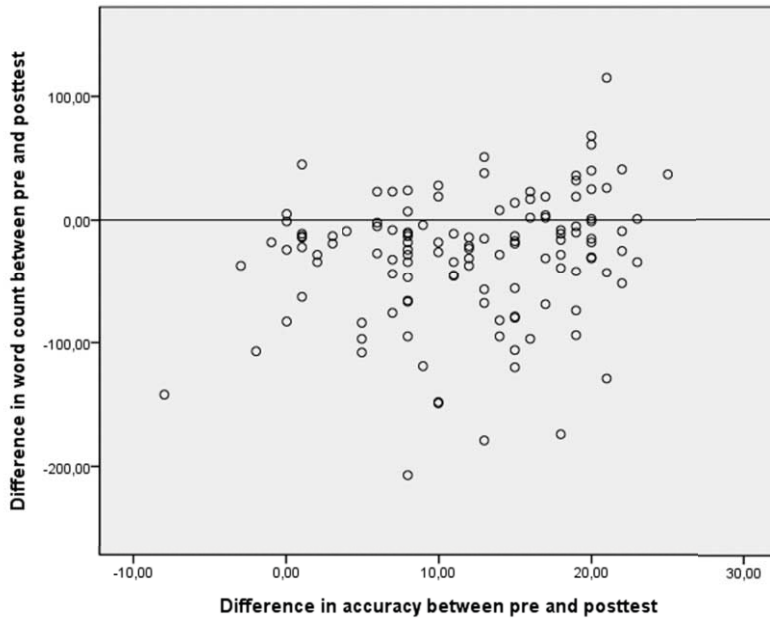
<sup>b</sup> The total includes all pictures. If participants assessed the no failures pictures correctly, they received all possible points.

Generally, this drop in the number of words was achieved. When one looks at either the learning objectives individually, by summing all the words for every response on every picture before and after, a decrease of on average 13 to 46% can be seen. The highest drop is similar to the results on accuracy related to diagnosing; the lowest drop is equally similar to the results on accuracy related to reporting. Moreover, overall the decrease is—with a percentage of 25%—more or less completely similar to the increase in accuracy.

These results raise the question whether a decrease in words and an increase in accuracy go hand in hand. Then one might reach the conclusion that less is more. Contrary to these expectation, on the pre- as well as the post-test it turned out that it was the other way around. The more accurate participants used more—not fewer—words,  $r_{\text{pre}} = .34, p < .001$ ;  $r_{\text{post}} = .41, p < .001$ .

If we look at the differences between the two tests, by subtracting the pre-test results from the post-test results, it turns out that a small yet still noticeable relationship exists between being more accurate and using more words afterward,  $r_{\text{dif}} = .20, p = .024$ . A closer look at the data reveals that about 25% of the participants used more words afterward and these tend to have high scores (Fig. 8.9).

How is this possible? One could theorize that greater knowledge results in more thorough expression. In addition, by looking at the learning objectives individually, it was noticeable that for reporting in general a medium to strong relationship could be found between using more words and accuracy. This is understandable. For reporting, it was often necessary to mention more than one correct item—those that used more words most likely included more correct items. For the other two considered learning objectives, observing and diagnosing, small to negligible but still positive relationships were found, still confirming that it actually seems more is more.



**Fig. 8.9** A scatterplot of the differences in word count and accuracy. Every observation above the horizontal line used more words afterward

This explains why certain people may have used more words; however, it does not explain the general decrease. A reasonable explanation for the latter is that people were using their words more sparsely and effectively. They wrote down what they knew was correct afterward, while beforehand they may have been simply guessing—something I widely observed during the interviews as well. Those people who became above-average knowledgeable about the subject, they just had a lot more to say than beforehand and applied their acquired knowledge.

Therefore, one can roughly divide participants in two groups: a large group with those that achieved less with more and a smaller group with those who achieved more with even more. Of course, individuals who seem to fall in between the two polls are found too: utterly verbose participants with little to no accuracy and very precise participants who in terms of accuracy performed similarly to the verbose high performers.

### ***Dispersion Decreased Variously***

A clear shared understanding comes with a shared vocabulary and so another criterion for judging the effects of the game concerns *dispersion*. This dispersion refers



to the variability of the responses who have been coded (almost) literally. The more variety, the more dispersed the responses are. With the game setting standards, it was expected that afterward the responses would be less dispersed.

**Table 8.5** Simple example of how the dispersion count and percentage work

Code	Example 1		Example 2	
	Count	Percentage	Count	Percentage
Code A	2	6.3	5	39.1
Code B	2	6.3	1	1.6
Code C	2	6.3	1	1.6
Code D	2	6.3	1	1.6
Dispersion count	4		4	
Dispersion percentage	25.0		43.8	

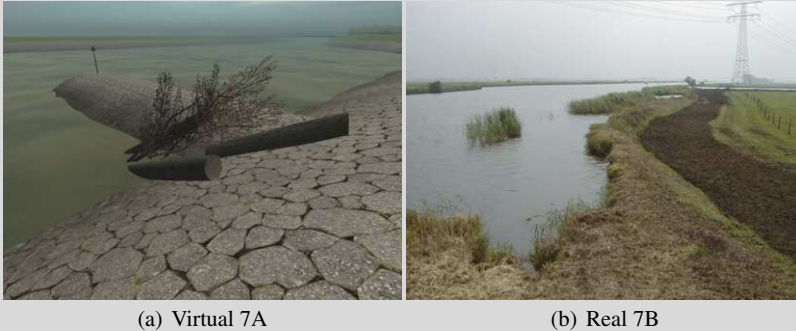
To see whether this is the case, two indicators were used. One is similar to word count and is for that reason called *dispersion count*. This indicator simply counts the number of codes used by all participants. The problem with this indicator, however, is that it does not give a proper representation of the actual dispersion. It does not take into account how many times a certain code is used. If before and after the game four codes are observed, but before each code is applied twice and afterward one code is applied five times and the others once, it can be argued that the dispersion is much less in the latter situation (Table 8.5). A greater amount of people have said something similar.

To consider the distribution of the codes among the participants another indicator was used, the *dispersion percentage*. To give more weight to codes applied more frequently, all code counts are squared. Then, all squared code counts are summed and divided by the square of the total number of codes applied. This multiplied by 100 gives the dispersion percentage. Mathematically, the calculation looks as follows:

Dispersion percentage =  $\left( \frac{\sum X^2}{N^2} \right) \times 100$ , where X = number of times a code is used and N = total number of times all codes are used.

The dispersion count results show an average for observing and diagnosing of 30 to 40% less variation. For only the type of items mentioned, a 24% decrease is seen. If we look at the configuration of the items, as in seeing what exact combination of items people mention, the decrease is 13%. Interestingly, for some pictures the number of configurations increased rather than decreased. A possible explanation is that participants may now have a richer understanding of the possibilities and various participants choose different configurations as a result. To elaborate, beforehand participants may have only said “damage” and afterward they either say “overtopping” or “settlement.” With this the report becomes more specific, yet also more dispersed.

### New failure couple



**Fig. 8.10** Pictures of the new failure couple

- *Assessing*: Virtual 7A is reportable. No damage can be observed, but this may quickly change if a strong current or wave hits the levees. About Real 7B no consensus was achieved, yet it was clear that this was at least severe if not even critical. Both assessments were considered accurate.
- *Observing*: For Virtual 7A floating waste; for Real 7B this concerns biological activity. The latter is hard to recognize, but one could recognize that this type of damage to the levee is typically caused by rats.
- *Reporting*: For Virtual 7A (1) the type of floating waste; (2) the (possible) damage that can be observed; and (3) an indication that it should be removed. For Real 7B (1) the damage that can be observed and (2) the length and width of this damage should be reported.
- *Diagnosing*: Erosion outer slope.
- *Taking measures*: For Virtual 7A the objects need to be removed; for Real 7B foil with sandbags on the outer slope will do the job.

In terms of the dispersion percentage, the smaller effect on reporting becomes even more evident: it only became 4% less dispersed. However, the dispersion percentage further makes the difference between observing and diagnosing clearer. Diagnosing has decreased with 28% almost twice as much as observing, which only decreased by 15%. The results on dispersion are thus consistent with the results on the other criteria so far.

### *Vocabulary Tripled*

For considering a change in vocabulary I first looked into the percentage of frequent words, such as the English equivalents of “the, is, there, from, that, a(n), and, of, on, in.” It became clear that these remained consistent between pictures and over all

pictures before and after the training: before as well as after the training on average 29% of the written responses were frequently used words. The word “the” (in Dutch, *de*) was the most used of all, with “of” (in Dutch, *van*) the second most used. The use of frequent words can therefore be considered constant.

Subsequently, I considered the *Levee Patroller* vocabulary, which are keywords that are mentioned in the the game. This vocabulary use remained, much similar to the frequent words, consistent as well between pictures. This time only a big difference can be observed with the vocabulary used before and after the training. A total of 176 words were considered and it appeared that before the training on average 9% ( $SD = 1.9$ ) of the words came from this set. After the training, this increased to 26% ( $SD = 3.1$ ). Thus, the vocabulary almost tripled.

Taking the 29% of frequent words into account, we see that the amount of words based on the game becomes very dense after the training. The most mentioned words are erosion and settlement, followed by inner and outer slope, sand boils, soil, and water. For some pictures some of these words were mentioned more often than one of the frequent words.

## Lessons Learned

This level explained and presented the results of the sensemaking test, a test specifically developed for the game-based training with *Levee Patroller*. In essence, the test asks participants to answer a number of open questions regarding several pictures—some real, some virtual.

The resulting data from the test is very rich and enables researchers to look at the data from various perspectives. For the purposes of evaluating the game, the data was especially used to determine the sensemaking performance by considering the accuracy by which participants made sense of pictures. In general it becomes clear that on average performance by participants and on pictures increased with about 26%.

Of course, differences between learning objectives and pictures are to be noticed. The watery slope failure, for example, which was the least performed on failure in the game, was also the least performed on failure in the test. In contrary, and much similarly to the game as well, the stone damage concerned the best performed on failure.

And here too we see not much improvement in terms of reporting and assessing and a major one regarding diagnosing, suggesting that what happened in the game was important and/or that certain learning objectives are more difficult to improve upon. However, playing the game definitely improved participants' accuracy, and therefore the game seemed to have an impact on sensemaking performance. From the analysis it becomes further clear that the game has the clearest impact if the content is connected in a meaningful way to the game mechanics.

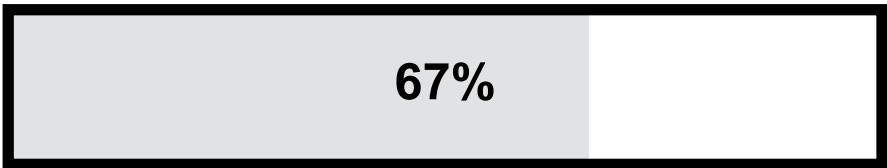
The game also had an impact on communication, as we can tell from the decreased word count, decreased dispersion, and increased vocabulary use.

## Level 9

### It Has an Exit Button, Right?

*We are busy with our second youth rather than our first. This is a proof of that, because we stayed!—Participant at DB4*

*DA1-#1: We do not play enough games! DA1-#2: Then we should start doing that! Both: Ha ha ha ha [laughing out loud]—Two participants at DA1*



67%

The training/evaluation with *Levee Patroller* involved exploring both individual opinions as well as group consensus. To explore the group consensus, I organized a *discussion* at each end-meeting. In total, I trained 11 groups and so I had 11 group discussions: four at Organization A, five at Organization B, and two at Organization C. These discussions are the focus of this level.

A wealth of literature and advice exists on having group discussions (e.g., Galanes & Adams, 2011) and on making them more “effective” by means of the use of computers (e.g., Fjermestad & Hiltz, 1998; Pinsonneault & Kraemer, 1990). I decided to keep it simple. By then the participants had been sitting long enough behind computers and so on purpose I made sure that all computers were removed before the discussion even began.

Another reason was rather practical. That way we were able to leave the location somewhat in time, ensuring the security personnel could go home too (they were waiting on us) and—not less importantly—we would be home and asleep on a somewhat decent time (usually about 1:00 a.m. with speeding ticket; about 1:30 a.m. without one).

The goals of this level are to describe

- How the discussion was further organized and also analyzed;
- The group consensus about the current game-based training; and
- The group consensus regarding the future of the game-based training.

## Setting Up a Discussion

To encourage discussion, I made use of a common technique used in focus groups: presenting statements (Kitzinger, 1995). On the topics that I thought were relevant I created a short statement and presented these one by one by means of a laptop and projector on a large screen.

I had to be economical with the number of statements. The discussion was for a logical reason at the complete end of the training and much time of the end-meeting was taken by playing the end-exercise and filling out the post-questionnaire, leaving little time for discussion. Up front I reserved about 30 minutes for the discussion and depending on the progress of these first two activities and the willingness of the participants to continue the discussion, I shortened or extended this time.

On average the meetings took 35 minutes ( $SD = 6:40$  min). The meetings at Organization A took a little less than the initial 30 minutes planned. This was due to the setup of having two sessions on a single evening. The meetings at Organization B were the longest, but unlike Organization C, the meetings were held in close proximity to their homes and so they would be home quickly. In contrast, the participants at Organization C had to drive at least 30 minutes before they got home and, therefore, the willingness to continue the discussion was somewhat less.

As for the statements, due to time I decided to restrict myself to five discussion items at Organization B and C and to three at Organization A. This would leave six minutes per item, which to me seemed the bare minimum to have a proper discussion.

I led all 11 discussions but tried to stay on the background and only intervened to move onto the next discussion item or to suggest whose turn it was to speak. When necessary I asked a question to elaborate on a point made by a participant or asked a general question to foster the discussion. I prepared some questions up front, such as asking how the participants experienced playing the game at home.

To see if a group consensus was reached, I occasionally summarized the points made and then asked if the group agreed with this summary. To make sure participants would not necessarily agree with a statement, such as the first discussion item, I stressed that they did not have to agree with it or simply asked participants to raise their voice if they did not agree.

Therefore, I acted more or less as a conventional facilitator. The reason I point this out is to emphasize that I tried to stay outside the discussion as much as possible. I did not want to persuade the participants or defend the training or game. I wanted to know what they thought of it.

Because I expected that participants would fire away with comments about the design, I explicitly told them that the design would be a discussion item later on. The idea was to focus on the discussion item presented at the screen. Despite this very explicit comment, I had to repeat this several times before participants would finally save their frustrations for later.

### **In-depth explanation: recording the discussions**

The discussions were recorded in two ways. The first way was a short transcript made by my co-facilitator. He or she wrote down the main arguments on each discussion item, enabling me to quickly grasp what the discussion was about and to have backup in case for some reason something went wrong with the second way of recording the discussion.

The second way made use of an audio interface consisting of four microphones spread over the meeting room, which were connected to a portable audio recording equipment. The recorder was on its turn connected to a laptop with the audio recording software *REAPER* installed on it. During each discussion four audio files were created, one from each microphone. With the recording software I could easily switch from one line to another if I could not hear a participant very well. Although it was somewhat of a hassle to bring all this equipment with me, this way I was ensured I would be able to record discussions in almost any meeting room.

Using the recordings, I made a complete transcript of each discussion. Because I noted the positions of the microphones and the positions of the participants in the meeting rooms, this made it further possible for me to know who said what. However, recognition was far easier for Organization A and C compared to Organization B and for a logical reason: I spoke to a relatively larger number of people one-on-one at Organizations A and C.

Because I was not completely sure about who said what, the quotes presented in this level will not include the participant number. I used the names for my own interpretation only. If I was confident about who said something and this is important for interpreting the quote, I added this information before or after the quote (such as with the rebel from Organization C). The codes include the organization name (A vs. B vs. C) and the group number (1 vs. 2 vs. 3 etc.) and have as prefix the capital D, which stands for "Discussion." A discussion quote from the first group from Organization B will thus be represented as "DB1-#1." I included the number behind the dash as to number the participants during a single discussion part.

This level is written similar to how I have led the discussion. Mostly, the participants do the talking. I have only organized their words according to the discussion items and the discussion groups. In between I give the context and meaning of their words.

Overall, the discussions were pleasant, full of jokes, and also passionate. The participants really cared about the conclusions of this pilot training and except for a few almost every one participated in these discussions. I tried to portray this atmosphere in describing the discussions.

Although the essence of the discussions will be reflected in what I have written here in this level, it is a shame that much of it will go lost in translation. First, I had to translate the discussions from Dutch to English. I tried to stay true to the original wording of the participants, but without any doubt the original connotation of the wording will be slightly altered.

Second, and much more unfortunate, is that I am unable to accurately translate the sort of Dutch language the patrollers use. They speak in different dialects and in a type of slang that is highly specific to particular parts in the country and Dutch

society. The reason I find this unfortunate is that the language used expresses a great deal about their culture and background and their line of thinking. The translations now read as if all of them spoke proper English. In reading my analysis, one should keep in mind people who value “deeds, not words,” and who like to be outside and work with their hands. This mental picture will help the reader appreciate the comments made.

## **The Impact of *Levee Patroller***

The first three out of five discussion items were about the current game-based training and its impact on the participants. The discussion items were:

1. *Levee Patroller* improves the inspection of levees (effectiveness).
2. *Levee Patroller* is a tool for the future generation of patrollers (target group).
3. I look differently at my environment now (awareness).

The second and third discussion items were not asked at Organization A and were constructed based on my experience of the first sessions. The second item was based on what I heard two dropouts say after the start-meeting. They were of the opinion that this was something for their children and not for them (Level 4).

The third item is based on a variety of responses, during the discussion and outside of it, in which I noticed that participants became much more aware of their environment. In other words, playing the game influenced their *situation awareness* (Endsley & Garland, 2000; Klein et al., 2006a). This is something I will elaborate on. For now it is important to note that because of this insight I decided to give it some more emphasis by including it as a discussion item.

## ***Playing Improves the Inspection***

I started the discussion with the most relevant issue: Does *Levee Patroller* improve the inspection of levees? Overall, participants seemed to agree with the statement. Yet, if we delve into the deeper meaning behind why people think it improves the inspection, we find a multitude of answers.

### **Diverging opinions**

In only one situation I had a disagreement (DC1). This occurred at the first group discussion at Organization C, which involved the so-called rebel referenced before (Level 4). I hardly started the discussion by asking who agreed or did not with the proposed statement and the following discussion was a result:

*Rebel:* Totally not.

*DC1-#1:* Why not?

*Rebel:* Because it does not make any sense.

*DC1-#1:* Why does not it make any sense?

*Rebel:* First, I have no feeling with that thing. Second, the images are such that, well, if I go on patrol, I do not walk with such a mouse button [sic] in my hand to look if it is a crack.

*DC1-#1:* OK, but is it because of inexperience with the computer or...

*Rebel:* It is both.

Here the rebel tacitly acknowledges that his own incompetence in using the computer drives his opinion about the game. He already acknowledged this to me over the phone, yet he kept on persisting that there is more to it. In his eyes the game is useless because the computer is not used in practice. I tried to hear the opinion of others, but quickly enough the rebel tries to take over the conversation again. This time his co-rebel joins the fight. Like the rebel he cannot see how the game is useful to the practice of levee inspection and thinks that he is already knowledgeable enough to inspect levees.

*Casper:* Does anybody else have another opinion?

*DC1-#2:* I have never had that [levee inspection] course and never ever went on patrol and now I do see what happens if you take the wrong measures or if you assess the severity incorrectly. I most certainly learned much from it and, in addition, how many times does it happen that people find a real failure mechanism? Now something exists that they can recognize.

*DC1-#3:* I think that it is also about the way of reporting. Not specifically to diagnose, because it is above all about I see a crack or a seepage and how do I deal with that? If you do not report this correctly to the Action Center, you will get the question back. That is how it works in practice too. For this reason, I think it is good for levee patrollers.

*Co-Rebel:* That works faster in practice. Behind such a thing it is a never-ending process.

*DC1-#3:* But it is about the procedure that needs to be followed: the location, the report, what do I see. I think that this is the most important contribution...

*Co-rebel:* I do not have any trouble with that!

*DC1-#3:* [Ignores co-rebel and continues]...but diagnosing and taking measures? That is not up to the patrollers. It is part of it [the game] and that is fun but you cannot game this.

Participant DC1-#2 is a young employee at the water authority and his words contrast to that of the older, very experienced employee DC1-#3. He is one of the heads of the levee inspection organization and preoccupied with it on a daily basis. He started to convey his opinion about the game cautiously at first ("I think that"), but along the conversation he stated it more firmly. In his eyes the game was above all useful for learning how to report. I have heard the same opinion by many other levee inspection experts during the design of the game.

In addition, because his patrollers do not need to diagnose and take measures, he is of the opinion they do not need to learn this. The game setup does enable to turn this off, because we have heard this desirability before and included it to give users an option. I only had good reasons to include it (Level 3) and one of the other volunteers (DC1-#4) seemed to agree with me.

*DC1-#4:* Well, it might be useful to know the story behind my report. I mean we see a crack and I am just a volunteer so then I call the Action Center and say something is happening.



They tell me fine, it is noted and then something happens. But now with this simulation we are able to see what really happens.

*DC1-#5*: You have to see the idea behind it. But reporting well, that is important.

*DC1-#3*: For levee patrollers that is most important.

*Rebel*: But reporting in an exercise or in practice, you do not do that with a computer do you?

*Everyone*: [say somewhat agitated and in a loud voice] NO!!!

*DC1-#3*: It is an aid to learn it.

*DC1-#6*: It is a learning tool, that is how you need to look at it.

*Rebel*: If you need to learn it, you have to do that in practice.

*DC1-#3*: No, because the procedure is the same. You need to have a location, you measure a crack...

The atmosphere started to worsen, because the rebel kept his foot down about his ideas about the usefulness. At this point I had the feeling I needed to intervene, because the rebel could not see the game's purpose or did not want to see it. I asked if anybody else had another opinion and interestingly enough, another commander (*DC1-#8*) pointed out that the game might be more than learning about procedures.

*Casper*: OK, I think it is clear now. Does anybody else have another opinion on what you possibly learn from playing?

*DC1-#7*: Some theoretical knowledge.

*DC1-#8*: Well I think that those who have never patrolled and have been on a levee are able to recognize failures now.

*DC1-#3*: You have to see this as a learning aid. You cannot compare it with the actual practice. [...] it is a training to learn "how does this work?" Practice is different. That is true. Procedures remain the same.

*DC1-#1*: It is an extra tool that you use besides the exercise to get confronted with your profession throughout the year.

*DC1-#4*: Most certainly for volunteers. We do not do this every day.

*DC1-#9*: Also for new people I think. You can call and say "Hey, I see a crack" but then you get the question back of "What is its length" and you think "Oh yeah, I should have thought of that before."

This discussion shows diverging opinions on what function the game fulfills as well. We see a divide between the senior, more expert participants and the new employees and volunteers. The first sees its use in particular in learning the procedures of levee inspection, whereas the others value recognizing, assessing, diagnosing, and taking measures too. They want to get the complete package in order to get the complete picture.

As for the rebel and the co-rebel, their word choice already says much. They continuously spoke of "that thing" instead of game, simulation, or computer. This language use was not reserved to only these two participants; in the first discussion at Organization A one of the participants also spoke of "that thing."

I cannot work on such a thing. But practice I do know. I know that very well. If I am in the polder and have a radiotelephone I will do much and much better.

The rebel and co-rebel further had trouble seeing how the game could benefit the inspection and for two reasons. First, they take the game very literal, in its use for the actual inspection ("I do not have that thing with me") and how it portrays the actual inspection ("That works faster in practice").

Second, they do not see the problem! Both said not to have any issues with reporting failures, which indicates that both did not have a real desire to learn something. They saw themselves as experts already. The participant with the long beard (#144) who even did not try to play the game at home belongs to the same category. He even told me he knew everything.

That does not mean these people are not valuable to the organization and maybe they are right. Maybe we are exaggerating the need to professionalize levee inspection and should we just depend on the local expertise of the people who work and live around the levees. However, what is clear is that game-based training is not meant for everyone.

### **It remains a game**

In the other discussions, I did not find anyone who publicly voiced his dislike against the game. However, in discussing how the game improves the inspection some immediately highlighted the reservations they have. In their eyes the game is beneficial to levee inspection, but it remains different from reality in several ways.

*DA1-#1:* You do get insight into the possible failures you need to pay attention to, but the real situation with vegetation is still something different than in this simulation...It remains a game.

*DA1-#2:* I believe in practice the image is much wider than that this simulation can give you. If I just walk in the polder on a levee, I can see multiple sides at the same time. Here you have to go every time to it [a failure]. That is what he [the Action Center] tells you to do and behind your back the levee breaches! Then your performance is insufficient, it is as simple as that! But I could not see and report it at that time.

*DA1-#3:* Indeed, if I walk on top of a levee [in the game], you cannot look to the left and the right, because the image is not wide enough.

*DA1-#4:* And you see something here, but it happens there.

In another discussion, similar comments were made about how the environment is different from that of practice.

*DC2-#1:* In practice the levees are never all mowed. In practice if it rains or the wind blows you have different circumstances than now. Now you sit dry and warm behind a screen playing a game. You see a virtual image. That is a lot different than when you walk outside and there is dusk or it is dark and the wind blows.

*DC2-#2:* The advantage of practice is that you know your levee segment very well. You know what the weak spots are.

*DC2-#1:* The seepages always return at the same spots.

*DC2-#2:* That is a big difference with the game. With the game you are continuously searching for something. In practice that is a bit different.

*DC2-#1:* Take those pitching stones. Berry bushes and other mess grow over them, so you do not get very close. In fact, the whole bottom part of the levee you will not walk in practice. In the game you can do this. For volunteers who never work outside it would be good to go out with bad weather. That is a different experience than sitting in your room.

Participants DA1-#1 and DC1-#1 are right. The failures are not as detailed and the environment is not as well. That is a trade-off we made and one that was some-

what enforced too by the limitations of the used technology. What is more interesting is what the others were saying. They first accuse the game of restricting their view, making them unable to see failures they would otherwise have seen.

Second, they assume the actual world is less uncertain than the gameworld. This contrasts with what the coordinators have told me. They spoke of highly uncertain situations. In reality it may very well happen that you see something here and it happens elsewhere.

That same evening the second group discussion also talked about the restrictions of the game. Here somebody intervened and indicated that the game is valuable. Interestingly enough, the one intervening concerned the dad (DA2-#2) of the participant making the comment about the restrictions (DA2-#1).

*DA2-#1:* I think it says more with actual pictures of the environment. If you have a crack [in the game], then I just look at it and know that is one, but if you just walk with long grass [in practice] and it is over it [over the crack], then you will not see anything...When I see a crack on my computer, then I just do click click click. At least, I do click click click [looking at all the older people in the room].

*Casper:* Are you saying the game is too easy compared to practice?

*DA2-#1:* I think so. I do not really do this for so long.

*DA2-#2:* But a crack is not as bad. Only if water comes into play. With a bit of wetness you should be more alert. You see this quickly and you feel it when you walk over it. So in this regard I think that it is [the game] helpful and that it is useful for practice.

*DA2-#3:* ...Now you also know the consequences of one or the other, of different things...it is now more clearly visible with the simulation. You can recognize certain things and know their consequences and what you need to do with them. What we received so far was only what they told us and this was of course little. With this system at your disposal you can refresh at times. You can get some theory.

*DA2-#2:* I think it most certainly helps. At least it gives you some idea. I have been almost all my live with the levee inspection and only once I sat at a post, when it was storming, and then I did not even dare to go out! Other than that I have never even walked around.

The young son (DA2-#1) makes a somewhat similar comment to that of participant DA1-#1, arguing that the failure images are far too simplistic compared to pictures of real ones. In addition to the trade-offs I mentioned earlier, this simplicity is also the power of those failure images, which is illustrated by the comments of the dad (DA2-#2). In his response we see that he learned something: if a crack is to be seen together with water, we have to be extra careful.

Both the dad and Participant DA2-#3 comments reflect what the game's overall purpose is. It is to give "you some idea." Beforehand hardly anybody had much experience with levee failures and as DA2-#3 said it becomes "now more clearly visible," because they can "recognize certain things and know their consequences and what you need to do with them."

The game has never pretended to model the actual world in every detail and as designers we have made decisions to actually distance ourselves from it. Some participants recognized this and could look beyond the restrictions and see its purpose and contribution; others had more difficulty with this. And of course practice is much different but that does not mean the game is not any useful. This is nicely illustrated by this comment from another participant (DB5):

...if you are in a real situation and you have to search for this one crack in the dark with a flash light on a levee slope with high grass, then this situation is far away from what what you have learned in the simulation. But it is good to become aware of what you could possibly find.

The reservations were not only restricted to the visuals of the virtual environment. In one of the discussions doubts were raised about the fit of the game procedure with that of practice. This was a major issue of concern during the development of the game too (Level 2). Each water authority wanted to have their practice realized within the game.

*DB2-#1:* It is a precondition that the way you report and inspect in the game are similar to reality.

*DB2-#2:* I agree. The game needs to be adequately adjusted to the procedures of every water authority. As it is right now, I have my doubts.

Participant DB2-#1 is the coordinator at Organization B. He emphasizes that the game's procedures should mirror Organization B's reality as closely as possible, something Participant DB2-#2 agrees with. After that he adds the following:

*DB2-#1:* We are discussing with the other water authorities how to train and educate the levee inspection organizations in a consistent manner and this means that you have to have consistent procedures and agreements.

What the coordinator added here is that the procedures are not uniform among the different organizations and that they are trying to standardize them. Although one needs to be careful in simulating different procedures, simulating procedures that are about to change is not useful either. In addition, the game could be seen as a vehicle of standardization. The procedures in the game are based on one of the first standardization attempts (Level 2) and now participants from different organizations have been trained according to the exact same inspection procedures and agreements.

Because of the reservations with the game as a training tool, during the same discussion, the suggestion was made almost right away to see the game as a support to the education of patrollers and not as the main instrument.

*DB2-#5:* If you see it didactically as a basic instrument for educating levee patrollers I would say no. You have to tell a group that with some sense and a good story with examples. For support I would say yes.

*Casper:* Are you saying people first need to get more theory before they can start playing?

*DB2-#5:* First they need to get some training and then this as support and added value.

*DB2-#6:* If you send people now outside you get a strange image, because if you see a piece of a crack in an asphalt levee—and you see this in practice often—then nothing is wrong at all.

*Casper:* So putting someone behind the game without any previous training is not a good idea?

*DB2-#7:* I think it is: to let that person learn what could possibly happen.

*DB2-#5:* No. Only related to an education it is useful.

I have heard this in other discussions too, but it is not a widely shared opinion. One reason for this diversity of opinions is that opinions seem to be dependent on

the background of the participant. This is what one of the participants seems to recognize:

*DC2-#1:* I notice a difference between you and me. You come from the region and you know the levees. You also know the locations and everything else. I sit here in an office and I hardly get outside. For this reason I would want to participate with this, because for me it works very well. I can really immerse myself that I am walking on a levee and that I see failures...For people with outside experience it might work better with walking outside and pictures, because the emphasis is on reporting and communicating failures, but for us who sit inside it is more about recognition.

*DC2-#2:* I also think it is most important for us to understand how we have to communicate. We say it is a seepage. Then what is important is whether soil flushes and other things like that. We need to put that on paper, so it is useful to others.

What is remarkable is that one of the heads of the levee inspection at Organization C argued intensively that the game is most useful for learning about procedures, something *DC2-#2*, another person with much outside experience, agrees with as well. Participant *DC2-#2* further said “we say it is a seepage” with so much confidence, it is as if there is no doubt about his observation. He skips “seeing,” he can immediately “say” it. No further processing of the information is required, because he made already sense a long time ago. So hereby he implicitly confirms that for him observing is not an important learning objective. It is about reporting.

For Participant *DC2-#1* on the other hand the game’s power is its visualization. It allows him to get a sense and feeling of what it is like to walk over a levee and get an idea of what to look for. For him observing is an important learning objective.

Those who are closely involved or knowledgeable about aspects of the game find the game to be restrictive. For others these restrictions were not noticeable or problematic.

## **Failures are more inside your head**

I will now turn to the positive remarks. I coded their positive remarks into four categories, which each represent one or more learning objectives: observing, reporting, vocabulary, and other. Let us start with the first. Observing seems a more important learning objective for those who have hardly walked over a levee. But even for those who did walk over a levee a number of times the game seems to be valuable.

*DB3-#1:* In October we have that exercise I think and then you walk over a levee. I think that is a good idea, because you have to know your own levee segment, to know how it looks like. I walked twice and I have never found anything and then you do not see what really happens. And I find this [the game] very realistically constructed and you learn much from it. You see different failures and that makes it easier to apply it in practice. I have also done the course last year and then you get from 8 p.m. till 10:30 p.m. a bucket load of information. After that you go home and look once more at your book—if you even make this—and that is it.

*Casper:* Do other people share this?

*DB3-#2:* Without any doubt.

*DB3-#3:* Absolutely.

*DB3-#4:* I have done the course two years ago but that does not stick very much.

*DB3—#5:* I think you cannot do it in any other way. It is hard to show it in reality.

*DB3—#6:* [In a sarcastic voice] Well, it is possible...

*Everyone:* Ha ha ha ha [laughing out loud].

Of course, this is exactly one of the reasons why the game was developed. Failures occur rarely and by simulating them inside a virtual environment we are able to let people get experience with failures. An alternative is that when walking over a levee participants are confronted with signs with a failure picture on them, but compared to this alternative the participants seem to find the game more useful.

*DB5—#1:* You see it really in front of you, at least some of the failures. You can walk over a levee in such an exercise, but then you get a sign with a picture. Now it says more.

*Casper:* You mean it is more visual?

*DB5—#1:* Yes, it is more visual because you see what happens.

*DB3—#2:* True, the beauty of it is that you see how it develops over time. If you get a picture then you get that picture and that is it. Now you get something and after 15 minutes you think Let us look at it again to see what is going on. That is what is so fun about the simulation. You see the consequences if you do not do anything.

From these comments we get that although the pictures are virtual, they speak more to the imagination of the patrollers than a plain picture. Failures speak more to the imagination because they are dynamic and because participants look at them in a situated activity. They sit behind a screen but this still gives the idea that they are there together with the failure and look at it with their own eyes. They are also able to look at the failure from various perspectives, which is not possible with a plain picture. This makes a virtual failure “say more” than a plain picture according to some of the participants. Not everyone agrees on this. The young son in Discussion DA2 suggested making use of actual pictures instead.

The game not only shows players what to look for, but also how to look if one finds something. A plain picture would arguably stimulate this less. You look at it, but you do not actively go somewhere. In another discussion (DB2) one participant commented about this.

If you see a crack at a certain moment then you also have to look at the bottom of the levee, because what happens at the inner slope?

Although the failure development over time and how signals relate to each other are important ingredients that make the game a valuable learning instrument, the major idea behind the observing learning objective is that participants get to know rare failures. This person at the fourth discussion of Organization A (DA4) describes this essence:

I think that many of the failures you have never seen are now much more inside your head. You have achieved more insights into them. Definitely.

## **It is all about reporting**

The second frequently mentioned reason why the game is useful concerns reporting. For some it was the reason.

It is above all about reporting. When you do not make good or complete reports with this game, you will be noticed about that...This is a handy tool for judging and reporting failures.

This comment by one of the participants (DB4) is one I heard many times. Participants said that it was especially useful that the Action Center made them aware about what was required. Quite often I heard something like “I immediately called the Action Center to tell them something is going on but then what is going on?” They had a natural response to not first make sense of the situation, but simply call as soon as they saw something. Although that is their duty, it is hard for the person on the other end of the line to understand what the patroller is seeing if they did not have a careful look at the situation and are able to communicate this. Consider the following discussion:

*DC2-#1*: It works for how it gives feedback, because it has a certain structure in giving feedback. If the water authority knows that they would want to do it like this, then a game like this is helpful in explaining what you have to do if you find something.

*Casper*: So it is about the procedure, about the steps you have to take?

*DC2-#1*: Well, say you find a crack then you look at its length, width, and everything else. You first collect that and then you call. In this area it is educational.

*DC2-#2*: At a certain moment I did have something like do you have to measure those cracks again? For some failures you know what you have to communicate...eventually it becomes sort of a routine.

*DC2-#3*: The statement, that is simply true...I think it improves the inspection.

*Casper*: Why do you think that?

*DC2-#3*: Because of the feedback and in particular those reports. Finding failures is rather easy. But reporting and the feedback, that is really important.

Reporting is seen as useful because the game has a structured way of doing this. If players do not follow this, they are notified via feedback. At first players may not follow this right away, but at some point it becomes a routine and maybe even too much of a routine, because players start to skip certain steps, such as not measuring the signals anymore (which is understandable, because the failures are identical).

The provided structure by the game may have found so much value because of the inexistence of structure in levee inspection organizations and the need for uniform and consistent procedures. As I just discussed one can have reservations about the exact implementation of the game, it does provide for a structure and it apparently gives an impetus to think about having one.

*DB3-#1*: I sit often in the Action Center and get to see many reports. The variety of reports of failures we receive during an exercise is immense and you know it should be of one and the same failure. Now you have to interpret this behind your desk while you do not know what is happening out there.

If you report yourself you realize that that is your image of what you see. People in the field have that too and I need to translate that when I am at the Action Center, because we have to decide what to do with it: Are we sending a bunch of people to the failure, yes or no?

You notice that making reports uniform and making sure they are similar to each other is becoming really important and we should actually train this somewhat more, not because people do not do it correctly—the desire to perform well is everywhere—but because you have this translation from what a patroller has seen out there. That is really difficult if you are not outside and do not have those standard images.

In this regard I think that this standardization could be realized with this program.

*DB3-#2:* You work here according to a logical routine. If he [the Action Center] does not understand it, it is because you forgot it and then you get to hear that from that telephone operator with a nasty comment.

The comments by Participant DB3-#1 indicate that the game made him aware of his situation and that of the patrollers. It raised the urgency to think about uniform procedures and he sees, similar to Participant DB3-#2, the game's procedures as something that could work and—rather more interesting—he sees the game as a vehicle for bringing about this standardization.

However, not everyone is convinced about the game's usefulness to teach about reporting procedures as exemplified by the comments of the following participant (DC2):

I find that you learn that whole routine with the Action Center rather easily with one or two [field] exercises. For me this was not anything new and I thought it was pretty obvious too. I found it especially educational to see various failures over time...because now I have a better idea. When I see things I know what could happen next. That is the essence of the program for me.

The person's background seems to dictate what they consider valuable. Of influence is most certainly whether a person has much outdoor experience or not. It further depends on the interpretation of what is meant by observing and reporting. The comment above reflects especially a definition of reporting as in the communication between the patroller and the Action Center. The act of reporting involves much more, because it is also about the classification system that is used and what needs to be paid attention to when reporting. It is in addition about the terms used.

### **Feeling comfortable with vocabulary**

In examining people's responses, a third specific learning objective that I noticed was getting to know the terms used. It is closely tied to reporting, but so is observing and many of the other learning objectives.

*DB5-#1:* Reporting I find the most difficult. But people have to understand you without any discussion.

*Casper:* You say it is difficult, but do you think the game helps you with that?

*DB5-#1:* If you practice more, you get to know the terms very well and if you feel really comfortable with them, then I think your reporting improves.

The participant indicates that for him it was first necessary to get accustomed to the terms in the game. Some of the terms are very technical. Think of macro-instability and micro-instability. During the regular inspection course these terms were used already and so with the game we simply continued the use of the terms (Level 2). Some of the participants made clear that they had some trouble with these terms.

*DA1-#1:* ...for us these are technical terms and I am just a regular polder boy. But now they at the Action Center also know what you are talking about, because those are educated scholars.



*Casper:* Do you mean you can communicate better?

*DA1-#1:* Well, if we have to communicate in this way, I learned something from this. If we do this with a radiotelephone I will probably talk with a boy who knows my language. It simply matters how much the distance is between practice and the Action Center. If all kinds of educated scholars sit there, I am nowhere.

Participant DA1-#1 indicated a difference exists between what happens in practice and what happens at the office at the organization, where the decisions are made. At the organization standards and procedures are thought of, such as what vocabulary is used, and this may very well not find any useful application in practice.

The participant further indicates that because of the game he is now comfortable with the vocabulary, but he would prefer to talk in his own language and would do so whenever he finds someone who would understand him.

The participant's reasoning is largely based on assumptions, maybe fostered by the terms used in the game, which confirmed to him that standards and procedures are thought of by "educated scholars" who do not know what happens and what is useful in practice. It is based on assumptions because in practice patrollers have to either call their post commander or they speak to a phone operator at the Action Center. The post commander is a more expert levee patroller, but he is still one of the polder boys. He knows their language.

The phone operator at the Action Center is usually commanded by one of the secretaries of the organization. They know hardly anything about levees, which is a reason for concern. At Organization B they said they need to train them and at Organization C they already started doing this. I spoke to a number of them during a field exercise and saw them in action (Level 10). They told me the training they received was too short and they would like to get a more in-depth training. Reason why is that they do not feel comfortable with their role. They are unfamiliar with the topic and have no idea what the terms mean or relate to. When I was observing them during the field exercise, they showed signs of stress and information overload. The situation was not too demanding, but for them it was because they have not received enough training.

Some of the concerns are shared in another discussion on this matter.

*DA3-#1:* You get more familiar with the terms used. At the start of this training I had no idea about it, but with the last exercise it went much better. I knew their meaning.

*DA3-#2:* I wonder if you still know them in about three years from now!

*DA3-#3:* I also wonder what happens with a real levee inspection.

*DA3-#3:* What I wonder is: now we know it but do the people we have to call know it as well? We have to build a sandbag containment ring! Until a week ago I really did not know what this meant.

These are questions and concerns that need attention, just as synchronization of terms and definitions is necessary.

*DB3-#1:* I think it is important to make sure the idiom, such as words as inner slope, outer slope, and crest, is similar to the ones on sheets and other material at the posts.

What does become clear is that participants became more familiar with some of the terms used and although this was not mentioned too frequently during the

discussions, it was part of the discussions a number of times, enough to highlight it as something separate.

### Seeing the big picture

Earlier on, with the discussion of DC1, I concluded that it might be that new employees and volunteers want to get the complete package to get the complete picture. One discussion revolved around the idea that this game enables them to see the big picture. This discussion confirms that this is especially true for participants with little to no experience.

*DB1-#1:* Look, I just started. I have been involved for a year now and I could not have achieved a more complete picture of the material without this simulation. Regarding this I think that when we go out next time you will have a much better image of the complete picture.

*DB1-#2:* I agree with that.

*DB1-#3:* Me too.

*DB1-#4:* I think this is faster, more effective, and cheaper than putting people together. You never know what the discussion is about and now you can make your own discussion.

*Casper:* Make your own discussion?

*DB1-#4:* Well, often you sit together and what can you talk about? You can talk about all kinds of things. This gives a reasonable total picture of what could happen. A lot of starters exist like me and now we can observe things for which you have not any answers for and you have the time to get them.

Seeing the big picture is what I consider the fourth and final learning objective. But about what does this big picture consist of? The following comments provide some insight.

*DB2-#1:* I started this year. Recently I received some education and now I am here. If I compare how I filled out everything the first time with what I filled out the last time, then it shows that I know what I am talking about. I did not have this at the start. So I think most certainly that this improves your knowledge about levee inspection.

*Casper:* What sort of knowledge improved?

*DB2-#1:* Different kinds of instability, micro and macro, how you need to report, the procedures, what happens if you make mistakes.

Participant DB2-#1 sums up a wide range of learning objectives. He emphasizes reporting but also talks about failure mechanisms and consequences. Getting to know and assess the consequences is what another participant found valuable too.

*DA3-#1:* You can assess the consequences much better. If you see something you know what the next stage could be. What I see now might not be bad but the next one is going to be bad. I am able to assess this.

The big picture is about gaining more feeling with the subject of interest by understanding what is happening and what is relevant in the larger context of the activity. If patrollers are only concerned with their primary tasks they are not responsible for assessing the situations nor for determining failure mechanisms. For this reason some argue that this should not be part of the training.

In a similar vein, some have argued that it would not be necessary to train failures and situations that do not appear in their region. Others seem to indicate they actually like to have a more general view on levee inspection.

*DA4-#1*: ...if you always walk on river levees, then you have little to do with those small failures but you do get to see them here [in the game]. If you only do those small water ways, then you get to see here the large river failures. You get therefore a more general view on things.

The big picture is about looking beyond your standard tasks and responsibilities. In case of levee patrollers this includes assessing, diagnosing, taking measures, and dealing with failures and environments that you will not encounter normally. Some oppose this, because in their view time and effort is spent on the wrong things. In addition, they are afraid that people may think that they are responsible for this.

Some of the participants were however glad to finally see the complete picture or as one of the volunteers said, to know the story behind their reports. Of course these learning objectives should not have a major emphasis in a training, but they do not have this in *Levee Patroller*. These learning objectives were by far not mentioned as much as observing and reporting, highlighting that the game is above all about these two other learning objectives.

### ***Next Generation of Levee Patrollers***

After the first sessions at Organization A, I found it important to include another statement, one about whether this game should be used already or if it is for the next generation of patrollers. The use of games for education and training is especially linked to children—those that grow up with playing games. Whenever I talk about my research, the first thing how people respond is something along the lines of “That is very interesting, especially with all those kids playing games. This might be the future of education.”

Many scholars have written generations that are growing up digitally—among other names, the Net Generation, the Digital Natives, the Homo Zappiens, or simply the Game Generation. Although maybe a few of the participants could be considered as part of such generation, most of the participants are not part of it, if we of course assume something like this exists (Bekebrede, Warmelink, & Mayer, 2011).

It is, however, hard to deny that the younger generations are more literate and interested in digital technologies than the older ones—something this study confirms. We have also seen that certain participants are clearly not pleased with this game-based approach. This makes it relevant to discuss this matter further and that is what I did during the discussions at Organizations B and C. The first time I posed the statement I did not do so very tactfully...

*Casper*: [After showing the statement on the screen explains that] The average age of levee patrollers is, well, a bit on age and...[loud mumbling starts in the room]

*DB1-#1*: [Among the mumbling one shouts sarcastically] You have to continue like this!

*Everyone*: Ha ha ha ha [Laughing out loud].

*DB1-#2*: Is this a statement or an assumption?<sup>1</sup>

*Everyone*: Ha ha ha ha [Laughing out loud].

*Casper*: OK, it is clear. Next time I will approach this differently. But do you agree or disagree with this statement?

*Everyone*: DISAGREE!

*DB1-#3*: You are never too old to learn.

*DB1-#4*: This is also for our generation.

Although the disagreement with the statement was universal, reservations were expressed, relating back to the computer skills and game attitude of the participants, as illustrated by the next discussion.

*DB2-#1*: I totally understand that people exist that have difficulty with it. I am not a fan myself.

*DB2-#2*: That will not change in the future. Even very inexperienced patrollers could be like that.

*DB2-#3*: I never play games. Never!

*DB2-#4*: This is not a game. You should not approach it like that.

*DB2-#3*: Yes, but I do not have the required computer skills.

*DB2-#5*: Some are more handy than others.

*DB2-#6*: You cannot walk with such a thing all the time [pointing to his mouse].<sup>2</sup>

*DB2-#7*: Why not? Does it not go by itself? That is not so difficult.

*DB2-#6*: It is difficult to me.

*DB2-#7*: You just have to do it ten times.

*DB2-#6*: If I do it one hundred times!

*Casper*: For some people it remains difficult.

*DB2-#8*: It has an exit button, right?

*Everyone*: Ha ha ha ha [Laughing out loud].

Two other interesting aspects of this discussion is *a*) the suggestion that the situation might not change in the future and *b*) seeing *Levee Patroller* as a game or not. These two themes come back in the second discussion at Organization C (DC2). Here one participant (DC2-#1) plays a central role. He is one of the younger participants, but I already noticed at the start-meeting that his computer skills were not so great.

*DC2-#1*: Someone who works with a computer, who has computer knowledge, that someone has a giant leap ahead.

*DC2-#2*: You speak of a giant leap ahead. Do I get from this that you want to be the best? Do you see it as a game as in you want to belong to the best ten players?

*DC2-#1*: You just want to play somewhat well. Otherwise you can just sit on a levee and say guys do whatever. Even if the levees breach, I do not care.

*DC2-#2*: OK, but that is a different motivation. Saying you need to be more handy is different from saying that you need to be handy to even play this game.

*DC2-#1*: Take filling out something. For someone who does this every day it is type type type and it is there.

*DC2-#3*: It is about whether you are busy with the game or fighting against the keyboard. That difference you have to take away. Only then you train and learn.

<sup>1</sup> In Dutch this is a pun. The word for statement in Dutch is *stelling* and the word for assumption is *vooronderstelling*.

<sup>2</sup> DB2-#6 is Pieter (Participant #67), the participant who played all exercises and still was not able to play it (Level 5).

DC2-#1: I have to do it more. The first time you do such a game you have to discover how everything works...and if you do it much, you do get better.

From the discussion it becomes clear that participant DC2-#2—the coordinator at Organization C—would have issues if participants would have approached it as an entertainment game and would like to be the best in the game, not to learn something about levee inspection. This is a difficult tension, especially because *Levee Patroller* does include scores and could create for competition among practitioners.

Because I worked with different existing groups at Organization B I suggested to create a competition among these groups as a future possibility in one-to-one conversations and I got positive responses. In addition, I also got questions during the meetings about how well the group did compared to the other ones. The fifth group consisted of a mix of two groups (Level 4) and I noticed a healthy competition between them. Members were making competitive remarks to each other, such as “That must have been the people from region X” and “We from region X know how to do X, Y, and Z,” and laughed about it. They even suggested themselves to make it into a competition:

DB5-#1: You could make this easily into a competition. That seems fun to me.

DB5-#2: Good idea!

However, when I actually proposed this during the fourth discussion at Organization B (DB4) I got a much similar response to that of Participant DC2-#2: “Then it becomes a real competition, a game. It is not a training anymore and I find that to be most important.” Clearly not everyone is fond of the idea of using game-like attributes in this setting, most probably because this would take away the serious nature of the task at hand.

The remainder of the discussion there revolved also around how to look at *Levee Patroller*.

DB4-#1: I know of plenty employees who refuse to work with a computer.

DB4-#2: I think it really depends on why you do it. You [referring to me] just talked about the game about Hugo Grotius and that does not interest me at all. I try something like that a few times and then I quit. It does not have a purpose, but this [referring to *Levee Patroller*] has a purpose and that makes it acceptable to me.

DB4-#3: A number of times the word *game* is used and I do not have such a game feeling. It is more of a way to learn about what it is about.

DB4-#4: No, you should not call it a game...it is a training.

DB4-#5: I think that if you call it a game, you undermine it.

DB4-#6: Plus the fact that the youth does not find the game any interesting. Just this morning I was playing an exercise with my son. He said “Geez, that is kind of boring. Do you have to walk like that on a levee? Then I do not want to ever become a levee patroller!”

Casper: Did you experience this too? [asking this to the youngest participant that evening who was 18 years old].

DB4-#7: I would not want to call it a game, because it is way too slow and boring, but in terms of content it is much more fun. In fact, that is what makes it most fun and gives you a reason to put even more effort into it.

The term game has strong connotations, an important reason why some practitioners and scholars prefer to use other terms, such as sims or simulations (Aldrich,

2009). The participants do not want to refer to it as a game because it has a clear purpose in reality, it has serious content, they are learning something useful, and it does not have typical game elements that define most games.

They are completely right about these differences compared to the stereotypical entertainment game, but I have never introduced the game as such. I have always referred to it interchangeably as a serious game, simulator, or simulation. The discussions show that participants struggled with what they were using and how to perceive it themselves.

What also becomes clear is that participants have an internal motivation to play. They see its purpose and are intrinsically motivated to do their best. For this reason I have previously argued that it is conceptually important to make a distinction between motivation and engagement. Motivation is driven by content and engagement by game elements. Both could reinforce each other, positively or negatively, such as with participant DB4-#6's son. He was not motivated and not engaged.

Motivation to learn is not sufficient when considering game-based training. Like participant DB4-#1 indicated, some people simply refuse to sit behind a computer. That is what the following discussion with the rebel and co-rebel will illustrate.

*DC1-#1:* In fact you should not do this with the old generation...what I mean is that people still exist that have difficulty with a computer.

*Rebel:* I do not have any difficulty. I just have completely nothing with it.

*DC1-#1:* Yes, but should we not offer it because you are too old or should we say Let us do it right now because the levees could breach tomorrow and we should know our business.

*Casper:* That is indeed the question: should we implement this right now or should we wait another five years or so?

*DC1-#2:* No, just do it now.

*DC1-#3:* But I find that the people that cannot deal with this game might very well be a very good patroller.

*DC1-#4:* You can approach it differently. Let us take Jan. Jan rides on a boat and he is computer illiterate. He tries and you also need to want to try. If you do not want, well, then there is nothing we can do about.

*DC1-#3:* There has to be a necessity.

*Co-Rebel:* Yes, you can also do it in another way.

Although age most certainly plays a determining role (Level 7), just like participant DC1-#4 indicates, it is also how you approach it. Many participants are even older than the rebel and co-rebel and they take a different approach. Next to Jan (#125) another example of someone like that is Hendrik (#70). Hendrik is a participant who I had to teach over the phone how to send an e-mail. This is what he had to say during the discussion:

*Hendrik:* I had much difficulty with it, because I do not know how to work with a computer at all. I cannot even type. So I started this with an enormous amount of energy just to get something on that computer screen.

*DB3-#1:* But you did it in the end.

*Hendrik:* Well, I sink my teeth into it and then I want to continue and finish.

*DB3-#1:* That is the most important thing.

*DB3-#2:* That has nothing to do with generation.

*Hendrik:* Well, I am retired so I belong to the oldies.

*Everyone:* Ha ha ha ha [laughing out loud].

*DB3-#2*: It is mentality.

*DB3-#1*: Yes, that is the most important thing. The other guys also did not have anything with computers and they finished it too.

That is the other approach Participant *DC1-#4* was referring to. It is about having the right mentality. A person may not have such mentality, because they refuse to get involved with computers and other digital equipment as many of the participants suggested. Although technology should not be forced onto anyone without a good reason, times do change and this would require them to adopt these new technologies, as some of these participants were suggesting:

*DB5-#1*: I do not agree with that statement, because if you are standing on a levee these days, you have to work with cell phones, radiotelephones, and satellite phones too. You do not do that with pigeon post anymore.

*DB5-#2*: Indeed, you also do not come with a horse and wagon to a levee. You take a car.

*DB5-#3*: If you endorse this statement, you have stopped developing yourself.

What these participants mean is that patrollers already make use of much high tech equipment. In the future this will most likely increase—also with the inspection. Certain organizations are for example already experimenting with smart phones and tablets. The game is therefore no exception but part of a larger movement in which new technologies are used to improve the inspection.

However, and that is what the co-rebel referred to and what we have concluded before, a person may not have such mentality because he or she simply does not see the purpose or thinks other and better ways exist to learn the material. In those circumstances it is difficult to motivate someone, especially if this requires much effort and energy at the start. A person may see the need to use a satellite phone but not to sit behind a computer as has become clear from the first discussion at Organization C with the rebel and co-rebel. Those who eventually did invest the effort and energy into it seem to not regret it.

*DB5-#1*: I belong to the older generation and I am glad I did this.

*Casper*: Why are you glad?

*DB5-#1*: Because this was really useful.

*Casper*: It was not too difficult?

*DB5-#1*: In the beginning it was difficult for sure but that is with everything. At a point you get into it.

The conversation at *DC1* took an interesting twist. Here some participants started to argue the reverse as well. As the world gets more digitized some people lose touch with the real world. They need to go out and actually walk over the levee.

*DC1-#1*: I agree with the statement but only if the real-life exercise does not become under exposed. I have grown up with computers so for my generation this is ideal to learn the basics. But you should not say this instead of.

*DC1-#2*: You should also pay attention to one thing and that is that more and more people exist that do not know what a levee is or an outer polder...and at some point a lot of people will join the water authority after their studies and they have may have seen a levee but never lived in a levee area...They have no idea of what is happening...Then this digital thing is not enough. You have to walk over a levee with those guys.

Although participant DC1–#2 might be exaggerating, I think he makes a valid point by saying that we should not neglect the actual environment. This should be a concern, because modern society becomes ever more digitized and the water authorities continue consolidating, making the organizations and their constituents more distant from the land they are responsible for.

Some of this might explain the contrasting results when I posed the third statement, whether participants started to look any differently at their environment.

### ***Looking Differently at Environment***

I also explored whether participants looked any differently at their surroundings during the training. An important reason I included this third statement was this comment made by one of the participants at Organization A (DA3):

I know to what I need to pay attention to. Coincidentally I drove past by a quay with damage and reported that. I was also at another quay and I saw an ugly damage there too. Normally I might have dismissed it, but now I reported it because I thought “Wait a minute, that is a damage!” I have become much more alert.

This is not a unique phenomenon. When one starts running, one sees runners everywhere. Or when one is pregnant, one sees pregnant women everywhere. This is referred to as situation awareness (Endsley & Garland, 2000; Klein et al., 2006a): you become more aware at your environment when you focus on something. Weick (1995) would consider this as the enactment of the environment—it is how you look at the environment, what you take into account. Based on previous experiences, current activities, and what preoccupies one’s mind the environment takes shape.

In the first discussion group at Organization C (DC1) there was an awkward silence. Then suddenly one participant said “No, I am at the levees every day and I have not viewed them any bit different.” The room was filled with employees who deal with levees on a daily basis and volunteers who live on or close to a levee. All of them nodded in agreement. At the second discussion group at this organization (DC2) the following conversation took place:

*DC2–#1:* I see everything in squares now.

*Everyone:* Ha ha ha ha [laughing out loud].

*Casper:* But did you start to pay attention to certain things?

*Everyone:* No.

*DC2–#2:* I did become more curious to just walk over a levee and to look some more...I have become more curious.

*DB3–#1:* These past two days I have not been in the vicinity of a levee, but I do think it makes a difference if you work at the water authority or not. When I started working there I suddenly saw ditches everywhere.

Participant DB3–#1 was also one of the younger participants and she was one of the participants who played all exercises in about two days—two days before the end-meeting. That is why she has not been to any levee; she was sitting at home playing the game. She points out one of my suspicions, that it makes a difference if



participants are confronted with the material more often, which is what happens if you work for a water authority.

At Organization B I did not get awkward silences or no's. What may possibly explain this huge contrast is that most participants at Organization B were volunteers who did not live on or close to the levees.

*DB1-#1:* Yes, I look different at it.

*Casper:* And how?

*DB1-#1:* You see all kinds of cracks in a levee or in the asphalt. Then you think "Look at that, those are all cracks." Perpendicular and horizontal cracks, you see them everywhere.

*DB1-#2:* I bike once in a while over a levee and although I am on the bike I start to automatically look at both sides.

*Casper:* You did this before the training?

*DB1-#2:* No.

At the other discussions at Organization B, the reactions were similar. One participant said he had seen every crack in his environment. Another talked about how he recognized signals from the game in his environment. A third person said to started looking for cracks instead of birds, something he apparently usually does.

Participant DB1-#2 comment indicates that it is not only about what people might observe. It is also about *how* they observe. This participant started to approach his environment from multiple perspectives—by looking to the left and right as well. This difference was a central theme in another discussion.

*DB3-#1:* You notice sometimes loose stones and normally I do not do that much with it, but now I have something like "Wait a moment, that could have some consequences." The awareness increased.

*DB3-#2:* I do not think I look that much differently at my environment. I do think that if I walk over my levee segment tomorrow, I will look different at it. You have become much more aware of those pitching stones, floating trees, and so forth.

*DB3-#3:* That probably happens because the simulation shows you what can go wrong. You would have never seen that otherwise.

*DB3-#4:* I walk now with different eyes.

*DB3-#5:* You see clearer what happens.

*DB3-#6:* I think that the next time you walk your levee segment you indeed look different at it. I also think that it makes a difference where you walk. You know now exactly where failures will appear. Sand boils happens in that ditch in the hinter land and that crack appears on the inner slope. Otherwise you would just walk over the crest, look to the left and right and think to yourself "It will be alright."

The stated confidence by Participant DB3-#6 is something other participants also highlighted.

*DB5-#1:* I have got more self-confidence. I drive every day passed by the levees and I now have something like let that crack come! I am ready!

*DB5-#2:* I want some action too!

*Everyone:* Ha ha ha ha [laughing out loud].

*DB5-#3:* In the beginning you think is that a crack or a puddle? I got more peace on my mind knowing that I am practicing.

Of course the participants were somewhat joking around, but such confidence could be a reason for concern. The game does not portray every possible situation

and reality is different in several ways. This concern was stressed by one of the more expert employees.

*DB2-#1*: I am afraid that people who have no feeling with it—and it is great that they all come and help—that they think that if they do it like this on the computer that they could do it just like that in practice. But when they actually walk outside in the middle of the night they think “Oh, how do we need to do this?”

On the other hand, it is encouraging that patrollers have become more aware of what to expect and more confident in their abilities. They further know better what to focus on, which environmental cues they should respond to (see Participant DB3-#5; “see clearer”).

Additionally, the raised awareness might be a result of knowing what the consequences are, as Participant DB3-#3 suggested. Beforehand I figured that it was not too hard to think of the consequences and that people who join a levee inspection organization would be very conscious of this. Yet the results suggest otherwise. It is important to show patrollers what the consequences are and to let them think about failures. To manage the unexpected, you need to be preoccupied with failures (Weick & Sutcliffe, 2001).

However, for those preoccupied with it already it does not count. I refer to the expert employees, the ones who on an almost daily basis are confronted with it. Those expert employees at Organization B indicated just as the participants at Organization C that the game did not make a difference in this regard.

We cannot talk about that. It is my occupation at the water authority, so you do not say you look different. You already look with various backgrounds at your environment.

Expert employees were already aware and had a mental model. The game simply added to the “various backgrounds.” This means that the game has made participants more situationally aware, but only if the participants were not preoccupied with it already because of their occupation or because of their living environment.

## Suggestions for Improvement

The final two discussion items are two questions related to the hypothetical situation that the game will be improved and further used:

1. Suppose *Levee Patroller* will be further developed. What three improvements are necessary?
2. Suppose the use of *Levee Patroller* will be continued. How should it be used?

In terms of game design, I asked specifically for three improvements, because I wanted participants to focus on the main issues they experienced and not talk about minor issues, such as textual inadequacies or a small glitch in the virtual landscape.

### ***Three Improvements? Make It Four***

Many possible improvements were raised during the training. In discussing the results of the game questionnaire I highlighted already some of the frustrations participants had with the game (Level 5). Most evident were the communication with the Action Center, the orientation in the gameworld, and of course the notorious drifting mouse pointer.

The emphasis on mentioning “three” improvements did not help much. Most of the participants started laughing when I said that, because they were ready to present me a laundry list with improvements. Leaving the mouse issue and other control and interface issues aside, I coded four important themes, the first of which we are already familiar with.

#### **1. The bureaucratic action center**

When I announced the improvements statement one participant immediately said “At the very least it should have another Action Center” (DB2) indicating his enormous frustration with it. The reasons why participants felt frustrated with the Action Center became clear from the game questionnaire, as summarized here:

*DA3-#1:* Worthless part of the game is that when you observe something, report that to the Action Center, and then report that again ten minutes later again, he tells you “Why do you call me?”

*DA3-#2:* That is not reporting but communicating.

*DA3-#1:* Yes, but it is stupid that they say “Do you call me for a chit-chat or something?”

*DA3-#2:* Such a response by the Action Center might very well be close to a real response, because if a levee breach is really about to happen then many people will call them. If nothing is happening, you can expect to get a remark.

*DA3-#3:* It also happened that I felt nothing changed and he still wanted me to go back. I found that really annoying.

*DA3-#4:* You do not have to follow his orders. You can neglect it.

*DA3-#2:* What I found more annoying is that when you are about to report something new, you cannot do it because he sends you to something else.

*DA3-#5:* I had this once. The Action Center says you can take a measure but then it was not necessary according to that levee expert. That is kind of contradictory.

Some frustrations are based on misperceptions, such as the latter. As I explained in Level 5, the levee expert is out there in the field and sees the failure, while the Action Center does not. An improvement would involve something to make these roles clearer.

Although it certainly has not been our intent to make people think that they could neglect orders by the Action Center, players do not have to go straight to a failure when they are requested to. Players could finish what they are working on and then go to the failure. But also here the game might not be clear enough.

Beyond these known frustrations, the discussions revealed some further insights into participant’s issues with the Action Center.

*DA2-#1:* I think it is set up too much in a bureaucratic manner. You call the Action Center and he tells you it is not correct what you are seeing. I think to myself: are you here or am I here? Then at a sudden moment a gigantic amount of water runs over a levee and I call to report that. He tells me that I did not fill out the report correctly!

*DA2-#2:* A report is not important at such a moment.

*DA2-#1:* I thought so too!

*DA2-#3:* I do not think it is bureaucratic. It works that way because it is a game. The computer has to react on certain things so it is just the game that works that way. You give a reaction and the game has to react to it.

Here the participants indicate that the game works too much “according to the books” and not how it would work in practice. Another participant experienced the same problems as Participant DA2-#1.

*DA4-#1:* I had once that the levee was settling and I thought that it was about to breach. I called the Action Center and he tells me I still have to measure and that I have to do this and that. I thought to myself there is nothing I can do anymore. I quickly put down a measurement marker but this flowed away immediately.

Participant DA2-#3 seems to understand why it works according to the books. He is more or less saying that it is a game and a game differs from reality. It works according to a set of rules that are strictly enforced, even under circumstances in which we normally would let go of the rules.

I would say it is rather because it is a learning tool. We could have implemented rules that could take care of such exceptions. We decided not to, because we wanted players to make proper and detailed sense of the failures. Otherwise it would come down to simply reporting that a huge problem exists and it needs to be taken care of. Few would not be able to do this in extreme situations. It is more valuable to teach people what is actually happening.

Maybe the largest frustration people had with the Action Center is that they cannot respond back. As one of the participants ended one of the discussions, “In reality you can at least yell back at them” (DC1).

## 2. Simpler and easier

A frequent remark was that players had to perform too many actions, making playing the game tedious. This might be an underlying reason of why some people grew frustrated with the Action Center too.

*DA2-#1:* I was about 1.5 to 2 hours preoccupied with it. Most certainly the first two times. If you missed something, then you had to redo the whole routine. I would say that if you have that guy on hold that you just say “Hey, this is not correct” and then you have such a button that enables you to quickly change it. That way you stay in the conversation and you do not have to go back all the time. You forget this...oh shoot!...but then click and ready. Or if they [the Action Center] think that is not possible that you can say immediately “Sorry, you are right, it is not critical.”

It was a punishment for people to continuously go back and forth. For determining the location, which people forget often, we already implemented that if players

forget this that they go to the location menu immediately after the phone call. This eases playing the game. For many other issues, such as forgetting a reporting item, they still had to go back to the report themselves, change the item, and then call back. Participant DA2-#1 would rather have seen that he could do this during the conversation.

Although such a solution makes the game less tedious, this tediousness ensured that players would not make such a mistake too many times. I have heard many players say that they made sure to not forget to report the location before calling the Action Center after having to go back and forth a number of times. If forgetting is forgiven rather easily by pressing a button, I do not think players would learn much.

To prevent the monotonous routine with the Action Center, another participant suggested that they should be able to hang up the phone at any time during the conversation so as not have to endure the pedantic remarks from the Action Center. This is a remarkable comment considering the target group's desire for a highly realistic game.

All of this does not take away that the game is complex, especially in its reporting structure, that it will require time and effort to get used to it, and that as a result it is experienced as tedious at times and in particular at the start.

*DB3-#1:* The first couple of lessons you're really busy to understand the game and then the game misses its purposes. You're not busy getting to understand how to report. You're busy getting to understand how the menus work. It could be simpler.

Learning to play the game before learning from the game is inevitable to some extent (Level 5), but one does need to make sure that the game does not miss its purpose. According to another participant, the game misses its purpose and needs to be simplified for other reasons.

*DA4-#1:* I think it has to with some kind of competition. Maybe that is part of it to make it into a game or something. I think you need to let this go. You have to keep it simple. Simplify. Simple. You have to be able to take the time to see how you could do something. Now I went through it extremely fast. You can also show circles to indicate to search in this area or search in that area. You're lost too many times. You're searching for too long. You have to keep that in the background. That is not the sport. Not in this story.

*Casper:* Do you mean you have to make too many actions?

*DA4-#1:* You have to minimize those too. You see a failure, you make a picture, and bang you submit it with GPS information attached to it. Then they [the Action Center] immediately know where it is...you show a circle in a region to indicate that is where you need to be. Then you accelerate the game and you put down a marker to indicate that is where it is. You immediately see the sizes, how large it is, what width, so you do not have to measure...just bang and you get it. That is it and ready. It needs to have some acceleration.

Participant DA4-#1 goes much further in what needs to be simplified or eased. He wants support in several ways, so more or less the only thing the player has to do is to search in a demarcated area for a failure. This high tech way of inspecting levees would be great in reality. In terms of learning it basically comes almost down to a sophisticated way of search the differences in a picture. It would be focused on observing only.

Yet, a single exercise does take up a lot of time and certain game procedures too, such as measuring.

*DB1-#1:* Working with those measuring markers is time-consuming. At some point I have not used them anymore. It takes up too much time to measure everything and especially if you have four failures. You do not make that in 24 minutes.

The measuring markers have been a major design issue and they continue to be an issue. Another one concerns the responsibilities of patrollers. Some expert employees (DC1-#1) find the inclusion of more complex tasks, such as diagnosing and taking measuring, not necessary. Volunteers (DC1-#2 and DC1-#3) seem to like it.

*DC1-#1:* This is meant for levee patrollers and so I think you should leave taking measures and diagnosing out of it. That is confusing for many people.

*DC1-#2:* Well, I actually found it fun.

*DC1-#1:* I also say this because if you are computer illiterate it is easier to deal with such a program if you do not have to take measures and so forth.

*DC1-#3:* OK, but you can also increase the game in difficulty.

Adding responsibilities is most certainly another variable by which the difficulty of the game could be increased. I decided not to do this as part of the training, because I was afraid such a change would be confusing to players and I already tweaked a number of variables (Level 4). However, getting rid of these responsibilities does make the game simpler and easier.

### 3. More, please!

A nice contrast between simplifying and making it easier to play the game is that a number of players wanted more. Most clearly they wanted more feedback, although not everybody agreed with this.

*DC2-#1:* I think you need to improve the feedback in the game. It is not clear why you get sixty, seventy, or eighty percent. It makes you think what did I do right or wrong? You have to make this clearer. You also have to enable to return to that program [the end-game feedback screen]. If you cannot go back to it.

*Casper:* To clarify, do you also want more feedback during the game or only at the end?

*DC2-#2:* Both please, because then it sticks better. At least for me.

*DC2-#3:* I do not agree with that, because if you look at the statistics tool, you get a good clue if you are in the right direction. Or you make a difficulty increase in which you provide more feedback at the beginning and decrease this later. I personally only have a need to get it at the end. There I did have once or twice that I did not understand why I had less than 70% correct...But that was really the only gap.

*DC2-#4:* Well, I do not agree with you. You are probably a water authority person I assume and we are not, so we need that feedback earlier than you, because with you it is already in your brains and with us it is not.

*DC2-#3:* I am not a levee expert at all.

*DC2-#4:* No?

*DC2-#3:* No, but I know everything about ditches.

The game has two feedback mechanisms. The first mechanism concerns the statistics tool from the inventory which allows players to keep track of their performance throughout the game. This tool does not tell you exactly what you did

right or wrong, but if players check it frequently (which I recommended them to do) they will get a good feeling when and why they receive points. Participant DC2-#3 seemed to agree with this. Others seem to have liked to see more and clearer feedback right away.

The second is the end-game feedback. This is an overview of where the failures were located and a description of how players dealt with them. Unfortunately, the button used to go from this overview to the game questionnaire was labeled "Next." Participants thought that by clicking on this button, they would go to the "next failure." Instead, they went to the game questionnaire and could not return to the end-game feedback. This explains the comment by Participant DC2-#1.

For both types of feedback it is true that it is difficult to trace what participants did right or wrong when it comes to reporting. They only know that a report is considered accurate if 70% or more is filled out correctly. Only the better players, such as DC2-#3 and DB1-#1, seem to make this specific comment. Others spoke in more general terms about not understanding what they did right or wrong (see the other comment by Participant DC2-#1).

*DB1-#1:* I got some low accuracy scores with my reports. I could not find anywhere what I did wrong. You do not learn much from that.

Besides more feedback participants gave me a list of aspects that they wanted to have included or more of. Here's the list:

1. Weather and water information, such as the water height.
2. Mechanical failures, such as a sluice that is not closed.
3. Other regions to practice in.
4. Inspect in the dark with a search light (and then one person commented, "You can also just turn off your screen").
5. Include the remaining failures from the manual, such as animals and floating waste.
6. More variety with the failures.
7. More refinement of existing failures.

Most of these additional wants and needs were made by the expert employees. The game may not have provided them enough challenge and/or they know there is much more that could be included. Regarding variety one of the experts said:

*DC1-#1:* If you see a crack in the inner slope, then it always had a puddle. After three exercises you have that image, and so I only look at that puddle. The crack I hardly see.

Variety with the failures is about subtle or not so subtle changes to failures. It could involve variation in length and/or width of a signal. It could also involve the inclusion or exclusion of signals. Sometimes the puddle does occur and sometimes it does not. Such variety might challenge players more or encourage them to maintain their attention.

Another desire that needs some explanation is the refinement of existing failures. This is best explained with the words of the expert himself:

*DB1-#1:* The failures need to be elaborated upon. If a ditch is closed off by soil, then you are supposed to see signals over the complete levee. Now you only see a crack at the crest. That is true for other failures too. If pitching stone disappears, then you have to see those stones somewhere and you expect a settlement at the top of the revetment.

Except for the request about improving the feedback, the mentioning of additional features did not take a prominent place during the discussions. Participants preferred simplification over expansion of features.

#### 4. A trial-and-error exam

One interesting improvement had to do with how to learn by playing a game. Quite a number of participants did not like the way the current game was set up, which to them too much resembled an exam or test.

*DB4-#1:* ...on the one hand I find it a beautiful thing...it has gorgeous animations...On the other hand I found it less fun than I expected, because to see how such a mechanism works and changes over time, what signals it has, you cannot see that really well. The way the game is designed is that you start with taking an exam. You have to look for failures right away and report about them.

*Casper:* You spoke of an exam?

*DB4-#1:* Yes, you take an exam immediately. What I expected is that in such a virtual environment, such a simulation, you can follow the whole process...That it shows what happens from the very earliest beginning till the eventual collapse. You cannot follow this very well at the moment. You get at a point and you see something. You report and then you go somewhere else, because multiple failures exist. So you have to leave the area where it is happening but in the meanwhile that process continues and you cannot follow it.

*DB4-#2:* If you are on a levee you cannot stand still. You have to go on.

*DB4-#1:* But that is reality! With this program it would have been nice to see how it evolves over time.

*DB4-#3:* If you go back to failures you do see them get worse.

*DB4-#4:* I understand what this sir is saying. Basically you should have a number of separate failures. With those you can see how failures develop and learn how to report them without being accounted for it.

*DB4-#5:* I think the current manner forces you to think about what these failures are. You have no choice...If you do not know it, you have to delve into that manual. That way you get the required knowledge.

*DB4-#2:* And you are forced to continue. I find this important too.

*DB4-#4:* I think that if this is a training that you have to—like the sir is saying—see such an image a number of times. You have to be able to stand still and look at how it develops to recognize the different stages. You can use this recognition when you play the game. Therefore I say you need to have an additional learning aid that has five or six modules that you can play to see how it develops.

*Casper:* do not you see the complete development of a failure at the end? [the end-game feedback shows how each and every failure develops over time]

*DB4-#1:* A training or an education or whatever you want to call it consists of a module information and at the end you get assignments. That is a normal setup of a training.

*DB4-#6:* While playing you do not know how it progresses and in the beginning I knew little about it, but then you do see how it progresses, you make a number of errors, and when you're done you know much better what you did than watching a movie. That is the major advantage.



Participant DB4-#1, a distinguished and well-spoken gentleman (and who was strikingly referred to as “sir” by another participant), started the discussion right from the start and it went on for a bit more. What it came down to is that Participant DB4-#1 did not see the game as an opportunity to learn, but as a test and for such a test he wanted to be prepared. He wanted to have information and insights and he did not have these or could not find this—despite his participation in the levee inspection course and his attempts to look up information in the manual.

That is correct, because some of the knowledge can only be retrieved by playing the game, meaning that players may not do it right the first couple of times. Players will learn from trial-and-error, discovery learning, or from experiential learning, which is at the essence of game-based learning. Participant DB4-#6 understood this well.

Other understood it too. By including scores players are forced to pay attention. “You have no choice” as Participant DB4-#5 put it and thereby “you get the required knowledge.” If players only look at the development of a failure the effect may not be much different than watching a movie.

This discussion may highlight a difference in learning between participants, which may or may not relate to a generation gap or experience and understanding of game-based learning. And Participant DB4-#1 was most certainly not the only person with issues regarding game-based learning.

*DA1-#1:* You had to find out everything yourself. It would have been better if you discussed everything step by step as in if you find this you have to handle so and so and if you encounter that problem you do it like this. Now you had to wrestle through it by trial and error. After playing a number of exercises you knew what to do and it went much better, but still, every exercise was a bit different and my scores were extremely poor. They were all insufficient.

Participant DA1-#1 highlights that also he wanted to have clear instructions on every game aspect before “taking the exam.” Something else this highlights is that participants are sensitive to poor scores and especially if these do not improve. That is understandable and probably reinforced their dislike to learn this way.

But it above all reflects a different view on their role in the organization. In their eyes they should receive clear instructions on what to do and they will follow this. They do not see it as their duty to deeply think about failures—not in practice and not in the game. The next suggestion exemplifies this view.

*DA2-#1:* I just want to report to the guys here’s an emergency. You should actually have a plastic sheet with a picture and which indicates what kind of failure it is.

*DA2-#2:* YES! You should definitely have a sheet so you can say to the guys that is it...it is too much sorting and unraveling. Before you made up your mind he is already calling you with “Where are you?”

*DA2-#1:* You should do it as at the Chinese [restaurant]...You look at a picture and then you could say it is number eight. That way you can shorten it and then an explanation appears about sand boils or micro or whatever it is called. I just see the levee settling and what it really is...

*DA2-#2:* [finishing DA2-#1’s sentence]...that is not our business!

These participants essentially want a sheet with all the answers on it. Playing the game then turns into a guess the picture test or a fill-in-the-blanks exercise. Although

in practice it might be better to give patrollers these types of aids, a distinction should be made between practice and the game. In the game patrollers should learn about levee inspection and then it is necessary to force them to think hard about what they see. That will not happen with all the answers on their lap.

These comments may also reflect that the game might have too steep of a learning curve for some players. Although I paid special care to make the training incrementally more difficult, some player would have benefited from smaller increases in difficulty.

*DB1-#1:* Maybe you could do the setup somewhat different by for example not going from one failure to four right away. You first do a couple with one or two and then the program has some benchmark moment that measures if you understand it and only then it adds a failure or two...now it went immediately boom!

For similar reasons participants suggested giving extra feedback at the start, work with a coach who gives tips and suggestions, make it possible to replay previous exercises, or let them decide when to progress to the next difficulty level. Some of these solutions exist already. In fact, even the failure modules that allow players to see the development of a single failure as suggested by Participant DB4-#4 have been developed. I disabled or excluded this in the research version for experimental and educational reasons (Level 3).

I think such scaffolding supports will help people to get over their test fears and feel more comfortable to learn in this interactive environment. Yet, people will remain that will not like to pursue what they consider a trial-and-error examination. I think this relates to deep views and beliefs on their role within the organization and on how and what they should learn.

And what needs to be done in terms of learning support remains a trade-off. Providing too much scaffolding support and/or information might detriment the power of game-based learning. One of the participants warns against this.

*Casper:* So you need to get some more theory before you go and play?

*DA3-#1:* Well, I do not know. This might take away the surprising effect. It is fun to learn this by playing.

Some see this as an exam or an assessment. Others like DA3-#1 see this as play.

## *Using Levee Patroller*

The last discussion statement was about how to use *Levee Patroller*. The game itself may need some improvements, but it might be even more important to think about how it could or should be implemented within the organizations, assuming of course its usage is continued.

I found this statement also relevant because how the training was set up concerned an experiment in and of itself. I reasoned about a possible efficient and effective usage of the game (Level 4), yet I had no idea of this would really work. I would not have been surprised if hardly anybody finished the complete training.

Yet, the participation rate was beyond my expectations. This already showed that the setup was successful. After posing the statement some of the participants confirmed that this was a good way to use the game.

*DB1-#1:* You should in fact do it in the same way as we did. You enroll. You get an introduction of an instructor and you play 2/3 weeks and after that You have a closing like now.

Others thought that it could not be done better.

*DB5-#1:* I just found this a very good setup. I would not know a variant that is any better.

This does not mean nobody made implementation suggestions. Especially the participants at Organization A made many suggestions that were easy to change and I implemented some of them for the sessions at Organizations B and C, such as an improved manual, a demonstration of how to play the game, and a peek sheet (Level 4).

It also does not mean we did not have any discussions. The major issues we discussed concern to what extent it is necessary to meet, whether to make it obligatory, what the role is of the game within the education of patrollers, and how it could be integrated within such education.

## **Together alone**

We know from the interviews that some of the volunteers are drawn by the social aspect. They hardly see each other throughout the year. It is also known that people learn from each other, by sharing and discussing information. These are good reasons to let participants meet each other during the training. Some even suggested that it could be an idea to play together.

*DC2-#1:* I think that with this type of virtual inspections it might be easier if you sit with two people behind a computer. That way you can complement each other and ensure you get a bit of a normal score. All alone you doubt much if you do something right.

*DC2-#2:* With two you might learn more. And in practice you also walk together.

*DC2-#3:* I do not agree with that. If I play such a game by myself I learn faster than together. But yes in practice it is different.

*DC2-#4:* If you play together you make a compromise. If you play alone you really give your own opinion. That gives a better view [of your own performance] than if you play together.

Of course, this idea was not shared universally.

*DC2-#5:* I personally do not need a start-meeting. I also did not attend this but I did not miss it too. I easily understood it and I think that people with the same skills would prefer to read this at home and then go for it and rather not attend a meeting. Maybe you should give people a choice. You do not make someone like me happy to request them to attend a meeting...I figured this out in fifteen minutes and otherwise I am two hours away with the same information, so for me it is not efficient.

This comment was from a young employee with good computer skills. Other participants who did not attend the start-meeting said the opposite. They said that they should have attended it. Overall most found the start-meeting essential and actually requested that this would be even more extensive.

*DC1-#1:* Better and more practice at the start [is needed]. For us it was too short. It is especially important for the computer illiterates to know how it works with the controls, because this gives you frustrations at home, especially since you are all alone. So it is good to know how it works and that you have some feeling with the images and the mouse and keyboard.

This is exactly what we practiced among others at the start-meeting, but considering these comments, it was insufficient for some. In fact a few argued that we should only play it together at meetings.

I would not recommend only playing together. First, participants would not practice more than what the time of the meetings allows for and it is this practice that makes the game powerful (at least that is my assumption and one that I will address in Level 11). Second, organizing these meetings is the big problem. More than 20 participants per meeting is not possible—in terms of hardware, assistance, and location—and at organizations with hundreds of patrollers many of such meetings need to be organized. Organizations would want to minimize such meetings. Besides this, almost everyone else experienced playing the game at home as “perfect” (Level 5).

Considering the overwhelming positive responses on playing at home and the difficulty for organizing meetings, I think this is something worthwhile. The start-meeting is worthwhile and especially with the current target group. Most need guidance before they are able to play at home. The low participation rate for those that did not attend the start-meeting proves that this meeting is essential.

If a large group becomes or is as literate in playing games as Participant DC2-#5, it is possible to give people a choice to attend the start-meeting or not, but making the start-meeting optional would very much lower the participation rate. Participants may underestimate what is required and feel less committed to the training (Level 3).

Commitment is a good reason to keep the end-meeting too. This end-meeting was especially organized for research purposes, yet it forced people to complete all exercises before that time. It further committed the participants to participate at all.

From the perspective of minimizing the meetings the end-meeting is however a sacrifice that can be made. The challenge is to find a proper replacement that would commit participants. One can think of combining the virtual training with another event for which patrollers already have to get together, such as a field exercise. Incentives can be thought of too from negative ones (you cannot participate with the field exercise unless you finished all exercises) to positive ones (you-finished-the-exercises-on-time achievement).

What does become clear from the discussions is that the participants liked playing alone at home but they wanted to have some togetherness too, to practice and share information. A number of individuals suggested even ways for being together virtually by making it possible to send messages to other participants for example. So the participants wanted more or less to play “together alone.”

### Voluntary or obligatory?

A significant number (5% to 10%) chose to discontinue or were unable to continue the training. This raises questions about whether or not the training should be obligatory. Organization C decided to make it obligatory right away because they wanted to stress the importance of being trained.

*DC1-#1:* For us it is a check that you are knowledgeable. If we send you on the levee later on, then we want you to be well trained.

*DC1-#2:* [who's a volunteer] I assume that a polder man goes with me who knows that exactly. I only go because of the environment.

As I explained in Level 4 the volunteers at Organization C were not too happy about the organization's decision to make it obligatory. From the previous discussion we see in addition that the volunteers and the organization have different expectations of each other. The organization wants well trained and knowledgeable patrollers, whether they are volunteers or employees, and the volunteers expect the employees to be knowledgeable. They are only there because they know the region.

At several discussions at Organization B I asked the participants whether the training should be obligatory. I received mixed responses.

*Casper:* Should we make this obligatory?

*DB1-#1:* You should not do that with volunteers.

*DB1-#2:* Well, I really think you should do this at least once every five years.

*DB1-#3:* For people who have trouble with the computer you should offer an alternative evening...otherwise you are sidelining those people.

*DB1-#4:* Maybe you should not make it compulsory, but you can make the necessity of it very clear.

The latter suggestion by Participant DB1-#4 is exactly what the volunteers at Organization C said too: do not make it obligatory but rather stress that participation is essential. At the same time volunteers also stress that some pressure is most certainly needed to make the training successful.

*DB3-#1:* I think that this setup should be extended to everybody, but does everybody need to go to a start- and end-meeting? I do not think so. I would just sent the CD and then via Internet submit a code to start and then another one when they are ready.

*DB1-#1:* Yes, but from another perspective it becomes then too non-committal and people will not do it. Then they are too busy with other things.

*DB1-#2:* I believe that too. I enrolled for this and then you go and do it.

*DB1-#3:* Yes, it functions nicely as a big stick.

*DB1-#4:* I think you should put some pressure on it. It should not be non-committal.

*DB1-#4:* You have to have some triggers, just like how we received those e-mails.

This shows that the current training setup had enough triggers to commit people to participate and finish the training, something the participation rates indicate. However, according to other participants the pressure was not large enough.

*DB4-#1:* I think it needs more of a big stick. I participated because I thought that it was kind of funny but then I have ten other things I need to do. I still participated but then I hear all kinds of people who made the effort to look up things and I think that is also its purpose. If you do not do this the levee does not have to breach, because I got a score higher than

50 or 60% every time. But then you did not run the thing for all it is worth and you are not being punished for it.

*DB4-#2:* This is of course very non-binding.

*DB4-#3:* If it means 50% that is no certificate or something else, you might be more fanatic.

*DB4-#4:* Yes, I wish that were true. That is my issue with the levee inspection. It is important we all know it. Also that one little chain link, because otherwise we still go under. So if we do not use this as a means to convince everybody what they ought to do, it is obviously not waterproof.

*DB4-#5:* I think it is also a bit of personal motivation...if your goal is to learn something because you do not know enough or you should know more, that should be a motivation on its own.

*DB4-#6:* Maybe you should do this as an exam as part of the levee inspection course...and then look how people perform.

*DB4-#7:* Like pilots go to a flight simulator to get their pilot license.

*DB4-#4:* This sounds heavy but if we really have trouble you should understand what lives are dependent on just a good report. That is something we do not think about too often, because it goes well. But if the threat comes around you have to understand it is life threatening. If you then go outside and you do not see all those things, something will go wrong.

*DB4-#8:* I do not think this is too intensive...[and]...If you start the levee inspection and people do not know it and something happens you are like a deer in the headlights...Maybe it needs to have some pressure. That is too bad but it has to be part of it.

What the participants in the latter discussion are more or less suggesting is that certification, something which is common with pilots and doctors, should be pursued. Without being certified patrollers are not allowed to inspect the levees.

The problem with such certification is that the games becomes even more of an exam and that is something that some participants already experienced as troublesome. For others who did not perceive it as such increasing the pressure might take out all the fun of playing the game. It will further make players even more critical of what happens in the game—any distracting or performance decreasing elements will be judged negatively. But maybe, as Participant *DB4-#8* was suggesting, this is a sacrifice that has to be made.

And then we have all those computer illiterates and unwilling and cybersick people that have to be dealt with. We further seem to be dealing with a disparity in expectations of what people should know and the relevance of learning certain knowledge.

Clearly most participants see the need and usefulness of the game, but how it should be approached and embedded, that remains an open question.

## **A repetition instrument**

What helps in determining how the game should be approached and embedded within the organizations is to decide first on its role. When I started the discussion about the use the participants at Organization A—who have not done anything for years—directly stressed that they would like to go outside.

*DA1-#1:* I think that next to this you should have an inspection in the field.

*DA1-#2:* What is happening in a storm they cannot show or only after the fact. But seepage and other similar things they can easily show and explain.

*DA1-#3:* The previous trip I really enjoyed. If some rough overgrowth exists that you can still see a leakage. Those types of things. You do not realize that and if somebody points that out that is incredibly valuable.

Going outside is important, because knowledge of the surroundings is valuable, especially when it is dark and the weather is bad. During another discussion participants tried to define the role of both activities.

*DB5-#1:* You should pay attention that the future generation of patrollers not only play the game and think they can inspect a levee after that. They should go just as much—or in fact just as little—to a levee as we do.

*DB5-#2:* It is a supplement.

*DB5-#3:* It is both.

*DB5-#4:* I think it is very effective. Where else you need two evenings to walk over a levee, you do it here in fifteen minutes.

*DB5-#5:* It however remains theory just as is walking over a levee, because that is theory too. At that moment you do not have a storm or high water.

*DB5-#6:* I think the program is a supplement to make it easier to report...If you do not do it, you get there too, but if you do it, you get there easier.

*DB5-#7:* But if you compare it to the levee inspection course and I did this last year, then I found this much more valuable.

Although they agree that going outside is important, the game's role remains unclear. Some say it is a supplement to going outside; others find it just as important or even more important. Participant *DB5-#5* made a comparison with the levee inspection course, which is the second alternative to the game. Although participants find the game valuable, here again they do not seem to agree on whether the game is a supplement to this course or not.

*DB1-#1:* I do not find this a supplement. It is much better than the course.

*DB1-#2:* That course gives you much theory and this is more like practice. That is more educational. It has a lot of repetition.

*DB1-#3:* But when you start with this blank, then this does not appeal, right? This is really a supplement...I think that if you do not do the course and start with this, then you will think "Oh my god, what did I get myself into!"

From all discussions it appears that the field exercise and the levee inspection course remain important and should not be replaced by the game. So what is the role of the game in that case?

*DB3-#1:* Do you see this as supplement to the course or as a replacement?

*DB3-#2:* Not as replacement. I think you should do the course before you can start with this. Otherwise you have something like what is sand boils and so forth.

*DB3-#3:* I think this is a very good supplement to the exercise we have done to explore your own levee segment. By making these [virtual] exercises at home you get to really see something and you retain your knowledge. I think this is with most of us: You take the course and then it is fresh but as soon as you start the [field] exercise you think "Sand boils? Hmm..."

*DB3-#4:* A repetition instrument!

*DB3-#5:* Yes indeed. Simply repeat repeat repeat.

Although many participants have stressed its enormous value, it seems that the larger consensus among the various participants is that the game concerns a supplement or supporting tool that allows to practice at home. The two alternatives, the field exercise and the course, have their merits too and should not be replaced.

The game has other values too, such as getting familiar with the failures, reporting, and the vocabulary, and seeing the big picture. But when discussing its actual role the notion of “repetition instrument” as proposed by Participant DB3-#4 seems to best describe how participants perceived its role. That is also how they continued to perceive it when describing how it would fit into a possible levee inspection curriculum.

### Quick follow-up and continuity

Regarding the game’s use two important questions remained. First, how should the game-based training be provided to those who have not used it yet? Second, how should those who have already played it continue to use it? As for the first question this was a much heard comment:

*DB4-#1:* It would be great if this follows quicker after the theoretical course, because you have to think in categories and I find the repetition it has really good.

*DB4-#2:* Yes, this is a great follow-up. I only did that course. I learned much from it, but if you do not maintain the information it will fade away. I also have to say, now I have been busy with this I indeed delved into the theory. I grabbed my books again.

*DB4-#3:* So it worked!

*DB4-#4:* Absolutely. For sure. But if I did not have that course I would have never been able to work with that thing. Absolutely not. It is not a replacement but a follow-up, a reminder.

Many participants indicated that one should first receive the course and after that play the game, because the game does require them to have some background knowledge. It is not impossible to play it without having such a course. Many participants did not receive this and they succeeded with it. Then again, it requires much more effort and they did complain that they would have liked to receive some more background information before starting to play.

The comment by Participant DB4-#1 reflects a vision that the game concerns a follow-up too (and of a repetition instrument). It should be started quickly after this course as to make sure all the information gets a place inside the brains of the patrollers (“to think in categories”). Others suggested that as a matter of fact the game-based training should become part of the course.

*DB4-#5:* I think that it is sensible to keep it as this training but then make it a part of the course. I think that I learned more from this than from the course. There you are trying to understand all that theory in one single evening. That really does not stick. Now you are busy with it much longer, it sticks much better, and you get a better understanding why you need it for. So I think it would be clever to make it a part of the course.

Comparing the game-based training to the course is not fair, because as Participant DB4-#5 indicates, participants are preoccupied with the training much longer. Yet, this again shows that it should not be seen as a replacement but as a way to



maintain and retain the information. Participant DB4-#5 unfortunately does not explain how it should be part of the course. Another participant makes such an attempt.

*DB3-#1:* I experienced that course as a lot of information and now I am a year further along the road and you get this. You learn more from this than from that single evening back then. But I think you need to cut that evening in two parts: The first evening you give a bit of theory and the second you repeat that and give an explanation of this program. At the end you give CD with this program and tell the participants to practice with it. You do not have to get together a third time to evaluate it. By e-mail you can tell how everybody performed.

The suggestion by Participant DB3-#1 does suffer from possible commitment issues, but he is right that it would be better to divide the evening in two parts. It is impossible to combine the current levee inspection course with an explanation of the game—at least not with this target group.

Participant DB3-#1 is not very complete as well. He does not mention what number of exercises people have to play. The number I picked was considered (Level 4) and I wanted to know if this was too little or too much and so occasionally I asked this.

*Casper:* Were the number of exercises too many or too little?

*DB2-#1:* Too many for my agenda.

*DB2-#2:* For me too but that is personal I think.

*DB2-#3:* Oh I did them all last weekend. Went perfect that way.

Nobody said the training had too little exercises, because generally they found it time-consuming. Although many proposed to only do two or three exercises others always reacted by saying that this would not help.

*DB4-#1:* Maybe you should do two or three of these exercises.

*DB4-#2:* I actually think that because of the number of exercises you learned something. The first exercise I forgot to report the location two or three times. That most certainly did not happen the fifth or sixth exercise...but I have been rather busy with it. It is quite time-consuming if you have to do this every year.

*DB4-#3:* You do not have to do all six of them. Every year you can say Let us take that one and that one and then maybe you take one that you play together.

*DB4-#4:* I think that you forgot all of it in that case. I think you should play all of them after each other. First as a starting course and maybe later you have to repeat every so many years.

Most discussions agreed that it is necessary to first have a similar type of game-based training as they received as a base. Playing two or three exercises after the course or over the year was not a widely shared opinion. However, the idea of playing over the year after receiving this base course was something that found a wide consensus.

*DB2-#1:* I do not think you are done with this all at once. You have to repeat this yearly. Maybe at the start of the storm season. It surprises me how fast things fade away if you do not do it for two months.

About how the yearly repetition of virtual exercises should proceed, many ideas were suggested. A frequently heard one was to do this just before the field exercise, as to refresh memories and to provide for a sense of urgency.

*DB3-#1:* You have to do this every year to keep it fresh.

*DB3-#1:* If everybody went for the first time then everybody can do it from home. You can just say you will get a code and you have to do these exercises because soon will have a [field] exercise.

Another suggested possibility was to inform people whenever a new exercise is available. Every year a couple of exercises should be made available as to make sure people get enough practice.

*DA2-#1:* I thought this setup was good. The repetition helps. Now you have to say "Hey we got another one and try to play this within a month." This keeps you fresh. Now it goes well and if you do it every half year or so you will stay awake.

A final suggestion was to create separate modules each with a theme and that consist of more than playing game. The modules could focus on failure mechanisms or regions.

*DC2-#1:* If we are going to use this year round to keep everyone fresh, maybe it would be more attractive by elaborating on a specific failure mechanism every exercise and combine this with some practice examples. Beautiful images and videos exist of levees that slide away...If you combine those with the exercise, everything comes more alive and you get the practice you ask for.

Not everyone is so excited to play the game year round:

*Coordinator:* My idea is that you go onto the levee during the winter and then two, three times a year you get an e-mail from the water authorities asking you to play an exercise on the computer.

*DC1-#1:* Well, I do not have such a need at all.

*Rebel:* No, me too. Totally not.

*Coordinator:* I just say something...The reason why is that we want you to be more committed...we are too far away...we only do one exercise per year and we try to send you a newsletter four times a year to keep a connection. In my opinion this would give an extra binding and that is why I would like to ask you to play two, three times a year...

(much murmur and grumbling in the room)

*DC1-#1:* I do not need that. You should rather organize an excursion within the area. Show something there what is useful, but to play a game...

What this part clarified is that if the game-based training is continued a similar setup should be combined with the existing levee inspection course. Continuation is further desired to maintain the knowledge and this could be done in various ways and a good way is to do this just before the yearly field exercise (if this is organized).

That is how the game-based training is largely perceived: as a follow-up of the inspection course and as a precursor of the field exercise (or excursion). This is how it could be integrated into a life-long learning levee inspection curriculum.

But we cannot speak of a clear consensus on all these matters. Opinions differ on the amount of exercises to be played and whether continuation should be pursued at all. With game-based training, it is not a one size fits all.

## Lessons Learned

Except for a few who clearly did not like the training, most participants were quite positive about it. They found it a valuable experience and provided little critique on the current setup. Participants found it only intensive—some even too intensive.

What became less clear is how the game-based training should be pursued in the future. Participants stressed emphatically the importance of outdoor activities, such as excursions and the field exercise, and those familiar with the levee inspection course indicated that this is essential too. The general pattern is that the game is seen as great follow-up to the course and an excellent precursor to any outdoor activity.

The clearest consensus was found on whether the game should be used with the current target group. Although everybody understood that not every member would appreciate this way of learning and that people exist with little computer skills, they were convinced that the game was something for them too.

Beyond those who do not like the game and never would like to engage with it again, it becomes clear that participants value the game for different reasons. Some stress its importance for observing failures; others for its reporting procedure. Then a number of participants talked about getting familiar with vocabulary or getting to see the big picture. Based on participant responses, I reasoned that

- Expert employees and volunteers with much outdoor experience value above all the game for its reporting procedure; and
- Regular employees and other volunteers emphasize especially observing and also getting the big picture.

The discussions made me realize another purpose as well, that of becoming more situationally aware. The situational awareness seems to only affect those who have little affinity with levees. They are not busy with it for work and do not live close to it.

What struck me most from these discussions is that the participants contradict themselves and create a series of dilemmas. These are the ten primary ones in order of descending importance:

1. They find the game very valuable and indicate that they have learned more from this than anything else, but then largely continue to speak of it as a “supplement” to other activities.
2. They want a realistic game environment, but they want help and support from the computer in any way possible (i.e., the mini-map) which actually makes the game less realistic.
3. They want to learn and without pain there is no gain, but they want the training to be shorter, easier, and simpler.
4. They do not want to be forced to play and participate, but say that they need pressure and enforcement to participate and continue to play.
5. They want to test and improve their performance by practicing, but they do not want to take a trial-and-error examination which involves learning by practicing.

6. They want to increase the importance of the game-based training (by certifying successful players), but realize that this would diminish the fun.
7. They do not want pedantic comments and other remarks that highlight they did something wrong, but they also indicate that these comments and remarks ensured that they did not do it again.
8. They want more feedback, but they do not take the time and effort to look at the feedback that is available.
9. They want to standardize the inspection and make reporting uniform, but they also want to play a game that is specific to their own procedures and environment.
10. They want to play at home and do it by themselves, but they also want to share information and play together.

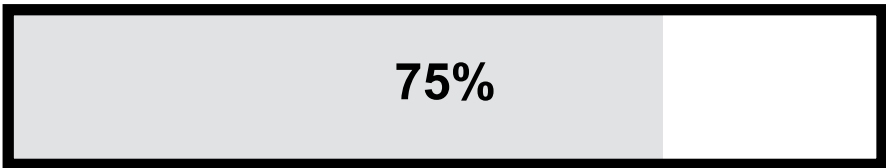
This is a list of contradictions for which no clear answer exist. However, these dilemmas do provide guidelines on what to think about when improving the game and the training.

## Level 10

### Picture That!

*Better than all that paper work. You can learn levee patrollers in no time everything that is necessary—A high school student*

*But I do wonder what will happen if we have a real levee inspection—Participant #29*



75%

We have examined the use and evaluation of a game-based training. But crucial questions remain unanswered, in particular about the real value and validity of the evaluation results. I employed four validation studies, each for a specific reason. This level details those studies.

First, I attempted to set up a training with *teenagers*, who in general are more computer literate and digitally well-versed than other age groups, but also complete *novices* in terms of the roles and responsibilities of the levee patroller profession. Including this group enabled examining how computer literate and complete novices performed in terms of game and test performance compared to the patrollers.

Second, a group of *super experts* was formed by an existing levee expert committee and they filled out the sensemaking test too. Inclusion of this group enabled to benchmark patrollers' test performance to an ideal situation, that of the knowledge level of a super expert. In addition to a better understanding of patrollers' performance, the comparisons with teenagers and super experts served another purpose too, which is validate the sensemaking test. If super experts performed worse than the group of 100% novices, we should be warned.

Third, I interviewed a number of participants before and after the training. The *interviews* served as a verbal sensemaking test, because It might be that participants are not able (or unwilling) to express their knowledge in writing, but are able to do this verbally. I also used alternative methods to get an idea of their knowledge. The alternative methods concern a knowledge elicitation and a cognitive mapping technique.

Fourth, Participant #29 wondered what will happen if a real levee inspection occurs. I wondered that too and the closest I could get to the “real thing” was a *field exercise*. In collaboration with Organization C I observed and compared the group of participants that participated with the game-based training and a group that did not.

The goals of this level are to describe

- How digital literates and novices performed in playing the game;
- How the levee patrollers performed in comparison to novices and experts;
- What alternative knowledge tests tell us about the effects of the game-based training; and
- What the effects of the game-based training are in another, more lifelike setting.

## Playing with Digital Literates

The first validation study concerned how digital literates and novices worked with the game. About 10% of the population sample of levee patrollers was younger than 30 years old and only 1% was younger than 20 years old. As illustrated in previous levels, on average the sample is relatively old and not so computer literate. This computer literacy is of importance, because players first have to learn to play the game before they learn from the game. Having little computer literacy will block users from moving to the second stage.

On top of computer literacy, game literacy in particular seems important. Game literacy asserts that playing games requires skills much similar to reading and writing. We should be careful with lumping all game literacy together. Someone might be very game literate in one genre and minimally so in another. With *Levee Patroller*, for example, I expected that players with much game experience in First-Person-Shooter (FPS) games would be better in picking up the game because in terms of visualization, movement, and control, it has close affinities with this genre.

We know that few levee patrollers had any game experience and so we can consider this another barricade to move to the second stage of learning from the game. The comments about how much the game is experienced as a trial-and-error examination are an indication that not every participant was as game literate (Level 9). Experienced game players would consider this a natural learning process and do not have any trouble in engaging with this.

## In Search of Teenagers

In looking for teenagers I considered three aspects: level of education, fit between educational background and levee inspection, and integration within a curriculum or another educational activity. I expected that most patrollers would not have received

the highest education available and then comparing their results to those of university students would be biased. Eventually I tried to recruit on almost all possible educational levels (but being aware of this bias).

If a fit exists between the students' educational background and the subject of levee inspection, I thought that they would be more intrinsically motivated to play the game. It would even be educationally relevant to them; therefore, I searched for studies that had commonalities with levee inspection. Such a fit would allow also for a possible integration within a curriculum or another educational activity.

I made four serious attempts to set up a training in parallel to the training at the three organizations, which resulted in three failures and a modest success.

### **The big fail**

My first attempt was with a University of Applied Sciences (in Dutch, *Hoger Beroepsondewijs*; literally 'higher professional education'). This school has a number of bachelor degrees in delta technology and had expressed their interest in the game already. They received the demo version of the game (called *Levee Arcader*) and used this so far for promotional purposes only.

I thought the training would be an excellent way for them to see if the full version was worth the investments and if it would fit in their education. In addition, I would be running the training and taking care of everything except for the location. The students were unfortunately on an internship during that time and so any integration within a curriculum was impossible.

In collaboration with the coordinator there, I invited the students to participate with the training as an extracurricular activity and I did not get one response. Then I put my initial discomforts about the compensation aside and doubled it to 50 euros—and still I did not get a single response.

My next attempt was at a vocational school (in Dutch, *Middelbaar Beroepsondewijs*; literally 'secondary professional education'), the type of school most patrollers attended. However, only two participants expressed interest, so I moved to my own university: Delft University of Technology. At the Faculty of Civil Engineering the first year students were not on an internship; they did take an introduction class into soil engineering. At the end of one of the lectures with an attendance of 200 students I invited them to participate. I also posted it as an announcement on the course's website and the professor sent an e-mail to all students.

Three students indicated that they would want to participate, but with all kinds of what if's, but's, and other additional conditions I had to bear in mind. I had my own conditions as well and so I decided to consider this a failed attempt too.

### **Some success with high schoolers**

One attempt occurred by mere coincidence. We received a request from a teacher at a Montessori high school if they could visit us to play some games and listen about

what we do. The visit would be part of a school excursion to Delft. We agreed to this if the students would contribute to my research.

On March 10, 2010, in total 21 students ( $M = 16.9$  yr,  $SD = .66$  yr) visited us and after a short introduction about levee inspection they played the training and start-exercise (and some even played the first one in addition to these). These students were in their fifth year at the pre-university level (in Dutch, *Voorbereidend Wetenschappelijk Onderwijs*). This is the highest variant in the Dutch educational secondary educational system.

The students played the game enthusiastically and seemed to have a good time. The atmosphere of the session was only much different compared to the sessions with the levee patrollers. The session with the high school students was one big chaos. The students were talking to each other left and right and were even shouting their game experiences across the room. Some even got up and went to someone else—to ask for help or to explain something.

The students played the game enthusiastically. The atmosphere of the session was only much different compared to the sessions with the levee patrollers. The session with the high school students was very chaotic and the students often interacted with each other during the training, in stark contrast with the levee patrollers, who were completely focused on the screen and their own game experience. They hardly shared experiences. Once I asked employees who work together if they communicated with each other.

*Casper:* Did you speak to each other about the game?

*DA3-#1:* Oh, were we allowed then?

*DA3-#2:* Well, often we did ask about each other's performance but never about the content, such as where the failures are.

This type of sharing happened continuously with the high school students. They further posed many questions, one after another. One high school student almost asked more questions than all of the patrollers of a single session together. Many of these questions were about the content of the levee inspection, because the controls of the game were of no issue to them. With the patrollers their questions were mostly about the controls. The students were surprisingly eager to do well. One student said "I am reading everything," while expressing a genuine interest in the material.

Because of the success of that day the teacher invited me to give the same adjusted start-meeting with two of his other classes. These classes consisted of 33 students ( $M = 16.1$  yr,  $SD = 0.84$  yr) who attended a variant just below the pre-university level (in Dutch, *Hoger Algemeen Voortgezet Onderwijs*; literally 'higher general continued education'). Here I made use of the computers at the school and, unfortunately, technical issues with the school's computer system impaired the study and interfered with data logging.

At the end of the sessions I asked whether students were interested in pursuing the complete training and out of the 21 pre-university level students seven indicated they were. With the other groups not a single person was interested. All seven attended the end-meeting which also took place at the school. This time I brought my own equipment with me. Of the seven student-participants, only one completed



everything. Two others played four out of six exercises. A fourth participant played only one exercise. The remaining three student-participants did not play at all.

Similar to my other attempts, the game-based training failed. The students are, of course, not the game's target group. The game is too specific to have any clear relevance for the students and not fun enough to sustain their interest for three weeks. Despite the failure of the training, it did offer some insight into how computer and game literates work with the game, and I retrieved 54 pre-questionnaires to assist comparing the students with the patrollers.

### ***Students Get Better Scores***

Only an hour was set aside to play the game and this is what most patrollers needed to play the start-exercise. Within this time the students played the training, start-exercise, and 12 out of 21 students played even more than the start-exercise. Because most did not finish those other exercises I will focus on just how they played the start-exercise and compare this to the levee patrollers. I ended up with a sample of 19 students. Two never played the start-exercise, because they played Exercise 1 instead.

Let us start with the scores. Here we see that the students ( $M = 76\%$ ,  $SD = 13\%$ ) outperform the 137 patrollers ( $M = 50\%$ ,  $SD = 32\%$ ) at length,  $t(56) = 6.75$ ,  $p < .001$ ,  $r = .67$ . The variety among patrollers is also greater than among the students, most likely because the computer (and game) skills are more equal between the students.

The difference in scores is especially caused by the levee patrollers not finding the failures. All students reported both failures, whereas 32% of the patrollers did not find the pitching stone failure in this level and 38% the boiling ditch failure. In reporting failures the students ( $M = 1.21$  non-failures,  $SD = 1.27$ ) seem to report on average more non-failures than the patrollers ( $M = 0.78$  non-failures,  $SD = 0.92$ ), highlighting a trial-and-error approach, but this was not significant.

In dealing with the failures that were found, no differences exist between the two groups, on an overall level per failure and for virtually every criterion or aspect. Only for assessing the severity of the pitching stone failure,  $t(104) = 2.22$ ,  $p = .029$ ,  $r = .21$ , and reporting the signal of the boiling ditch failure,  $t(81) = 2.04$ ,  $p = .045$ ,  $r = .22$ , the students performed (somewhat) better. A probable reason why the students outperformed the patrollers in these areas is that the patrollers had to unlearn their original ideas. The students, however, were starting with a blank slate and in a position to be immediately receptive to the best answer according to the game. They also passed answers to each other, which probably explains why, for example, everyone chose the same (and correct) answer for reporting the boiling ditch signal. Their collaborative learning approach should be kept in mind because it likely confounded the comparison.

The non-difference confirms a number of earlier hunches. First, it shows that patrollers had to learn how to play the game. The students were more literate and

thus needed less time to “read” the game. Second, even students were able to report the failures appropriately. This confirms that either dealing with the failures is not too difficult and/or the reporting procedure is supportive in dealing with them.

At the very end—with the end-exercise—the gap is closing between the students ( $M = 77\%$ ,  $SD = 8.5\%$ ;  $N = 7$ ) and patrollers ( $M = 61\%$ ,  $SD = 21\%$ ;  $N = 125$ ), but the students still outperformed them clearly,  $t(11) = 4.19$ ,  $p = .002$ ,  $r = .79$ . Like with the start-exercise the students found all failures (five in total); only 63% of the patrollers were able to find all of them. Three out of seven students played a good number of exercises at home and so it shows that even with little practice, the students were able to do well on the most difficult level.

However, the students did not outperform the 63% patrollers who did find all failures. Finding failures is crucial in terms of getting a good score. Without finding one, the players miss all points associated with one failure. The sample population of the patrollers is much more diverse and these results show that not being computer literate and game literate seem to have an influence until the very end.

### *Student Perceptions*

Before playing the game, the students filled out an adjusted pre-questionnaire. The perception statements (Level 7) and sensemaking test (Level 8) remained the same; the background questions were different. Some questions were left out, such as the one about organizational commitment, and others were added. I asked for example if the students had heard of inspecting levees before. Of the 54 students in total 20% indicated to have heard of levee inspection before they played the game.

The idea behind using the questionnaire was to use the students as a benchmark. The questionnaire had never been validated and although the answers of the patrollers did give us an idea of their perceptions regarding their computer skills, games, and inspection, interpretation becomes better if we can compare the results with a group who has never been involved with levee inspection at all. The variables that were considered for comparison are those from Figure 7.1 and are listed in Table 10.1 as well.

Just as expected, the students perceived to have more computer skills and also play more digital games. Only two students thought they were somewhat proficient; none thought they were not proficient at all. More than half (67%) considered themselves proficient or even very proficient. With the patrollers much less participants (31%) considered themselves to have such skills.

In terms of playing digital games the expectations are also confirmed. Compared to students levee patrollers played much less games. A whopping 46% of the students indicated to play a digital game every day. That is a big difference with the 5% patrollers who do so. Only 6% of the students said to play these games rarely or a couple of times per year. With the patrollers this concerned the majority (68%). If we look at the genre of First-Person-Shooters (FPS) specifically, we find that ex-

actly half of the students play this regularly. Just five students (9%) never played this genre. With the patrollers a large majority (77%) never played such a game.

These differences do not extend to playing analog games. We see here that students tend to play them more often—only 30% said to play them rarely compared to 40% of patrollers—but this is not large or strong enough to speak of a difference. The differences do extend to the components, because students have more of a positive game attitude than patrollers: they find games more fun and believe to learn more from them. They have an equal belief in the usage of games for serious purposes, such as training or education.

**Table 10.1** The comparison results between students ( $N = 54$ ) and levee patrollers ( $N = 136$ ) using Mann-Whitney tests for items and independent t-tests for components

Variable	<i>U</i>	<i>t</i>	<i>df</i>	<i>p</i>	<i>r</i>
Computer skills	2049			<.001	.37
Playing analog games	3326			.069	.13
Playing digital games	1013			<.001	.59
Game attitude		7.67	191	<.001	.49
Success potential		-8.50	73	<.001	.71
Pre-knowledge perception		-7.07	182	<.001	.46

*Note.* Due to some missing data the sample sizes were sometimes less.

Although the students may have a more positive game attitude, their success potential is lower. They expected the game to be less fun and learn less from it compared to the patrollers. Understandably, they lacked the motivation to learn about levee inspection.

Understandably as well, the students perceived to know less about levee inspection than the patrollers. This means that the patrollers did not perceive themselves as complete novices. Even after excluding the expert employees the results remained more or less the same.

With only seven students, we cannot make any firm statements, but what is rather interesting is that no differences can be found in how they judged the game and about their knowledge perception after playing the game. The seven students perceived to know just as much about levee inspection as the actual patrollers. The results on the sensemaking test will show that the students are not so wrong about perceiving to have this knowledge.

Before I will detail the sensemaking test, i will conclude this discussion by saying that the questionnaire results have strengthened our earlier findings in seeing the patrollers as not so computer and game literate. They are nevertheless motivated to learn and already have some initial knowledge (or at least perceive to have this). That is what the comparison with the students makes clear.

## ***Student Sensemaking***

I also asked the students to perform the sensemaking test, to compare their performance with that of the patrollers. The students have no background in inspection, making them a perfect benchmark. Of course, students had some difficulty filling out some of the questions, in particular regarding reporting, diagnosing, and what measures to take. They were forced to rely on common knowledge and common sense. Many provided silly answers, but it was apparent that some were able to make educated guesses (or had some very basic familiarity with levees.) These are some of the silly answers:

- That you should not get wet.
- That you have to get out of there!
- That you have to dial the right number to report your findings.
- Shotgun (about what measure to take).

I will now elaborate what I found by benchmarking the patrollers with the students. Although I looked into a number of specific points of interest, such as the non-failures, with this benchmarking I especially focused on the learning objectives and the overall scores.

### **Students use almost half the words**

It turns out that the 51 remaining students wrote much less. This is consistent among the two sets. For the core Set A, the students used 41% less words compared to the patrollers and for core Set B the results are with 48% less words about the same. If we put the results of Set A and B together, it becomes a difference of 43% and one that is a strong,  $t(148) = 6.56, p < .001, r = .47$ .

If we consider the difficulty for students to answer some of the questions and that they very likely have less motivation to fill this out, this difference could have been expected. What it further suggests is that the patrollers have much more to tell and this could be an indicator that they were knowledgeable from the start—at least more so than the students.

If we are more strict and only consider students and patrollers who filled out at least four complete failures out of five from the core sets, the word count difference between students and patrollers becomes with 37% somewhat less (Table 10.3). About 32 students remain after this selection, because a good many students thought more than one failure was not a failure.

### **They are far more laypeople**

The amount of words used by students and patrollers indicate that patrollers might be already knowledgeable; by looking at the accuracy scores their knowledgeable-ness is confirmed (Table 10.2). Except for assessing, the patrollers outperformed the

students on all learning objectives and the total scores. What is remarkable is that the difference is constantly about 13–15%. The patrollers have the same (slight) advantage on all objectives—except for assessing.

Regarding assessing the students perform still somewhat better than one would achieve with random guessing. The fact that their performance cannot be separated from the patrollers tells us that overall patrollers were not any better in assessing a situation compared to the average student.

The results also show the importance of training. With no background in levee inspection a person can get far, but we cannot expect this person to make sense of failures very well.

### **Even compared to volunteers**

The group of patrollers consists of both employees and volunteers. The employees could be further subdivided into regular employees and expert employees. This begs the question how the students compare to the volunteers. The volunteers outperformed the students on everything, except for—of course—assessing. However, we should consider that the patrollers wanted to do well on this test and (most) students did not care so much. Even accounting for their lack of motivation, I believe this discrepancy would still hold.

### **No exceptions except for the non-failures**

In terms of performance on the individual pictures, it turns out that we cannot speak of any difference with the non-failures. With the sheep non-failure (Real 5A), 85% students marked correctly as a non-failure compared to 83% patrollers. With the parked car non-failure (Virtual 5B) the numbers are 79% students who correctly saw this as no problem compared to 73% of the patrollers. This confirms my earlier hunch that it was relatively easy to identify the non-failures. Even students are able to do well on this!

The students' performance on all other pictures compared to those of the patrollers are listed in Table 10.3. Here we see that the patrollers did far better on every picture, except for the two non-failure pictures and two with which they had trouble with. The troubled pictures are the virtual illegal driveway (Virtual 4B) and the real watery slope failure (Real 6A). I noticed patrollers had a hard time interpreting both pictures. This explains why no significant difference is seen with these two.

With the illegal driveway in general—whether real or virtual—the majority of students (71–72%) did not see a failure. With the patrollers roughly half of them did not see a failure. What is interesting is that a number of students made the same strange interpretation: they too thought the levee was settled. Others indicated a not so strong road or spoke of “mud” that does not belong there. Not one suggested that driveway should not belong at the levee or that it damages the revetment.

**Table 10.2** Benchmarking the pre-test results with the students ( $N = 51$ ) and the post-test result with the experts ( $N = 14$ )

Picture	Students			Patrollers		Experts		
	<i>n</i>	<i>M(SD)</i>	%	Pre, %	Post, %	<i>n</i>	<i>M(SD)</i>	%
Assessing	51	1.53(1.06)	31	37	56	14	2.36(1.01)	47
Observing								
Accuracy	34	2.47(1.89)	16	31***	59***	14	5.93(1.00)	40
Words	34	31.70(21.1)		+28*	-84*	14	57.4(39.0)	
Reporting								
Accuracy	34	1.50(1.76)	10	24***	34	14	6.14(1.23)	41
Words	33	28.88(16.8)		+42***	-122*	14	96.3(69.9)	
Diagnosing								
Accuracy	34	0.00(0.00)	0.0	13***	54	14	8.29(1.49)	55
Words	33	11.5(12.4)		+37**	-82*	14	17.9(11.7)	
Core								
Accuracy	34	5.65(3.48)	11	24***	50	14	22.7(3.41)	45
Words	32	71.1(37.3)		+37***	-103*	14	172(110)	
Total								
Accuracy	51	5.08(4.01)	8.3	22***	45	14	6.66(2.4)	47
Words	41	75.9(41.7)		+42***	-125**	14	220(149)	

\*\*\*  $p < .001$ ; \*\*  $p < .01$ ; \*  $p < .05$  (two-sided).

The greatest difference is seen with the sand boils pictures. This is a well-known phenomenon in the field of levee inspection. For those not in the field it may seem uninteresting and not so dangerous. They considered it a “brown mess” or mud—something you may encounter in any ditch you look at. In fact, with the virtual one (Virtual 2B) half of the students even considered it a non-failure.

Their best performance was with the virtual stone damage (Virtual 1A). This was one of the clearer and least ambiguous pictures. Anybody could observe that stones are missing. They were furthermore able to guess a number of reporting items correctly, such as the size of the damage and the looseness of the surrounding stones.

What this especially tells us is that the patrollers did not outperform on one or two pictures specifically. With a few exceptions, they outperformed the students on every picture.

### Same post performance

Seven students completed the pre- as well as post-test. To reiterate, only one completed everything; two played four exercises at home; one just one; and the three remaining students nothing. Although they have not played it much, their performance on the post-test is above average. In fact, it is equal to that of the patrollers.

Whereas the patrollers obtained an average score of 50% at the end, the students scored 47%.

One of the student performances on the post-test is an extreme outlier, with a score of 14%. Excluding him, the remaining six students received an average score for the core set of failure of 53%. This is not significantly more, but it tells us that students might be able to get more out of the game with less effort. What we can definitely conclude is that with just a number of games students are up to par with patrollers.

But playing more does seem to pay off. The student who played everything (#154) got the highest score and was closely followed by the two students who played four exercises.

## Sensemaking by Super Experts

The students were used to benchmark the patrollers on one side, to gauge to what extent they are novices. We have come to find out that patrollers perceive to know more and seem to actually also know more. Ideally there would be a benchmark on the other side as well, to see to what extent the patrollers can be considered experts in levee inspection—before and after the training. To determine this and to assess the validity of the sensemaking test, I approached an existing expert committee on levee inspection to fill out the sensemaking test as well.

This expert committee consists of 14 members. Nine have worked at Deltares and helped with the development of *Levee Patroller*. They were therefore already familiar with the game and its content. As a matter of fact, the chair of the committee even initiated the development of the game. The other members were affiliated with other institutes. I agreed with the committee chair that he would send an invitation to the other members and that I would present the results on their next meeting. In addition, those who participated were promised a gift certificate of 12.50 euros.

### *A Super Expert Counts for Two*

Seven experts eventually participated. These seven participants filled out the 14 pictures all at once and to compare the results with the patrollers who filled out seven picture before and seven after the training, I made 14 cases ( $N = 14$ ), two cases by each expert. A super expert counts for two.

The sequence of the pictures was unchanged. They first filled out Set A and then continued with Set B (Level 8). As we will soon see, this sequence did affect their performance and is something we need to keep in mind.

The results are highlighted in Table 10.2 and Table 10.3. From these results we can deduce a number of insights.

**Table 10.3** Benchmarking the pre-test results with the students ( $N = 51$ ) and the post-test result with the experts ( $N = 14$ )

Picture	Students			Patrollers		Experts		
	<i>n</i>	<i>M(SD)</i>	%	Pre, %	Post, %	<i>n</i>	<i>M(SD)</i>	%
Virtual 1A	26	2.15(1.12)	22	33***	67*	7	4.86(2.91)	49
Real 1B	23	0.87(0.63)	8.7	18***	55*	7	3.86(1.21)	39
Real 2A	21	1.29(0.96)	13	38***	62	7	6.57(0.79)	66
Virtual 2B	13	0.92(2.0)	9.2	31***	55*	7	6.86(0.90)	69
Virtual 3A	26	1.31(1.26)	13	25***	50	7	5.29(0.76)	53
Real 3B	23	1.13(1.22)	11	27***	51	7	5.71(1.11)	57
Real 4A	8	0.25(0.46)	2.5	10***	39**	4	3.00(0.82)	30
Virtual 4B	6	0.67(1.21)	6.7	9	34	2	4.00(0.00)	40
Real 5A	27	8.67(3.26)	87	88	88	7	9.14(2.27)	91
Virtual 5B	24	8.00(3.99)	80	79	84	7	9.43(1.51)	94
Real 6A	27	1.52(1.01)	15	22*	39	7	4.57(2.15)	46
Virtual 6B	24	1.46(0.98)	15	26***	42	7	4.86(1.68)	49
Virtual 7A	17	1.94(0.83)	19	27***	39	5	6.00(2.12)	60
Real 7B	8	0.13(0.35)	1.3	10***	19**	7	5.14(1.35)	51

\*\*\*  $p < .001$ ; \*\*  $p < .01$ ; \*  $p < .05$  (two-sided)

### Experts use more than twice as many words

We have seen that the students use almost half the words compared to the patrollers. With the experts it is much the reverse (Table 10.2). They use more than twice as many words. This time the difference is much larger with the core Set A (124%) than with core Set B (87%). This can be explained. The experts filled out Set A first and then continued with Set B. At that point they might felt that they were repeating themselves. In addition, they could felt less inclined to elaborate on their answers. Whatever the reason, their responses are with an average of 103% still considerable more,  $t(13) = -2.92$   $p = .012$ ,  $r = .62$ .

Again, if we consider the amount of words as an indicator of knowledge, it shows that the experts are indeed more knowledgeable. If we compare this with the amount of words used by the patrollers before the training, then the experts use about 52% more words. This is the opposite of the students!

### Afterward patrollers perform equal to super experts

Although the experts used far more words, in terms of accuracy their performance is similar to that of the patrollers. That is a major improvement, because if we compare the super expert results with the pre-test results by the patrollers, the super experts outperform them on each and every area. This means that the training helped the patrollers reach the level of experts, perhaps beyond.



With assessing the patrollers tend to outperform the super experts. With observing the patrollers performed definitely better. If we consider the amount of words used, we can continue to argue that patrollers are even better: they are concise and accurate. This combination is of much value in a crisis situation.

However, if the super experts would have played the game—even if just for a little bit—I would expect them to get much better results. They would then have a better understanding of what is expected from them and they would have learned how to use the vocabulary of the game.

### **Employees are better than super experts**

With an average of 56% on the core set of pictures the employees outperform the super experts,  $t(45) = 3.45$ ,  $p = .001$ ,  $r = .46$ . If we consider all pictures, no difference is to be seen. Before the training employees had an average of 29% and this shows that if we consider pictures that are closely related to what was taught with the game, employees can even get better than super experts.

The employees are to be subdivided into regular and expert employees, because not every employee is fully preoccupied with levee inspection and we know that their performance on the pre-test differs. Considering this distinction, the super experts ( $M = 45\%$ ) still perform better than the expert employees ( $M = 38\%$ ), but not by large,  $t(19) = 2.14$ ,  $p = .045$ ,  $r = .43$ . After the training, no distinction exists between the employees and expert employees.

### **Good at standards and not so good in non-standards**

The mere fact that if we consider all pictures the difference between the employees and super experts diminishes to nothing, hints to a possible performance gap with the two excluded pictures: the non-failures and the new failures. This is not caused by the non-failures (Table 10.3). It is the new failures. The patrollers improved only slightly on these and did not get close to the level of the experts.

This confirms that overall the super experts remain the experts. Only in the areas the patrollers were trained at they were able to outperform the experts and on that very moment they had to take the post-test.

When we continue to look at the other pictures, it becomes clear that the patrollers outperform the experts on the stone damage pictures (Virtual 1A and Real 1B) and the experts the patrollers on the boiling ditch pictures (Real 2A and especially Virtual 2B). Earlier, I described the striking results—with patrollers but also with the students—regarding the stone damage pictures. By comparing these results to the experts it becomes clear that playing the game helped the players so much that they specifically outperform experts on this specific failure.

The good performance on the boiling ditch failure does not come as a surprise, because this is a well-known phenomenon and experts should know every bit about it. What is interesting is that they especially performed better on the virtual picture,

the one which the patrollers could have practiced with on a number of occasions. This actually tell us that for this specific failure the game could be improved. It does not get its players up to the level of experts.

Another picture where the patrollers outshine the experts is the real illegal driveway (Real 4A). A number of experts thought this was not a failure. During the discussion it became clear why. They upheld a very strict definition. Only when overtopping would occur or some other signal, they would consider it a failure. The picture portrays a potential failure, not an actual one.

The remaining experts performed remarkably poorly with both illegal driveway pictures. One explanation is that this failure was added after a road trip in collecting pictures for the game. It was a spontaneous addition by the design team, because the failure situation was not a commonly shown standard one. The experts were thus likely unfamiliar with it. This tells us that super experts excel in standard failures and fare worse in non-standard failures.

### ***Putting Results into a Perspective***

These results validate the sensemaking test. We would expect super experts to perform better than students and we would hope that the patrollers get to the level of experts by means of the game-based training. This is what exactly happened. The test is sensitive to the knowledge levels.

Some further scrutiny is however needed. Although the test might be sensitive to the expected knowledge levels, if the experts receive on average a 45% score this requires some elaboration. We need to put this test and its results into a perspective to understand the results of the experts.

Then I had a discussion with the experts about their results on assessing specifically. From this discussion it became clear that we need to put the results into a perspective as well.

### **A validated yet merciless test**

An average score of 45% for a super expert makes the test maybe somewhat questionable. We would expect the super experts to get a 80% score or higher, if not a 100% score. One answer is that super experts are super experts in a specific area: They are specialized in one or two failures. One expert might do well with sand boils and another with a watery slope.

A second answer is that the test is like the game black-and-white. Often and especially with assessing and diagnosing only one answer was considered correct, whereas we could have endless discussions about what else could be considered correct too. The answers are not set in stone and so the super experts (and maybe also the students and patrollers) might have done better than portrayed here.

The test is also merciless. If one makes an error, a participant loses many points. For example, by incorrectly assuming that a picture is a non-failure, this will decrease the total score with 20%. Some of the super experts mistakenly considered the illegal driveway as a non-failure and this had a major impact on their score.

Finally, experts had a major tendency to describe failures with generic words, such as fault, slip, or deformation. Such use is punished, because one needs to be specific. Per instance they were not specific enough it cost them 4 or 6% on their total score. When I informed the chair of committee about this, he told me this by e-mail:

...the word fault (in Dutch, *afschuiving*) is often used in exchange for macro-instability, even if strictly speaking a fault could belong to another failure mechanism. It is ambiguous jargon—personal communication, July 20, 2010

The super experts use ambiguous jargon. If the experts do not get it straight, how can we possibly expect the patrollers to get it straight?

### More complete and detailed

The super experts used many words and could have been more concise, like the patrollers were in the end. However, the information they provided was certainly useful. It just was not the type of information that was scored.

Waterside is partially missing on the “foreshore and outer slope” but also on the crest. Because of this the profile of the waterside is insufficient. Barely any crest width exists. This is probably caused some time ago by a wash—#203/210

This example belongs to Real 7B and the question “What do you see?” In essence his answer is “a missing part of the waterside.” That is not incorrect, but it is a descriptive answer. So despite all the information, this super expert’s answer was categorized as “1 = slightly accurate (SA).” Everything else is an elaboration on that answer—on the causation and consequences.

Like I explained in Level 8 some of the patrollers provided such information too, but it happened sparsely. Only with the very first picture on the very first test this was really noticeable. The experts, on the other hand, provided elaborations throughout. The above-mentioned example is a description of the last picture he had to judge.

Something important which I decided to neglect concerns the mentioning of the contributing signals. Three pictures contained a contributing signal (Virtual3A, Real 3B, and Real 1B). I neglected this because I could not observe any difference with the pre- and post-test. The experts seem to notice the contributing signals better, at least with one picture in particular, the real landslide. They mentioned its contributing signal liquefaction more,  $U = 133$ ,  $p = .019$ ,  $r = .27$ .

All this gives the impression that the expert’s answers were more complete and detailed. It may further mean that the experts’ answers are under appreciated with the current scoring system. This should be kept in mind in making a distinction between the patrollers and the experts.

### **No agreement whatsoever**

On December 13, 2010 the expert committee came together. One agenda item was a discussion of their sensemaking test results. On not one picture did the experts agree with each other.

Of course, with some pictures much more agreement existed than with others. Six out of seven experts recognized the non-failures and with Virtual 3A, Real 3B, Virtual 4B, and Virtual 6B five out of seven agreed about a certain severity. Choices were spread out with all other pictures. The experts were flabbergasted by these results. They assumed that they thought much more alike.

This resulted in a heavy discussion and the experts explained how they looked at the failures. One said in response to a comment about how dangerous the sand boils failure is:

You think this is a critical? You have never been out there probably. Once, back in 1994, I saw this everywhere. None of them caused a problem. I would say this is something to pay attention to and certainly not critical.

Although I think more discussion would help to align the experts, complete alignment seems impossible. Different interpretations and experiences—such as by that one expert with the sand boils—shape how people make sense of something and if that something is ambiguous and complex, we should expect different responses.

## **Interviewing the Levee Patrollers**

Before and after the training I interviewed a number of participants at each organization. I had two purposes with these interviews. The first was to get to know the patrollers better. The acquired information I used throughout this book. This constituted the first part of the interview and there I asked questions about what they do beyond levee inspection and how they got involved with it (see Interview Protocol; Questions 1–6). This was done to establish a better understanding of who these patrollers are. This part of the interview was semi-structured. I prepared some questions up front, but the conversations went into various directions.

The second purpose was to validate the outcomes of the training—which is the overall purpose of this level. This constituted the second part of the interview and all other questions were devoted to this (see Interview Protocol; Questions 7–15). This part of the interview was structured. It had to be, because this was more or less an oral examination. Structure was necessary to make the results comparable.

### **In-depth explanation: finding interviewees and analyzing the results**

Interviewees were randomly selected from a participant list. I kept certain criteria in mind, such as recruiting an equal number of volunteers and employees from each organization. Because I was familiar with the participants after the training, I picked their names blindly.

If a participant was picked and fulfilled the criteria, I called this person to ask if he or she was willing to do an interview. On occasions participants declined and then I selected randomly another possible candidate.

I scheduled the interviews for an hour: 30 minutes for the first part about their personal life and view on the organization and another 30 minutes for the second part which tested their levee inspection knowledge orally. Most interviews finished sooner than scheduled, because participants did not have much to say about levee inspection. In addition, what they had to say I mostly heard before and so I skipped some of the questions. With the interviews that took longer than an hour, I talked about items that were not on the Interview Protocol. On average the interviews took 51 minutes.

The interviewees were recorded with a digital voice recorder and took place at either people's offices or homes. In fact, all employees were interviewed at their office and all volunteers at their home. No complete transcript was made of the interview. For my purposes this was not necessary. The interview data was not my primary source of data. I used the interview data to triangulate my data of who the patrollers are and see how they performed orally.

I further deemed it counterproductive to let interviewees review the transcript. Highly likely the interviewees would have a different opinion and I was interested in what they thought right there at that moment I was interviewing them.

The interview dates for the pre-interviews do not vary significantly. All interviewees were interviewed right before their start-meeting, the earliest possible time. By doing the pre-interviews after the start-meeting the results would have been biased. With the post-interviews the dates do vary, to be able to explore a possible effect of time on retention.

I used the following interview protocol:

### **Personal**

1. What do you do in your daily life?
2. Why did you become a levee patroller?

### **Organization**

3. Could you describe the procedure of inspecting levees at your organization?
4. What are your tasks and responsibilities during the inspection?
5. How well do you know the other levee patrollers?
6. Has much changed in the past years?

### **Failures**

7. Suppose you go on a levee inspection. What failures could you encounter during the inspection?
8. Suppose you find failure X. (Here I randomly chose one failure from the list of failures mentioned by the respondent at the previous question and repeated this a number of times)
  - a. What will you see?
  - b. To what do you need to pay attention to when reporting this failure?
  - c. How could this failure develop over time?
  - d. What would you report to the Action Center?
9. Have you ever encountered one of these failures?

### **Failure mechanisms**

10. What is according to you a failure mechanism?
11. What failure mechanisms exist?

12. Could you draw for me a *a)* erosion outer slope, *b)* erosion innerslope, *c)* macro-instability, *d)* sand boils, *e)* micro-instability.

### Pictures

13. Suppose you find this failure (Fig. 10.4). (I first showed the real picture and subsequently the virtual one)
- What do you see?
  - To what do you need to pay attention to when reporting this failure?
  - How do you assess the situation?
  - What possible failure mechanism is causing this?
  - What measures could be taken to prevent the situation from becoming worse?

The structured oral examination part of the interview consisted of three parts:

1. *A knowledge elicitation test relating to failures:* I first asked the participants what failures they could encounter when inspecting a levee. Then I asked them to explain how they would deal with one or more of these failures.
2. *Drawing and defining failure mechanisms:* After asking interviewees to define what a failure mechanism is and what types exist, I asked them to draw the failure mechanisms to the best of their ability.
3. *Oral sensemaking test:* Similar to the sensemaking test on the start- and end-meeting, except oral. I showed a picture of a real and a virtual failure and asked the interviewees the exact same questions.

I interviewed 20 people. At each organization I interviewed four participants, two volunteers and two employees, before the training. My original attempt was to do the same at the end. Because I had been so much in touch with the participants and had various informal talks, the need to get to know the participants lessened. Two or three individuals were considered sufficient for validation. Therefore, after the training I interviewed three participants at Organizations A and C and two at Organization B.

When reading the results, please bear in mind that the two post-interviewees (IPpost-#91 & IPpost-#116) at Organization B were interviewed almost a year after the end-meeting. Then one was interviewed about one month after (IPpost-#4), two were interviewed about two weeks after (IPpost-#136 and IPpost-#147), and three about three months after (IPpost-#9, IPpost-#27, and IPpost-#123).

### *“It Seems Like an Exam”*

When I started the structured second part of the interview occasionally a participant shouted “It seems like an exam!”, which is more or less accurate. The first item of this exam concerned a knowledge elicitation test. I started this test with these lines:

Imagine you have to go on a levee inspection. What possible failure signals could you encounter?

After they mentioned what failure signals, I summarized what they told me and asked if there was anything to add. I repeated this procedure until the participants said that they could not think of anything else. I noticed a tendency of participants to answer in accord with their own situation. They started by listing what failures could happen to the levees that they are supposed to inspect. This is a logical response. Considering the hypothetical situation, one would consider what to do with regards to the organization he or she is working for. Because I wanted to hear about all possible failures in all possible situations, I stressed that the participant had to think of any situation—not just the one they expect to encounter at their own organization.

#### **In-depth explanation: categorizing the elicited failure signals**

I listed every failure signal the participant came up with. I did not judge these at this point. Then I categorized the answers by using the twelve signal categorization system from the game. I was able to classify the responses without any issues. Every response fit within the system. One important additional rule I used was to classify failure mechanisms with a signal that they are most clearly associated with. Thus, I associated the mentioning of sand boils with water outflow, erosion inner slope with grass revetment, and erosion outer slope with pitching stone.

The translation resulted in fewer signals per interviewee. Interviewees mentioned not infrequently names of signals that closely relate to one another. For example, they mentioned seepage, sand boils, and water outflow. Or they gave examples of signals that could be easily grouped together, such as mole corridors and rats. I categorized both cases as one signal, as water outflow and animal activity, respectively.

The numbers are, however, not drastically different. Without categorization interviewees mentioned 5.50 signals before and 7.25 after. This is slightly more than the averages after categorization (Table 10.4).

After eliciting the failure signals I randomly chose two or three signals and asked the interviewee what he or she would report upon finding this signal. So with the knowledge elicitation test I looked into observing as well as reporting.

#### **A richer signal repertoire**

Table 10.4 shows how many interviewees listed the twelve signals. This table makes clear that in terms of percentages the most frequently mentioned signals remain largely the same. Playing the game did not seem to have an impact on this. It did influence the variety of mentioned signals and the frequency of some of the lesser known signals. Some signals that were not mentioned before, were mentioned after and most notably the pitching stone signal. In fact, this turned out to be one of the most mentioned signals.

Some signals, such as human activity and floating waste, are mentioned much more after the training. With human activity this is less surprising, because one of

the included failures concerns a human activity. Floating waste, on the other hand, did not make any appearance.

**Table 10.4** The elicited failure signals of the pre-interviewees ( $n = 12$ ) and the post-interviewees ( $n = 8$ )

Signal	Pre				Post			
	A	B	C	Total, $n(\%)$	A	B	C	Total, $n(\%)$
Water outflow	3	4	3	10(83)	2	2	3	7(88)
Crack	2	2	4	8(67)	3	1	2	6(75)
Settlement	2	4	4	10(83)	3	1	3	7(88)
Human activity	2	1	1	4(33)	1	1	3	5(63)
Overtopping	3	3	1	7(58)	2	1	1	4(50)
Grass revetment	3	2	3	7(58)	1	2	3	6(75)
Animal activity	0	2	2	5(33)	1	2	0	3(38)
Liquefaction	0	1	2	3(25)	0	1	2	3(38)
Floating waste	0	0	1	1(8)	0	1	2	3(38)
Bulge	0	0	0	0(0)	1	0	1	2(25)
Pitching stone	1	1	1	4(33)	1	2	3	6(75)
Horizontal movement	1	0	0	1(8)	0	0	1	1(13)
<i>M</i>	4.25	5.00	5.50	4.92	5.00	6.50	8.33	6.63
<i>SD</i>	0.96	2.16	2.08	1.73	1.00	0.71	0.58	1.14

What we learn from the knowledge elicitation results is that participants have a much richer signal repertoire. On average participants are able to mention close to two more signals. This is also a lasting effect. The two participants at Organization B were able to get this higher average after a year and both of them were volunteers.

In addition to the results of Table 10.4, I looked into vocabulary usage. I assessed how closely the language used resembles that of the game. It turns out that on average the pre-interviewees used exactly two similar words and the post-interviewees used about four (4.25 to be precise). That is double.

It was furthermore harder as interviewer to understand what the pre-interviewees were talking about and how to categorize their answers as a coder, because their answers were more ambiguous and less articulated. For example, one of the interviewees spoke of a “hole in the levee” (IPpre-#5) and another about a “moved levee” (IPpre-#11). It took some questioning and puzzling to understand that they were talking about a pitching stone and horizontal movement, respectively.

### Richer reporting repertoire and more correct

Only four signals had been questioned sufficiently to be considered for a comparison. I deemed it sufficient if more than two pre- and post-interviewees had been asked to tell me what they would do if they encounter a certain signal. The sample size is already little and with two or less per group any conclusions would be ar-



bitrary. The signals I eventually considered are settlement, seepage (or sand boils), erosion (grass revetment or pitching stone), and crack.

In comparing the results I looked at the number of reporting items mentioned and their correctness (Table 10.4). A reporting item is anything that the interviewee mentioned. I did not judge or converted the responses. Typical responses are the size of the damage (length, width, and/or depth), assessing the severity, and reporting the location. Each is one reporting item. If an interviewee would have mentioned to report a failure's size, severity, and location, he or she mentioned three reporting items.

**Table 10.5** The reported items and their correctness by pre-and post-interviewees

Signal	Pre					Post				
	<i>n</i>	<i>M</i> <sub>items</sub>	<i>SD</i> <sub>items</sub>	<i>M</i> <sub>correct</sub>	<i>SD</i> <sub>correct</sub>	<i>n</i>	<i>M</i> <sub>items</sub>	<i>SD</i> <sub>items</sub>	<i>M</i> <sub>correct</sub>	<i>SD</i> <sub>correct</sub>
Water outflow (6)	9	2.78	1.09	1.33	0.87	5	4.50	1.76	2.33	1.03
Crack (4)	8	3.63	1.19	1.00	0.53	4	3.00	1.15	1.25	0.50
Settlement (3)	6	2.83	1.17	0.83	0.75	5	4.60	1.34	1.40	0.55
Erosion (3)	5	2.20	0.84	0.60	0.55	4	3.50	1.29	2.25	0.96

Judging the correctness of these reporting items follows the systematics applied with the sensemaking test (Level 8). This systematics involves giving one point for every item that is included in the game's reporting procedure. Because some items were rarely ever mentioned (such as the direction of a settlement) or were combined by others (such as length and width that were often referred to as size), some reporting items were grouped together. Table 10.5 lists the signals and the maximum points for each (between the parentheses).

The results highlight that here too the post-interviewees mentioned more. This time it concerned reporting items and not signals. Some of what the post-interviewees mentioned were not signal-specific items, such as the location, failure mechanism, and severity, but the pre-interviewees did this just as much. Additionally, the post-interviewees were more correct about what they mentioned. For the number of items as well as their correctness a similar pattern emerges: the post-interviewees performed about one-and-half times better (1.60 to 1.75 times to be precise).

Two exceptions exist. First, with erosion the post-interviewees performed better. It is much closer to four times the pre-interviewees' performance. One explanation is that the participants learned that soil flushing is an important indicator. This information stuck. Second, with the crack signal, hardly any improvement is to be seen and most certainly not with the number of reporting items mentioned. This decreased in fact. A reasonable explanation is the familiarity with the crack signal. Patrollers know that if they find a crack, they need to measure it.

The interview reporting results validate the outcomes of the sensemaking test. Here too the results are rather disappointing. I speculated in Level 8 whether the poor reporting results could be attributed to the method. This alternative method

shows no difference. Although it could be that elicitation—whether oral or written—is not an appropriate method to test this at all, this confirmation suggests that people have difficulty remembering items explicitly—unless they make a big difference. Soil flushing comes to mind as such a big change maker.

### **Patrollers were already knowledgeable**

Despite the disappointing reporting results I was generally impressed of the knowledge people exhibited throughout the knowledge elicitation test and in particular during the part in which they had to say what they would report. For example:

If you come across a seepage, you have to pay attention if it is just water or whether mud comes along. You also need to check if it grows. It becomes worse if the water level on the other side rises or if the flow through intensifies. You'll notice that sand comes along and a color change of what comes out of it. A crater formation more or less happens. Sand has to go somewhere—IPpre-#48

Another participant stated:

...how big is that crack? What is its length and width? This tells you if he goes or stays...if the levee becomes like this [shows this with his hands], then you see he wants to move. We have to put sand bags at the toe of the levee to prevent him from going—IPpre-#70

This confirms that the patrollers were already knowledgeable. Although they could not always articulate their knowledge very well, they clearly remembered something vaguely or had an idea of what is important. They had some sense. The following interview conversation illustrates this too.

*IPpre-#134:* You have dangerous and not dangerous seepages. Dangerous ones need to be contained with sand bags. The not dangerous ones are those close to for example tree roots and tree trunks.

*Casper:* You speak of dangerous and not dangerous ones. How can you separate one from the other? [the answer is if soil is flushing]

*IPpre-#134:* Hmm...[long break]...I thought it had something to do with gushing, but I have to look that up before I go on patrol.

Participant #134 had an intuitive understanding of what is important when encountering seepages and I noticed understanding among others too. They vaguely remembered something from a course or what they heard from others. Some interviewees could also endlessly talk about what they did know, something they experienced or heard of.

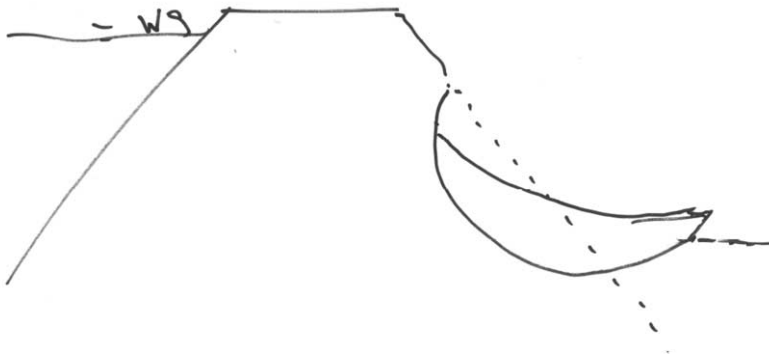
### ***Drawing Failure Mechanisms***

The next step during the interview involved drawing failure mechanisms. Inspired by cognitive mapping, brainstorming techniques, and drawing techniques in therapy,

I asked the participants to draw what image they have of a certain failure mechanism. The primary reason to include this was to force participants to make their thinking explicit. This helped them to explain me what their understanding is of macro-instability or sand boils. It helped me to ask them questions about their description.

As a matter of fact, many of the participants started drawing before I even asked them to, because visual representations helped supplement their verbal expression. When drawing I asked for clarification by pointing to the drawing and this enriched their description and my understanding.

A second reason was to investigate the pictures themselves. The game should give players various mental models of failures and I was curious to see how the game experience possibly enriched such models and what associations they had. This was not my primary purpose, because if firm conclusions were needed, I should have interviewed many more people. In addition, participants' drawing skills and creativity are of much influence in a method like this. Therefore, any results related to what was drawn should be approached with caution.



**Fig. 10.1** An example of a failure mechanism drawing, by Participant #133 (IPpre), who was confused about micro- and macro-instability

The procedure of this part of the interview was as follows. I first asked the definition of a failure mechanism. If they had some idea, I continued by asking what types exist. Then I mentioned the name of one of the five failure mechanisms as defined by us during the design of the game and requested the interviewee to draw his image of it. I proceeded with this until all five failure mechanisms were drawn.

### “Failure what?”

Based on the sensemaking test it became clear that participants were not so familiar with the term failure mechanism. Of the 12 pre-interviewees two made a brave attempt after some deliberation (IPpre-#5 and IPpre-#79), two gave a vague description and seemed somewhat unsure about their answer (IPpre-#90 and IPpre-#133), and only two gave a firm and correct answer (IPpre-#146 and IP-#149). Except for one, these answers came from employees. One employee and all other volunteers immediately said they had no idea. A typical response was “Failure what?” It seemed like they had never heard of it. This was a surprise to me. I had not analyzed the sensemaking test at that point in time, but the term is mentioned throughout the levee inspection course and I thought this was common knowledge for those involved with levee inspection.

One possible explanation for their unfamiliarity with the term is that patrollers seemed convinced that “those terms” were not going to be used in practice. They considered it terms thought of by highly educated gentlemen sitting behind their desks.

Mostly you do not use those terms. You just say I see this or I see that. If they want to put a label on this, I am fine with that, but I will tell them in this way. If they then say it is that term, I think “Oh yeah, that is true.” If it [a levee inspection] ever occurs—and I hope not—then I have forgotten all those terms—IPpre-#11

### In-depth explanation: judging failure mechanisms

Defining the failure mechanisms was easy to code. Most said they did not know and a few others were making a poor guess. These were grouped together as “incorrect.” I considered a definition correct if it clearly illustrated that it is something that causes a levee to fail and/or that what patrollers see on the levee are symptoms of the mechanism. The gray area of responses is when interviewees described in vague terms that damage appears because of the mechanism. A gray response example is:

This happens before the failure...I do not know...Something that turns into a failure later—IPpost-#116

Coding the drawing of failure mechanisms was harder. I used the drawings as well as the descriptions interviewees gave and I used the same coding scheme. An interviewee could be incorrect, somewhat incorrect, or correct. Interviewees did not have to fully explain the failure mechanism process to be considered correct. The game does not teach this. It was sufficient to explain what they would see and where on the levee this would occur.

The correctness was judged using failure mechanism descriptions and images from the game. The descriptions and drawings did not need to be identical to those of the game. Many alternative situations are imaginable. For example, erosion outer slope could also occur due to floating waste. If interviewees mentioned such an example—and some of them did—this was considered correct too.

If interviewees pointed into the right direction, but provided vague descriptions, were clearly unsure about it, or were partly erroneous, this was considered somewhat correct. This is an example of such a gray response with this task:

I connect this [macro-instability] with a settlement of the inner slope. Does water flow out? Does soil flush or not? You have to look around the ditch to see if that happens. If it is just a little bit, then it is micro-instability, but if you see cracks, then it is macro-instability—IPpost-#136

Participant #136 (IPpost) is somewhat correct, because he does mention signals affiliated with macro-instability as well as micro-instability. He just mixes and confuses the two and thought of an incorrect heuristic. In the game macro-instability as well as micro-instability have a crack.

Although after the training I did not get the typical “Failure what?!”-response, many interviewees told me that they forgot. Eventually, after some consideration, only one could not give me a definition. Another made a brave but poor attempt. The six remaining post-interviewees were split up in between those making a good attempt and those with a correct answer.

This shows that the post-interviewees have generally a better idea of what a failure mechanism is. Whereas the majority (64%) had no idea or was incorrect before the training, the majority of the post-interviewees (75%) was somewhat correct or fully correct. Even after almost a year the participants still had a notion of what it is about.

### “Drawing what?”

Although participants clearly were not familiar with the term “failure mechanism” before the training, when I mentioned each specific failure by name, they had a mental image and especially for the two erosion types and sand boils, because they were able to draw them. Sand boils seems the most familiar term.

However, the four interviewees from Organization A did not know this term and did not draw it. One was able to describe the mechanism, but confused it for erosion outer slope. In contrast, all other pre-interviewees were able to picture it and they did a fairly accurate job. This is not a coincidence. Organization A did not provide for a levee inspection course. Now the term failure mechanism may not have stuck, some of the material seem to have found its way into the minds of the people attending the course. This became clear in my conversation with Participant #146:

*Casper:* Was the story about failure mechanisms conveyed well during the course?

*IPpre-#146:* Yes, it was clear. And very insightful too with at least two experiments with a sandbox. That works for me. If you see something, it sticks better, not with what words its associated with but with what could possibly happen.

One could make an educated guess about a term, but doing so still requires an initial understanding. Such an educated guess did not work well with macro- and micro-instability, for which some interviewees made the false assumption that everything little would be considered a micro-instability, like this interviewee:

Is that not related to pitching stone? If a couple of them are missing, a small part of the levee is damaged—IPpost-#27

**Table 10.6** Performance on drawing failure mechanisms by the pre-interviewees and the post-interviewees

Signal	Pre				Post			
	A	B	C	Total, <i>n</i> (%)	A	B	C	Total, <i>n</i> (%)
Erosion outer slope								
Incorrect	1	0	0	1(8)	0	0	0	0(0)
Somewhat correct	0	0	2	2(17)	1	0	0	1(13)
Correct	3	4	2	9(75)	2	2	3	7(88)
Erosion inner slope								
Incorrect	1	0	1	2(17)	1	0	0	1(13)
Somewhat correct	1	1	1	3(25)	2	0	0	2(25)
Correct	2	3	2	7(58)	0	2	3	5(63)
Macro-instability								
Incorrect	3	2	1	6(50)	1	0	0	1(13)
Somewhat correct	0	2	2	4(33)	1	1	2	4(50)
Correct	1	0	1	2(17)	1	1	1	3(38)
Micro-instability								
Incorrect	3	3	2	8(67)	2	1	2	5(63)
Somewhat correct	1	1	2	4(33)	1	0	1	2(25)
Correct	0	0	0	0(0)	0	1	0	1(13)
Sand boils								
Incorrect	3	0	0	3(25)	0	0	0	0(0)
Somewhat correct	1	1	1	3(25)	2	1	0	3(38)
Correct	0	3	3	6(50)	1	1	3	5(63)

The error persisted after the training, which is one of the indications that the pre- and post-interviewees are not remarkably different from one another (Table 10.6). This suggests that the diagnosing scores on the pre-sensemaking test are likely an underestimate of the knowledge participants had. They may not have known the exact words, but they did have a better understanding than the test results suggest. Likely too is that the post-test results may have been an overestimate. Quickly after the training participants may have remembered the words well and were able to designate them appropriately. But the associations were weak and so after time they confused their meaning or simply forgot.

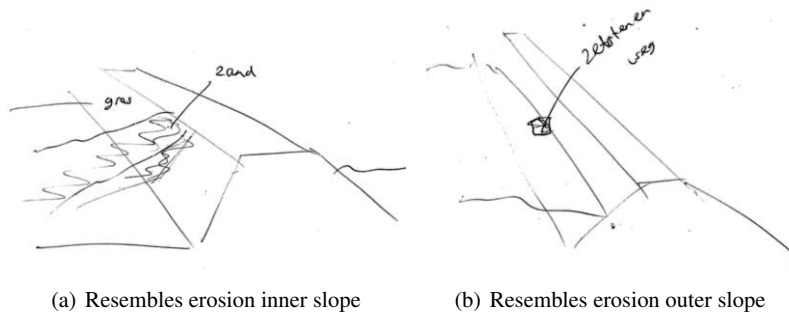
Although this performance by individuals contrasts with other findings, the performance results for each failure mechanism is consistent. Like elsewhere, performance on micro-instability is the lowest and on erosion outer slope the best.

### 3D drawings and a close resemblance

In investigating the drawings, there does not appear to be any striking difference at first sight. After the first failed attempt to detect any emerging patterns, I decided to look for differences in the perspectives used, the details in the drawing, and the

likeness of the drawings compared to the game. Again, the differences were not obvious, but I found a number of insights that I would like to share.

Almost everybody (16 out of 20) made at least one drawing with a crosscut of a levee (crosscut perspective; see Figure 10.1) and for half (10 out of 20) this was the only way of how they visualized the mechanisms. Using a crosscut is the standard way in which failure mechanisms are visualized in courses and textbooks. It is also used in the game.



**Fig. 10.2** Two failure mechanism drawing by Participant #123 (IPpost). The signals drawn are identical to the failure situations in the game and the drawings are in 3D perspective

Other perspectives that were used is to view the levee from on top (top-down perspective), from the side, facing only one of the levees slopes (side perspective), or in 3D (3D perspective). The top-down perspective and the side perspective were used by pre- and post-interviewee. The 3D perspective was only used by three post-interviewees and one pre-interviewee who had played the game before. One interviewee (IPpost-#123) made all his drawings in 3D. This was a young employee who had no experience with levee inspection at all. The game-based training was his first encounter with it. He was a blank slate and only had the game experience to answer the questions. His drawings accurately resemble those of the game and he did this about three months after he finished the training (Fig. 10.2).

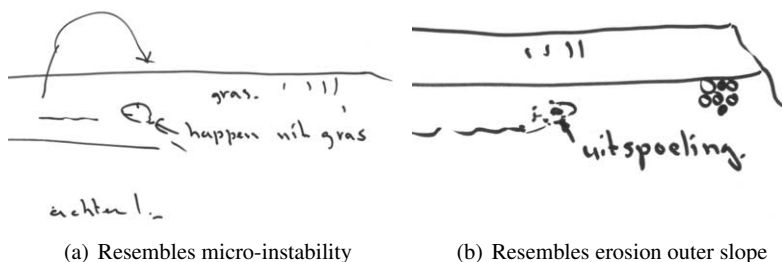
The drawings by the one pre-interviewee (IPpre-#11) with a 3D drawing resembled those of the game too. This individual had played the game two years ago. It seems like this game experience strongly influenced his drawings, since three out of five drawings are almost identical to the situations from the game (Fig. 10.3). I noticed this during the interview already and asked him if he thought the game contributed to his drawings:

Much stuck. Back then it was the first time. You immediately know what you have to pay attention to. This was really useful to me—IPpre-#11

Some made direct references to the game, such as the earlier mentioned Participant #123, who drew Figure 10.2(a).

That is kind of the same [as erosion outer slope]. Here is water. Here is the hinterland. And here you have something that I found a great example from the game. It is someone who made a road with sand. That is a human activity, right?—IPpost-#123

I suspect that such game referencing occurs when people have hardly anything else to build on and when the game provides for a clear reference. If people have other experiences to rely on, they integrate the game experiences within their existing knowledge.



**Fig. 10.3** Two failure mechanism drawing by Participant #11 (IPpre). The signals drawn are identical to the failure situations in the game

The drawings did not markedly differ in anything else. Playing the game did not result, for example, in more detailed drawings. Such detailed drawings could possibly indicate that players have a richer mental model of failures. This did not happen. But if we consider the game's graphics, then we have to conclude that these are not detailed either. The game has some details, such as windmills and other landscape objects, but the levees themselves are plain and simple. The lack of detail in the drawings might be result of this.

### *The Oral Sensemaking Test*

The third and final step of the interview concerned an oral sensemaking test. This is the same test as described in Level 8 except that it is done orally. This time I used two pictures—one real and one virtual picture. The real picture depicts a failure wherein a soccer field was designated as a new housing area. During the construction they dug too much soil away, making the surrounding levee unstable. On the crest of the levee a crack became visible and it also started to settle inwards. They quickly resolved the situation by providing a counterweight.

Although this real picture resembles the macro-instability failure, the exact situation was not practiced in the game environment. With the virtual picture this was



different. Participants could have seen this situation twice. The virtual picture shows some overtopping. The grass revetment of where the overtopping occurs is further—more missing. This indicates that the erosion of the inner slope started.

The reason I added this part to the interview was to validate the findings of the sensemaking test. A second but less important reason was to see if any differences become clear between the pre-interviewees and post-interviewees. Both failure pictures are different from what is used in the written sensemaking test and so they provide us with extra information about the impact of the game.

### Concerns about validity

One particular concern I had is whether participants would really elaborate on the questions and be able to convey in writing their thoughts. Except for a few participants, most participants kept their answer succinct on the written sensemaking test. This further increased my concerns with the validity of the test. The oral sensemaking test first of all shows that these concerns are misplaced. Here too most participants were succinct in their answers.

You see here a local overtopping. Revetment is probably flushed away—IPpre-#90

Others were even shorter and just said “settlement” or “crack.” Of course, just like the sensemaking test some people gave more elaborate answers. They were either motivating their answer or were looking for other possible but much less important indicators. Their actual answer came first. Look at this response by Participant #48 after asking him what he sees.

Crack on the road. A marker is positioned on the right hand side and that is most likely a point of interest. They might have crossed over this with a heavy machine. Based on all those signs I would say it is not accessible to public—IPpre-#48

With this example the answer is provided with the first sentence. The other sentences are additional observations. Such additional observations Participant #48 could have used to answer the other questions, such as what he would report and how he would diagnose the situation. Such thought processes happened very likely with the written test too. They could only not be observed—unless the participant decided to write this down. Some of them actually did and this confirms that the oral responses are largely not any different than the written responses. This means we can safely assume that the written responses are valid.

Another concern emerged throughout the training. Beforehand I made sure the pictures of the sensemaking test were large enough and still participants were complaining. They complained about the difficulty to make sense of them. Granted, some of the answers were based on an awkward interpretation. One that comes to mind is the interpretation that the illegal driveway represents a settlement. The participants who interpreted this thought the sand on top of the levee should be viewed as a gap.

Such misinterpretations recurred throughout this part of the investigation. They highlight a potential problem with this method, because those people with awkward

### Interview failure set



**Fig. 10.4** Pictures used during the interviews

- *Assessment*: The hometown failure is critical just like the another virtual failure.
- *Observation*: For the hometown failure a crack and even a small settlement can be observed; for the another virtual failure overtopping can be seen and a damaged grass revetment.
- *Reporting*: For the hometown failure (1) length and width of the crack; (2) the type of crack and if one or multiple cracks can be seen; and (3) the height and direction of the settlement. For the another virtual failure participants need to report (1) the accessibility of the area, (2) length and width of the damage, (3) the amount of water and whether infiltration occurred, and (4) if flushing of soil occurs and if so how much.
- *Diagnosis*: The hometown failure is an indication of macro-instability and the another virtual failure is most likely erosion inner slope, although it could also evolve into a micro-instability.
- *Measure*: For the hometown failure a sand berm is needed; for the another virtual failure foil with sandbags on the inner slope.

interpretations received a lower performance than what they might have been capable of. Therefore, for some the test may have likely been an underestimate of their actual abilities.

### Failure mechanisms and reporting do not come naturally

The reason to include the oral sensemaking test was to validate the written sensemaking test. Another validation opportunity could have been to validate the responses in accuracy too. I did not have this purpose with the oral test, because such accuracy validation would not be fair. Otherwise it would mean that two pictures decide over the fate of 14 others.

Two other issues made accuracy validation complicated. First, this test was provided after discussing the failure mechanisms. This means that the interviewees fa-

miliarized themselves (again) with the terms and we would expect them to perform better in applying the terms than during the written test.

Second, I was forced to neglect the reporting results. After listening to the interview recordings I concluded that most participants answered the reporting question insufficiently and that I as interviewer made insufficient attempts to make sure they answered this question accordingly. Because of this I only had a few interviewees who mentioned any reporting items.

#### **In-depth explanation: coding oral sensemaking test**

The coding of the oral sensemaking test is based on the written one. A transcript was first made of all oral responses. Then the same systematics were applied (Level 8). To reiterate, this involves the codes inaccurate (IA = 0), slightly accurate (SA = 1), accurate (A = 2), and very accurate (VA = 3) for calculating the accuracy scores for observing and diagnosing. In a nutshell, The very accurate score is provided if the answer is literally similar to that of the game; it is accurate if the answer is a good alternative; it is slightly accurate if it is correct but vague or descriptive; and it is inaccurate if the answer is incorrect or too vague.

For assessing, a simpler scheme was used: it is either accurate (A = 1) or inaccurate (IA = 0). This accuracy is also based on the game. For the real pictures the situation assessment is based on using the rules of the game in mind. For example, one of the rules in the game is that if a settlement occurs, it is always critical. This makes the hometown failure critical.

Scores have been calculated per learning objective and per picture. The first was calculated by adding the results for each learning objective over the two pictures. The score per picture was calculated by adding the results on assessing, observing, and diagnosing for each picture separately. The total score is a combination of the two scores per picture. The scores were converted into percentages by dividing each score by its maximum.

Paradoxically, the need to neglect the reporting results validates the issue with reporting found earlier. From the written sensemaking test we know that the reporting scores showed low improvements (about 10%) and the interviews highlight that answering this question does not come naturally. I had to pull this out of them—which I did not do, at least not consistently.

This does not mean we should ignore the accuracy results for both pictures completely. We could still learn something valuable. For example, if we look at the diagnosing scores, performance was better—as expected—by the pre-interviewees compared to the written test (Table 10.7; it was 13% on the written test). Although it improved, many pre-interviewees decided not to use any of the failure mechanism terms when I asked this question. The same is true for post-interviewees for whom the performance was more or less equal to the results on the written test. This was a bit of a surprise, because we just discussed the terms and I let them draw a picture for each one of them! It seems like using the failure mechanism terms does not come naturally too.

Remarkable too is that performance hardly improved for assessing. This result cannot be attributed to one picture specifically. With the hometown failure the situation looks less severe, but if a settlement occurs, it should always be considered critical. Interviewees have not considered this rule. They were led by the seemingly peaceful look of the situation and most judged it as reportable.

**Table 10.7** The results on making sense of the pictures during the interview by the pre-interviewees and the post-interviewees

Indicator	Max. <sup>a</sup>	Pre		Post		Improvement, %
		<i>M(SD)</i>	%	<i>M(SD)</i>	%	
Assessing	2	0.83(0.83)	42	0.86(0.90)	43	1
Observing	6	3.33(1.83)	56	3.71(1.80)	62	6
Diagnosing	6	2.17(1.40)	36	3.57(2.70)	62	26
Hometown failure	7	3.08(1.83)	44	3.86(2.34)	55	11
Another virtual failure	7	3.42(1.98)	49	5.00(2.08)	71	22
Total	14	6.50(3.50)	46	8.86(4.38)	63	17

*Note.* Max. refers to the maximum number of points participants could achieve.

With the “another virtual failure.” the opposite occurred. They were led by an implicit rule that is false and that is apparently persistent. One incorrectly assumes that as soon as water tops over the levee, the situation is out of control. An expert employee (IPpre-#149) explained during the interview that although overtopping needs to be carefully monitored, if the revetment is in good shape and the amount is not too large, the situation is under control.

With observing, the difference is minimal. This time the non-difference is caused by the hometown failure. Here the post-interviewees performed even slightly worse than the pre-interviewees. One of the post-interviewees (IPpost-#27) did not even see a problem. He saw a commercial sign, a boat, and that is it.

Observing performance did increase with the “another virtual failure.” In fact, performance increased for its contributing signal too. The main signal concerns the overtopping and the contributing one the grass revetment, because grass revetment damage is not an issue without the overtopping. Many more post-interviewees (71%) reported the grass revetment as well as the overtopping compared to the pre-interviewees (42%). This result contradicts what I discovered about contributing signals with the written test. There I found no difference between the pre- and post-test.

The difference in observing explains why the total improvement is larger for the another virtual failure. What is interesting is that participants practiced this in the game and so the post-interviewees should already have made sense of this. They did. It struck me how quick and to-the-point they answered my questions, even after a year:

*Casper:* What do you see?

*IPpost-#91:* That is *that* overtopping.

*Casper:* To what do you need to pay attention to when reporting this failure?

*IPpost-#91:* Location, length of the thing, and how much is gone.

*Casper:* What possible failure mechanism is causing this?

*IPpost-#91:* Erosion inner slope.

*Casper:* How do you assess the situation?

*IPpost-#91:* This is critical and something needs to be done immediately. It will not take long before it goes.

We cannot draw any firm conclusions based on the interviewee performance on these two pictures, but it does highlight that accuracy performance is very much dependent on the pictures—what signals the pictures contain, how many, if the pictures are ambiguous, and so on. We can further take away that failure mechanisms and reporting do not come naturally. Even after a training and a quick review on reporting and diagnosing, participants are still hesitant.

## Exercising on a Real Levee

I used pictures of real failures for the sensemaking test to approximate a real situation. Another approximation would be to look into behavior of participants during a *field exercise*. A field exercise concerns a training alternative where the patrollers go out and walk over the actual levees. On these levees the facilitators place markers with pictures of failures. The patrollers have to find these and communicate their findings to an Action Center.

To some extent, a field exercise overlaps with the game. It is also about recognizing, reporting, and communicating findings. However, such exercises have a stronger focus on communication, because participants have to make use of several forms of communication media, such as mobile, landline, and satellite telephones, and have to communicate within their team and with an Action Center. In the game these types of communication are very restricted, because players only communicate with a computerized Action Center and have just a few interaction possibilities.

Recognizing and reporting is in contrast much restricted with the field exercise. The pictures of the field exercise are static and provide already much needed information. The pictures show various hints as to what is to be seen on the picture (“quicksand” and “no cracks”) and they give away most answers for the reporting items (“25 meters”). All that patrollers have to do is to find a marker with a picture and then fill out a form by repeating what is mentioned on the picture. Just a bit of interpretation is needed, because a decision needs to be made about what signals appear on the picture.

The markers themselves are the most clear restriction. Players go out to find these sticks and not failures. Patrollers are not sensitized to look for failure signals this way and this makes the search process completely different from what they are supposed to do with an actual levee inspection. The game arguably does a better job at this.

Whereas the game further touches more onto judging and diagnosing failures, a field exercise relates more to the logistics of organizing a levee inspection and gaining regional-specific knowledge. Organizing a levee inspection with sometimes hundreds of patrollers requires practice too. They need to be instructed and handed over proper equipment. Subsequently, it needs to be ensured that communication equipment works and that the Action Center knows what to do and is able to handle all requests. These organizational matters are clearly not part of the game.



**Fig. 10.5** An example of a field exercise failure picture. Notice the provided information

Undeniably, it is necessary that the patrollers know their own region and especially because much of the patrolling would very likely happen at night with bad weather. It would also be useful if they know where the weak spots are located. The current game is set in a fictional environment and although it is not impossible to visualize existing levees in the game (and we accomplished this much recently with the use of GIS data), navigating the actual environment should stay an important training component. Virtual environments remain an abstraction.

Despite some of these differences in training objectives, it would be interesting to find out if any transfer occurs from the game to this approximation. The field exercise is an approximation because just like the game, it is a simulation of the actual inspection process. In collaboration with Organization C it was decided to investigate the game's effect by comparing two groups: the group who participated with the training and another, much similar group. The two groups are from here on referred to as the *Game Group* and the *Control Group*. Before the onset of the game-based training we chose which two out of six regions would be compared. According to the coordinators of Organization C two specific regions are very much alike in terms of size, regional characteristics, and mix of volunteers and employees. I trusted their judgment on this and we assigned randomly one to the Game Group and another to the Control Group.

### *Setup of the Field Exercise*

Each year Organizations B and C try to organize a field exercise. The general setup at both organizations is comparable. Pairs of two or three patrollers are assigned a levee segment and they have to go out and look for sticks with levee failure pictures and communicate their findings to an Action Center. One of the differences is that for Organization B, the patrollers communicate with their post commander who, on its turn, communicates with the Action Center. Another but lesser important difference is that at Organization B the patrollers keep their own equipment (flashlight, shoes, and coat among others). At Organization C this equipment is stored at the regional headquarters and handed over at the start of an inspection.

The two organizations try to learn from each other. Organization C invited a number of observers from Organization B and this happens the other way around too. Organization A has to my knowledge never organized a field exercise. They did organize field trips to the levees. During such trips patrollers get a tour of the region and receive some information about its weak spots and other peculiarities.

With an actual inspection patrollers have shifts. To make it possible to let everybody practice all at once two teams were created for each levee segment: Team A and Team B. Team A started on one end of the levee segment and Team B on the other. Each team had its own phone operator to prevent the pressure on the Action Center. They also had their own markers. The facilitator put about once every 15 minutes walking a marker somewhere on the levee for each team. The sign on the marker had a picture of a failure and indicated whether it was meant for Team A or B.

Upon finding a marker, patrollers first had to fill out a failure registration form (much similar to the game). To further lessen the pressure on the phone operator, only 50% of the markers had to be communicated to the phone operator—one or two with the satellite phone and all the others with the mobile phone (to either a landline or a mobile phone). When the phone operators received a call, they asked questions using the same failure registration form and they put the answers into a database system. The commander who sat in the same room as the phone operator had to judge this information and make a diagnosis using another form.

Although the patrollers were urging me to go outside with them, up front I decided it would be better for me to sit at the Action Center. There I could listen to the incoming phone conversations from various pairs and not base my observations on the performance of one or two pairs. Because each team had its own Action Center, another observer was sitting at the second Action Center. In addition to the observations, I handed out a very short pre-questionnaire before and after the exercise.

The Game Group consisted of 21 participants. This is much fewer than the initial group of 35 participants of the game-based training, but this group included six regional commanders and three inspection coordinators. The Control Group involved 28 participants.

### *Observing the Action from a Corner*

Differences between the two groups became apparent from the beginning—but not due to the game. At the start of the exercise participants had to get their gear and their instructions. When the Control Group participants arrived at the entrance, they had to go left or right depending on their team. On confirming their attendance, they received their gear immediately. Individual teams then received a short instruction on how to use the satellite phone and just before the training started one of the facilitators gave an instruction about the exercise in the general.

With the Game Group, one of the key people for preparing the exercise arrived late. Due to this, the exercise was more disorderly. The logistics of the building also did not help. Unlike with the Control Group, this time participants had to go all the way up to the canteen on the third floor and wait there until everybody arrived. Together they went down to the basement to get their gear and then went back up again to the canteen to receive the instructions. Because of the delay these instructions were rushed over. I noticed that the facilitator forgot to tell important information, something he did tell two days before with the Control Group.

Next to these logistics the weather played a role. On Tuesday with the Control Group the weather was excellent for a Dutch November. The sky was clear and the patrollers did not have to face any rain or wind. Two days later the weather was drastically different. It was pouring rain and the wind was strong. In fact, a weather alarm was given just before the exercise. We could speak of a storm on that day. The water authority also received phone calls from worried citizens. They thought the patrollers came out because of a serious flood threat!

With both groups I sat down with Team A and the other observer with Team B after the patrollers left the building around 7:00 p.m. What struck me most was not so much a difference between the two groups, but a difference between the two phone operators. My phone operator with the Control Group was calm and handled every phone call with ease. The one with the Game Group had much more difficulty:

I cannot see the forest for the trees. I get too many reports. I do not make it to fill these out in *FLIWAS*<sup>1</sup>.

Admittedly, the number of phone calls differed much. The Control Group phone operator received nine phone calls and the one at the Game Group 23—of which 17 were actual reports. The other phone calls were tests, incorrect, or did not get through. Also, the first phone call made at the Control Group's Team A was around 8:00 p.m.; with the Game Group it started from about 7:30 p.m. However, in that one hour time span the Control Group's phone operator spoke for 36 minutes (53% occupancy rate) and that is relatively similar to the 54 minutes the Game Group's phone operators talked in the one and a half hour time span (59% occupancy rate).

---

<sup>1</sup> The program that the phone operators used was called *FLIWAS*. This is the acronym for FLOOD Information & WARNING System.



**In-depth explanation: observing the Action Center**

During the field exercises two teams were used—Team A and Team B—with each their own Action Center. The Action Centers were set up in offices with computers, laptops, phones, and a map of the area. The offices were large enough so I and the second observer could sit in a corner without hampering the activities. I instructed the second observer about the exercise and levee inspection in general.

We noted the time and duration of each report and wrote down as much about the conversation as possible. For example, we wrote down what signals the participants talked about and how they judged the situation. It was difficult to disclose everything, because we could only hear the phone operator talk. On occasions we were able to hear the patrollers too, when they talked loud enough into the phone. In addition, sometimes phone calls followed one after another. This made it hard to keep track of the conversations.

The average phone durations were calculated as well as the *occupancy rate*. This is a percentage of how much time a phone operator was on the phone throughout the exercise: Time on phone / Total exercise time. To get a fair comparison, the total exercise time was determined by considering the first and last phone call of each phone operator.

Any other peculiarities were written down too, such as how the Action Center operated. Informal talks with the phone operators followed after the exercise.

A number of factors may explain this stress. First, the computer was not functioning well. The computer was needed to fill out the reports by the patrollers in a specific computer program.

Second, the phone operators received three short courses before the exercise, two on how to use the computer program and one on the procedures. In an informal talk with phone operators, each indicated that these courses were not sufficient to become familiar with the material and the procedures and they wished they had participated with the game-based training. The vocabulary was especially still unfamiliar to them. The Control Group's phone operator was more familiar, because she was closely involved with organizing the exercise.

Third, some procedures were not clear and/or were interpreted differently at each team. The Action Center of Team A and B consisted of a phone operator and the commander. Strictly speaking, the role of the phone operator was to communicate with the patrollers about the failures. The commander's role was to interpret the findings. However, there was a lack of clarity about what tasks needed to be done by whom. One such task involves who is calling the people for taking measures. The other observer noted that his commander at the Control Group did this, but his Game Group commander asked the phone operator to do this. As a matter of fact, this commander was of the opinion that all phone calls should be dealt with by the phone operator. When the phone operator line was busy, some patrollers chose to call the commander directly. Unlike the other commanders, he referred them to the phone operator.

Now his phone operator was also not so busy. She received eight phone calls and only five of them were about reports (32% occupancy rate). The other phone calls were attempts with a satellite phone, but these did not get through or broke off in the middle of a conversation. So all conversations happened with a mobile phone at this team. Later we understood that one of the couples did not call at all.

This explains the fewer phone calls. In contrast, the other observer's team during the Control Group exercise received 18 phone calls and 15 of them were reports (55% occupancy rate).

The commander at my Game Group's commander did not make clear agreements with his phone operator about the procedures. With my Control Group such agreements were made. The commander would deal with phone calls about taking measures and the phone operator would focus on the reports. I noticed that not having any clarity was a stress factor to my Game Group's phone operator. She seemed to have a need for something to hold onto.

Finally, the Game Group patrollers took initiative. The idea is that patrollers call and answer the questions of the phone operator. However, some of the Game Group called and immediately said what they saw (a crack or water outflow!) and gave their opinion:

This is a very serious report! Something needs to be done immediately! I have no time to fill out the form.

This led to some confusion. Patrollers spoke of terms that the phone operator was not familiar with and for which she could not find a form. One of these concerns the liquefaction signal. This term is used in the game, but Organization C decided to not use this in their own forms. In the same vein, they spoke of reporting elements that did not need to be filled out with Organization C's reporting forms, such as the height. The Control Group, on the other hand, neatly followed the communication procedure.

This latter factor is also the most important observed difference between the two groups. The other observer noticed this too. They further seemed more immersed into the exercise and were more playful about it:

*Patroller:* 87 points...that is a huge one!

*Commander:* Guys, it is just a game. Relax!

*Patroller:* Well, if we think like that...we have to be serious about it.

and

*Patroller:* We see an animal activity and it is critical. It seems an elephant walked over the levee.

Despite such comments and confusions during the phone conversations with the Game Group, their average phone call was not longer. In fact, a call from the Game Group took about 3 minutes and 2 seconds ( $n = 25$ ;  $SD = 81$  s) compared to 3 minutes and 24 seconds ( $n = 26$ ;  $SD = 52$  s). It is not shorter either.

Although both groups were excited about the exercise, the atmosphere differed. The Game Group seemed to be more enthusiastic and jovial; after all those hours of sitting behind a computer, they were finally able to do the "real thing." The Control Group was more formal about the exercise.

## ***Opinions and Perceptions***

Before and after the exercise I gave participants a short questionnaire. I gave the pre-questionnaire when they picked up their equipment and the post-questionnaire when they came back to the regional headquarters. Because I knew that my involvement in this exercise needed to be as unobtrusive as possible and they were working on a tight schedule, I restricted myself to six items for the pre-questionnaire. The first four items were identical to those of the questionnaires used during the game-based training. With the post-questionnaire I used four items and two open questions. The items and the responses are shown in Table 10.8.

In deciding on the four recurrent items, I included the item that asks about levee inspection knowledge in general (1). The other three items (2–4) were about the two most important learning objectives for patrollers: recognizing failures and reporting them. The remaining two items on the pre-questionnaire were to assess the participants' perceptions on their preparedness and performance expectations (5–6). Unlike the Game Group, the Control Group did not have any levee inspection activity throughout the year. The last time was about a year ago, when they were involved in the first field exercise. Common sense would, therefore, dictate that the Game Group should have perceived to be better than the Control Group on these matters.

On the post-questionnaire I asked the participants to judge the communication—within their own team and with the Action Center (7–8). Because the participants of the Game Group have had a shared experience already, they might have understood each other better and were able to communicate more easily with the Action Center. Then I also asked them to judge the exercise in general (9). Finally, I asked if the Control Group was interested in participating in a game-based training and if the Game Group still found their participation any useful.

I included two open questions about what went right and wrong during the exercise. I expected to receive more elaborate responses from the Game Group, because they should be more knowledgeable and should be able to notice details that the Control Group would not see.

The exercise started and ended somewhat chaotically and because of that four participants of the Control Group did not fill out the pre-questionnaire and two others the post-questionnaire. With the Game Group only one participant did not fill out the pre-questionnaire. I also excluded the new member of the Game Group from the analysis. Unfortunately, three participants of the Control Group jointly filled out the questionnaires and I decided to exclude them too. The total number of participants was consequently about the same for both groups (Table 10.8).

### **Control Group seems more confident**

Contrary to expectations, the Game Group did not perceive to *a)* have more knowledge, *b)* recognize and report failures better, *c)* feel better prepared, and *d)* perform better. In fact, the results show that the opposite seems to be far more true—and

**Table 10.8** The results in percentages of the pre-questionnaire and post-questionnaire by the Game Group ( $n_{pre} = 19$  and  $n_{post} = 20$ ) and Control Group ( $n_{pre} = 20$  and  $n_{post} = 23$ ) during the field exercise

Item	Group	Rating, %							Mann-Whitney			
		1	2	3	4	5	6	7	<i>U</i>	<i>z</i>	<i>p</i>	<i>r</i>
<i>Pre-questionnaire</i>												
1. I have much knowledge about inspecting levees.	Game	0	0	11	47	37	5	0	113	-2.26	.024*	.36
	Control	0	0	10	20	25	45	0				
2. I know what kind of failures could appear.	Game	0	11	74	16	0			105	-2.74	.006**	.44
	Control	0	0	45	50	5						
3. I can recognize failures.	Game	0	11	74	16	0			89	-3.17	.002**	.51
	Control	0	5	25	70	0						
4. I know what to pay attention to when reporting a failure.	Game	0	5	74	21	0			138	-1.87	.061	.30
	Control	0	10	33	57	0						
5. I am prepared for this exercise.	Game	0	21	74	5	0			135	-1.99	.046*	.32
	Control	0	14	48	38	0						
6. I expect to perform well during this exercise.	Game	0	0	11	16	53	16	5	138	-1.74	.082	.27
	Control	0	0	10	14	14	52	10				
<i>Post-questionnaire</i>												
7. The communication in my team was good.	Game	0	5	10	65	20			223	-0.20	.84	.31
	Control	0	4	9	65	21						
8. The communication with the Action Center was good.	Game	0	21	32	47	0			176	-1.22	.22	.19
	Control	0	17	13	70	0						
9. In general I found the exercise good.	Game	0	0	35	65	0			204	-0.74	.46	.11
	Control	0	4	44	48	4						

*Note.* For Items 1 and 6 a 7-points scale was used and for all others a 5-points. The 5-points was measured in one direction (from “Not well” to “Very well”) and the 7-points in two directions (from “Completely disagree” to “Completely agree”) for reasons explained in Level 3.

\*\*\*  $p < .001$ ; \*\*  $p < .01$ ; \*  $p < .05$  (two-sided).

especially when it comes to knowledge, recognition, and reporting (Table 10.8). The Control Group seems more confident without having done anything for over a year. This is a surprising result.

It is surprising, because the Control Group did not consist only of experts. I interviewed the commander of that region, who said:

The patrollers are all externals. The army is composed of civilians. They do not have most of the needed knowledge. Almost all of them recently started—IPre-#133

If we compare the results by the Game Group with their answers in the past on the pre- and post-questionnaire of the game-based training, it becomes clear that except for the second item the ratings subtly improved compared to the post-questionnaire answers. The Wilcoxon signed ranks test reveals, however, that they were only significantly different from the pre-questionnaire. So a half year after the training the overall perceptions by the participants remained the same.

On the post-questionnaire, no differences are evident between the two groups. Only with the communication with the Action Center is there again a tendency that the reverse of what was hypothesized might be more likely. About 70% of the Control Group thought the communication was good to very good compared to just 47% from the Game Group. It could be that some of their dislike of the computerized Action Center transferred to the field exercise, but it is more likely attributed to the operators at both evenings. As I explained before, the Game Group phone operator of the team I observed had much more difficulty with the exercise compared to the one of the Control Group. Nevertheless, based on these results we certainly cannot say that the Game Group perceived to communicate better.

### Game Group focused on substance too

Regarding the two open questions, we find that the Game Group may have been more critical—albeit in a positive way. It turns out that the Game Group did not write more words when asked what went wrong during the training ( $M_{\text{game}} = 6.44$ ,  $SD_{\text{game}} = 3.90$ ;  $M_{\text{control}} = 4.59$ ,  $SD_{\text{control}} = 3.54$ ), but they did write more when asked what went right ( $M_{\text{game}} = 5.80$ ,  $SD_{\text{game}} = 4.38$ ;  $M_{\text{control}} = 2.82$ ,  $SD_{\text{control}} = 2.07$ ),  $t(19) = -2.41$ ,  $p = .026$ ,  $r = .48$ . Overall, adding the words for both questions we see that the Game Group wrote about twice as much ( $M_{\text{game}} = 11.3$ ,  $SD_{\text{game}} = 7.11$ ;  $M_{\text{control}} = 6.63$ ,  $SD_{\text{control}} = 4.23$ ). The Game Group had most certainly much more to say about the exercise.

What they had to say differed too. The Control Group complained primarily about the lack of time. Exactly one-third made a comment about this. One of them said this was because of the travel time. Surprisingly, none of the Game Group mentioned this although they had to travel too. In terms of percentage, both expressed about equally about not getting enough information up front about the exercise and having difficulty with finding the markers and recognizing the pictures. Some found the pictures unclear.

#### In-depth explanation: coding the open questions

The post-questionnaire had two open questions. One asked about what went wrong during the field exercise and the other about what went right. First, the number of words for each answer was automatically calculated. Then I coded the answers by identifying *themes*. The identification of themes was relatively easy, because it was quite clear-cut about what the participants were speaking of. I asked another rater to identify themes as well so to be able to consider the reliability of my coding. I came to a total of ten themes. The other rater

only came to seven. Three codes were identical: time, communication, and collaboration. This is not surprising, because many participants used the exact same (Dutch) words in their answers.

Two codes were synonymous. I used “explanation” and “equipment” and the other rater spoke of “information” and “material,” respectively. The code explanation (or information) refers to instruction given at the start of the exercise and the debriefing at the end of it. Equipment refers to comments about satellite phones, flashlights, or other “material” the patrollers had to work with.

The remaining two codes did not have such an agreeable meaning, but they referred to the same answers. One I decided to call “organization” and this code is about comments that refer to how the exercise was set up. The second is “environment.” A couple of participants indicated that they liked to practice in the dark and to get to know the region. A couple of participants indicated that they liked to practice in the dark and to become familiar with the region. One said to dislike the weather. Both codes hardly appeared.

I had three extra codes. The other rater considered comments about “forms” as part of the information theme. I made it into a separate theme, because the reporting forms have a strong link to the game and I clearly noticed a difference between the two groups relating to this (sub)theme. For a similar reason I coded comments about recognizing failures into a separate code called “recognition.” The other rater considered dealing with failures as part of the “environment.” Then I made a separate code about having “difficulty” with certain aspects of the training and the other rater coded these comments as part of either the equipment code or the environment code. Some of the comments were about the pictures and others about having difficulty in finding these pictures in the dark.

The three extra codes especially made a difference in coding and were the major reason why the inter-rater agreement came to 79%. That is a reliable outcome and especially when considering that both raters had to come up with a coding scheme and apply their codes to the data.

Then the Game Group shared more concerns about the communication with the Action Center (Game Group = 29% vs. Control Group = 20%). Both indicated that the Action Center was not always available. In addition, some of the Game Group participants mentioned that a list of telephone numbers was missing. This is correct. Due to the chaotic situation at the start, some teams did not get this telephone list (and others received two of them) and this may actually explain why the Game Group tended to be more dissatisfied in their communication with the Action Center.

Another reason may have to do with the equipment. Contrary to the Control Group, far more participants from the Game Group were dissatisfied with their equipment and in particular with the satellite telephone (Game Group = 65% vs. Control Group = 20%). As we observed, the satellite phone also did not work well with the Game Group and especially at Team B. This is a contrast, because some participants of the Control Group indicated that the equipment—and in particular the satellite phone—was one of the positive aspects of the exercise.

Significantly, a fair number of participants from the Game Group (18%) complained about the reporting forms. They said that too many of them exist and it required much flipping through the pages to get to the right form. How it works with reporting with the field exercise is that each signal has its own form, much like the game. However, in the game players automatically go to the right form after selecting the signal. Such ease and comfort does not exist with the paper-based version.

Having experience with the computerized forms may have likely made the participants more critical about how it works during the exercise. The Control Group did not know any better. The only comment about the forms was a suggestion to combine the forms into a booklet instead of giving them separate forms.

Only the Game Group reacted positively to the forms. Almost one-third (31%) said something about the forms. They thought the forms were clear and that they worked. They further mentioned it was an easy way of reporting the signals. One even considered the forms the sole positive aspect of the exercise. Thus, unlike the Control Group, the Game Group had much to say about the forms—negatively but also positively.

Another striking difference is that the answers of the Control Group all related to the equipment, communication, and, especially, the collaboration. Except for the forms, the Game Group participants (31%, to be precise) also added comments about recognizing the failures. It seems like the Control Group focused more on the organizational parts of the exercise and the Game Group focused on such structures as well as the substance.

### **Game-based training was useful**

The post-questionnaire for the Control Group ended by asking about their willingness to participate with a game-based training. Most somewhat agreed (41%) with this statement and an almost equal number agreed with it (36%). All but one person disagreed with it. It seems that also this group has a fair amount of interest in participating in training with *Levee Patroller*.

I asked the Game Group to look back and consider if they agreed whether the game-based training was useful in participating with the field exercise. Two participants disagreed and one strongly disagreed (which includes the rebel and co-rebel). However, most agreed (45%) and a good number even strongly agreed with it (20%). The majority indicated that playing the game was useful for the field exercise.

### ***The Game's Effect on the Field Exercise***

The sample size is clearly not large enough to make strong concluding remarks. However, this was not the purpose of considering the field exercise. The purpose was to explore possible effects of the game-based training in another situation. In that way we could validate its outcomes to some extent—not firmly but tentatively.

Beyond the sample size, several factors played an important role in how the exercises functioned, such as the logistics, the weather, instructions, phone operators, and the equipment. The comments by the participants confirm their importance. These factors make it hard to speak of a fair comparison.

Then, also noteworthy, the exercise has different training objectives compared to the game. Some of them overlap and so the game could have an effect there, but it

is almost like comparing apples with oranges. The participants of the game-based training concluded this themselves too: the game is complimentary. It serves another purpose within the education of patrollers.

Being aware of the restrictions, I applied efficient research methods that would be unobtrusive, require little effort, and give a maximum return-on-investment. The downside to these is that they are less objective and rigorous compared to the more intensive methods, such as recording conversations and examining the filled out failure registration forms. The latter may have been especially non-rewarding, because participants missed markers and many forms went missing, making it again difficult to get a fair comparison. In addition, if the results from the game are to be transferred to the exercise, we would expect the Control Group to report the failures well from the beginning too.

Consistent with this assumption, the commanders who debriefed and evaluated the failure registration forms with the patrollers told me that they did not notice a particular difference between the two groups. Organization C was hoping to find a major difference in performance. This would give them a good reason to invest in the game-based training. On the surface, the groups did not differ however.

In fact, it could be argued that playing the game frustrated the exercise. Using their knowledge from the game, the patrollers confused the phone operators who did not have that knowledge. With the Control Group no such problems occurred, because here the phone operators were in control and simply followed the protocol of the reporting forms.

I would argue instead that this is a positive effect. Although a better alignment between the game and exercise is desired, it is an indication that the participants are knowledgeable. They are able to look beyond the failure registration forms. The confusion is undesirable, but this should be fixed on the side of the phone operators and not on the side of the patrollers. The phone operators should receive a comprehensive training too. Otherwise the failure registration forms are what predominate the crucial parts of the inspection. These forms are and should remain a tool. This point relates to the discussion about the role and responsibilities of patrollers: Are they passive messengers who merely pass information to others or are they active interpreters who understand what is going on?

Although the Game Group may have been more willing to participate with the questionnaires, this knowledgeableness speaks from their answers on the open questions too. They wrote more elaborate answers and talked about substance too. The Control Group concentrated on form only. These concern the organizational elements, such as communication and collaboration, and the equipment.

Surprisingly, the knowledgeableness of the Game Group does not speak from their perceptions. The Control Group in fact demonstrated more confidence, especially regarding what they perceived to know about levee inspection. It might be that they were knowledgeable and more so than the Game Group was at first. This is something I did not test. Even without such testing, it remains surprising, because the Control Group did not participate in a three week training and Organization C thought the two groups were very much comparable—also in terms of expertise.



Assuming the Control Group could not have been more knowledgeable, what may have happened is that their participants were *unconscious incompetent*.<sup>2</sup> They did not know any better. Another and complimentary explanation is that self-perception is based on the type of activity. With the game-based training participants need to acquire knowledge; with the exercise they need to apply it. For acquiring people may underestimate or undervalue their abilities and with applying they overestimate or overvalue their abilities. The latter has to do with self-confidence. Sport teams also do not go into a competition with the idea that they cannot win.

In that case it might also have happened that the Game Group participants were *conscious incompetent*. Because they know what is required for levee inspection, this may have adjusted the perception inflation. A sports team that consciously knows that they play against a far better opponent is content with a draw and does not go for the win necessarily.

This is—of course—mere speculation. What is certain is that the Game Group's perceptions remained stable. This shows that the game's effects sustained after even half a year. We can further be confident that the majority found the game useful for their participation in the field exercise.

Therefore, the comparison revealed factual and perceptual evidence for the game's effect on the field exercise. The evidence is only not very overwhelming. It certainly is too little to convince others to invest in game-based training. Then again, this may have been too much to be asked for with the many restrictions in mind.

## Lessons Learned

The first attempt concerns a (failed) attempt to set up the game-based training with students, but from which one can garner valuable insights. For example, we were able to confirm that the patrollers are computer illiterate in various ways—in perception and with actual gameplay. The digital literates got better scores and picked up the game much faster. They also played differently: it was a collaborative effort instead of an individual one.

The student results with the sensemaking test highlight that although the patrollers may not be as skilled in picking up the game, they showed themselves to be knowledgeable (at least more so than the students).

But not more so than the experts. The second attempt served to look into how experts make sense of failures and these performed better than the patrollers before the training. After the training the patrollers are equals. The game helped them to perform on the level of the super experts. In fact, employees turn out to perform even better.

---

<sup>2</sup> The terms unconscious and conscious competent have unclear origins, but are widely used in training literature as part of a four stages of competence model (e.g., Scannell & Les Donaldson, 2000).

The interviews—the third attempt—confirm what was found by benchmarking the sensemaking results between the patrollers and students: the patrollers were already knowledgeable. They further highlight that patrollers have trouble with reporting and diagnosing. Especially the failure mechanism terms are a problem and this continues to be a problem after the training.

The interviews furthermore suggest that the game has had a lasting effect. The exact terms may have been forgotten, the post-interviewees had a seemingly better hunch about levee inspection. This was even the case a year after the training.

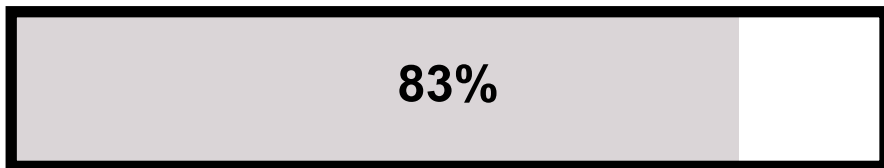
The fourth and final attempt was the field exercise. This showed us the surprising finding that untrained patrollers are more confident and that trained patrollers look differently at the exercise. They looked at the form and substance. But most importantly, the trained patrollers found the game-based training a useful precursor to the field exercise.

## Level 11

# Integrating the Puzzle Pieces

*The more you do it, the more fun it becomes...It is now just click, click, and go!—Participant #139*

*It does not make you go poo, but it does give you a headache—Participant at a course at Organization A*



The previous levels provided the puzzle pieces, insights gained from the training/evaluation with *Levee Patroller* that together are key to our understanding of the game-based training. In this level we put the puzzle pieces together to form a coherent picture.

If something should have become clear from the previous levels, then it is the enormous diversity among the players. For one participant the game becomes more and more fun; for another it gives a headache. These differences are reflected by the game scores. They pretty much range from the lower to the upper end. What is causing this? This level attempts to shed some light on this as well. What we in fact attempt to answer here is to answer the following two questions that have been part of the quantitative approach (QUAN) of the training/evaluation:

1. What is the effectiveness of the training with *Levee Patroller*?
2. What factors contribute to its effectiveness?

As a guide for answering these two questions and for integrating the puzzle pieces, I formulated 15 working hypotheses in Level 3 about the outcomes that determine the effectiveness of the training and variables that may influence the effectiveness (Table 11.1).

But even if we answer these questions we have not established the sought for understanding of game-based training. Although the puzzle pieces are integrated and make up a puzzle, we still need to figure out what the puzzle is about. For this

reason the training/evaluation made use of a qualitative approach (QUAL) too and this aimed to answer the following two questions:

1. How do participants experience the game-based training?
2. How do participants play the game?

Forthofer (2003, p. 710) suggests that “mixed methods designs are inherently more complex, and those that attempt any integration or synthesis of results across methodologies require an additional phase of ‘meta-interpretation’.” In my case the qualitative approach functions as an additional perspective onto the results of the quantitative approach: it tells us what they really mean (or could mean).

The goals of this level are to describe

- The effectiveness of *Levee Patroller* and its contributing factors;
- How participants experienced the game-based training and played the game; and
- A reflection on the results.

## Accepting the Hypotheses (or Not)

Table 11.1 summarizes the working hypotheses. A number of them were answered in the preceding levels. For the sake of clarity I will repeat these results here and conclude what this means for the hypotheses: can they be accepted or should they be rejected?

Other hypotheses have not been answered. These concern the hypotheses that require an investigation of the relationship between variables that were retrieved with different methods. To investigate this I took the structural model from Level 7 (Fig. 7.1) and added new variables to it. These new variables are the total sensemaking performance scores, the word count, and the average total game score.<sup>1</sup>

In constructing the more comprehensive structural model I followed the same steps as with the previous model. I first looked at the correlations between every two variables that were included. Then I considered the partial correlations and removed the relationships that were mediated by other variables. This process resulted in Figure 11.1. This figure shows the original correlations and not the partial ones.

In structuring the discussion of the results I make the same distinction as in Level 3 between hypotheses that relate to the outcomes, those that are about variables that influence the outcomes, and those that involve a difference in outcomes among participants. I will start with discussing the hypotheses about the outcomes.

---

<sup>1</sup> I used the sensemaking performance and the word count scores of the core set of failure pictures, which means I neglected the non-failures and new failures. The total core scores are a better representation of performance, because the core set of pictures were trained.

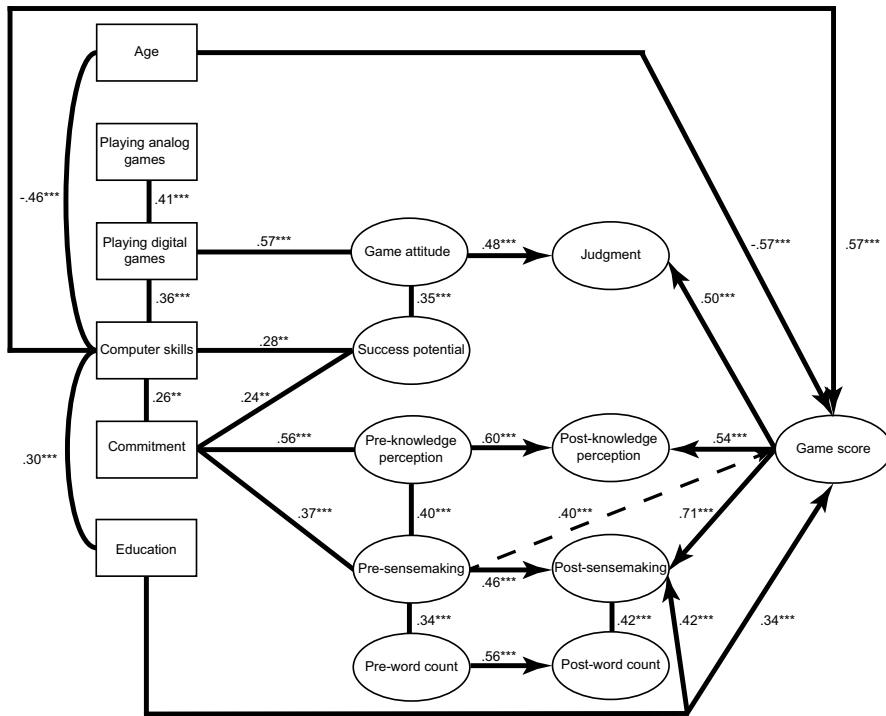
**Table 11.1** The 15 working hypotheses and their acceptance

H	Description	Accept?
1	Post-training knowledge perception (Hypothesis 1.1) and sensemaking (Hypothesis 1.2) will be higher compared to the pre-training knowledge perception and sensemaking performance, respectively.	Y/Y
2	Knowledge perception does not correlate strongly with sensemaking performance; the correlation coefficient will be lower than .50.	Y
3	Post-training sensemaking performance on virtual pictures is higher compared to real pictures.	N
4	Game judgments correlate with knowledge perception (Hypothesis 4.1) and sensemaking performance (Hypothesis 4.2); participants who evaluate the game higher will have a higher knowledge perception and sensemaking performance.	N/N
5	Post-training vocabulary use will resemble the games vocabulary closer compared to pre-training vocabulary use.	Y
6	Post-training word count (Hypothesis 6.1) and dispersion (Hypothesis 6.2) will be lower compared to the pre-training word count and dispersion, respectively.	Y/Y
7	The number of exercises played moderates the results on the main outcomes; participants who play more exercises will have a higher results on the main outcomes.	Y
8	Game scores moderate the results on the main outcomes; participants with higher game scores will have higher results on the main outcomes, respectively.	Y
9	Computer skills (Hypothesis 9.1) and game skills (Hypothesis 9.2) moderate the game scores; participants with higher computer and game skills will have higher game scores, respectively.	Y/N
10	Game attitude moderates the game judgments; participants with a higher game attitude will have a higher game judgment.	Y
11	Motivation (Hypothesis 11.1) and expectations (Hypothesis 11.2) moderate the results on the main outcomes; participants with higher motivation and expectations will have higher results on the main outcomes, respectively.	N/N
12	Results on the main outcomes are similar between volunteers and regular employees (Hypothesis 12.1); and the expert employees will achieve higher results compared to volunteers and regular employees on knowledge perception and sensemaking performance (Hypothesis 12.2); but the learning gains of volunteers and regular employees are higher compared to the expert employees (Hypothesis 12.3).	Y/Y <sup>a</sup> /Y <sup>b</sup>
13	Students (Hypothesis 13.1) and younger participants (< 40 years; Hypothesis 13.2) achieve higher game scores compared to older participants (> 40 years).	Y/Y
14	Patrollers pre-training sensemaking performance is higher compared to students sensemaking performance (Hypothesis 14.1) and less compared to super experts sensemaking performance (Hypothesis 14.2); patrollers post-training sensemaking performance will approximate the super experts sensemaking performance (Hypothesis 14.3).	Y/Y/Y
15	The Game Group has higher confidence before the field exercise (Hypothesis 15.1) and communicates better during the field exercise (Hypothesis 15.2) compared to a Control Group.	N/N

Note. H = Hypothesis; Y = Yes, accept; N = No, reject.

<sup>a</sup> This is true except for the post-sensemaking performance.

<sup>b</sup> With knowledge perception gains are visible but not significant.



**Fig. 11.1** The relationships between the main variables based on Pearson correlations (2-tailed). The arrows represent a causality we would expect based on the sequence of measurements. The other relationships are bidirectional. \*  $p < .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$  (two-sided)

### *It Clearly Did Something*

Before the implementation of the game-based training I identified three main outcomes: knowledge perception, sensemaking performance, and game judgments. In Level 7 I showed that with the retrieved empirical data it was possible to construct a summated scale for the first main outcome. The measurement of knowledge perception both before and after the training gave rise to two variables: pre-knowledge and post-knowledge perception. The sensemaking performance I defined up front. It concerns the overall score of how accurate participants made sense out of the pictures on the sensemaking test (Level 8). It also has two variables associated with it: pre-sensemaking and post-sensemaking performance.

The idea behind the game-based training was to have a positive effect on both these outcomes. It turns out that this game-based training affects strongly and positively participants' knowledge perception,  $t(111) = -8.49$ ,  $p < .001$ ,  $r = .63$  (Hypothesis 1.1). The same is true for the sensemaking performance,  $t(124) = -19.2$ ,  $p < .001$ ,  $r = .87$  (Hypothesis 1.2). We can therefore be confident about concluding that the game-based training had a positive impact on its two main outcomes.

The second hypothesis predicted that self-assessment is only moderately correlated to actual learning, based on the evaluation principle “more than the tip of the iceberg” and the results by Sitzmann et al. (2010). This means that I expected that knowledge perception, which is a self-assessment, does not strongly correlate to the sensemaking performance. A correlation is considered strong when it has a value of .50 or higher (Cohen, 1988).

This second hypothesis is supported too. Figure 7.1 shows the correlation between pre-knowledge perception and pre-sensemaking performance. This correlation is indeed moderate. The correlation is exactly similar for post-knowledge perception and post-sensemaking performance, suggesting a consistent, moderate relationship exists between the two main outcomes. What this additionally tells us is that the feedback provided in the game did not help to modify this relationship. According to Sitzmann et al. (2010, p. 172) “If learners receive feedback on their performance, they should modify their self-assessments to be more aligned with their actual knowledge levels.” This did not happen, which means that we need to be prudent in the usage of self-assessments in game evaluations or find ways to make use of more accurate ones.

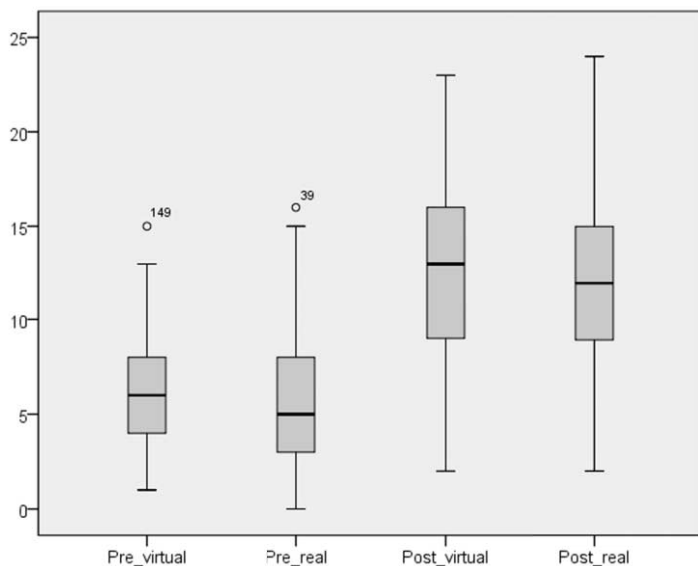
Regarding the increased sensemaking performance it is further important to consider whether it matters if the pictures are virtual or real. If it does not, transfer is strong. The game is then able to assist in making sense of real pictures. If it does, we can draw several conclusions. Either some transfer occurs. Participants still improve on the real pictures, but not as much as on the virtual pictures. This is a fair assumption: participants have been practicing with the virtual pictures, not with the real ones. Or no transfer at all occurs. In that case, participants simply improved on the virtual and not the real pictures.

I did not expect that no transfer at all would occur, but I did expect that participants would perform better on the virtual pictures (Hypothesis 3). I was incorrect; overall, no differences were found—before and after the training (Fig. 11.2). This is an exciting and unexpected outcome. Not only do participants improve in recognizing virtual pictures, they actually improve in recognizing real ones too. And as no differences exist, what is learned in the game is 100% relevant in the real world.

The third main outcome variable, game judgment, is like knowledge perception a summated scale. It is based on a number of items on the post-questionnaire (Level 7). I reasoned and therefore expected that this outcome would correlate with the other two main outcomes, because if participants are positive about the training, it seems more likely that they perceived to have learned (Hypothesis 4.1) and performed well on the sensemaking test (Hypothesis 4.2).

Judgment is related to post-knowledge perception,  $r = .29$ ,  $p = .001$ , and post-sensemaking performance,  $r = .37$ ,  $p < .001$ . However, as Figure 11.1 illustrates, these relationships are mediated by the game score. For this reason we should reject the hypotheses.

This is an unexpected outcome too, for which Figure 11.1 allows for possible explanations. The first has to do with its mediation by the game score. If participants do well in the game, they appreciate it more, because as I will elaborate on in a



**Fig. 11.2** Boxplots of the pre- and post-sensemaking performance scores before and after the training ( $N = 130$ ). These are the raw accuracy scores. Participant #149 is an expert employee and Participant #39 a volunteer with a civil engineering background and job

moment, the game score does affect the two main outcomes. They retrieved more value out of the game and appraised it higher.

Another explanation is that some participants appreciated the game-based training more than others. These are the ones with a positive game attitude. This tells us that although some may like it more than others, it does not mean that the intervention does not impact them equally well.

These explanations are confirmed when we consider the appraisal component, which is a component based on the responses on the game questionnaire (Level 5). This component indicates how valuable an exercise was to the player—how players assessed an exercise in terms of enjoyment, realism, and learning. The total appraisal component, which is a summation of the individual appraisal components, has only a unique and strong relationship with judgment,  $r = .59$ ,  $p < .001$ . Its correlations with the two main outcomes are just like judgment mediated by the game score.

The game may have an impact on communication (Hypothesis 5), which I consider a secondary outcome—an advantageous but not indispensable outcome. To consider the impact on communication I first looked into participants' vocabulary on the pre- and sensemaking test (Level 8). The assumption here is that when participants have a shared vocabulary, their communication improves (Level 2).

The vocabulary associated with the game almost tripled (from 9% to 26%). Knowing that consistently 29% of the words participants used were frequent words,



such as “the, is, from, that, and,” we see that a large part of the vocabulary draws from the same source, suggesting that communication improves.

To consider the impact on communication I also looked into two other secondary outcomes: word count and dispersion. Assuming that participants acquired an inspection vocabulary and became more accurate in making sense of failures, I expected that they would need fewer words to describe what they see. This increases communication efficiency. The same line of reasoning explains why I expected that the variety of responses would decrease too. Considering word count (Hypothesis 6.1), the post-training word count did decrease compared to the pre-training word count,  $t(124) = 6.35$ ,  $p < .001$ ,  $r = .50$ . However, word count positively correlates with sensemaking performance (Fig. 11.1), and even more so post-training, suggesting that those who are more accurate in making sense of the failures use more words compared to others. These two results do not bite each other. Generally all participants decreased their word count, but the better performers have a richer vocabulary and understanding that they used to express what they saw on the pictures. What confirms this idea is that the “super experts” used comparatively many more words to describe the failures (Level 10).

The results with dispersion are somewhat consistent (Hypothesis 6.2). Although we find less variation in what participants reported, a decreased dispersion is only noticeable with the observing (15%) and diagnosing (28%) learning objectives. This means that in telling others what they see, less variation occurs, but in describing what they need to pay attention to it does not. One possible explanation is again that participants achieved a richer understanding. Whereas beforehand they knew little about what to mention, they learned about various possibilities after the training.

Throughout the training/evaluation I measured other secondary outcomes about which I did not hypothesize and for which I cannot provide conclusive evidence. In sum, what else became clear is that playing increased player’s awareness—about the impact of failures and the difficulty in reporting them (Level 7). Some also became much more aware of their surroundings (Level 10). They started to look differently at their environment. This was not true for everybody. For example, this effect was not evident among participants who live close to levees or are involved with levees professionally. Participants’ confidence seemed to have increased too. They found inspecting much more of a routine and provided comments that they are ready for some failures and want some action.

Thus, the game-based training clearly had *some* impact. It positively influenced the two main outcomes and exerted considerable positive effects on the secondary outcomes.

## ***The Contributing Factors***

The previous hypotheses were about the outcomes especially, how the game influenced these and how they relate to each other. I further hypothesized about a number of variables that could affect the outcomes. In Level 3 I mentioned it is useful to

make a distinction into “moderators” and “mediators” to discuss these types of variables. Moderator variables influence the direction and/or strength of a relationship between two other variables; mediator variables explain it (Baron & Kenny, 1986). If we would control for the mediator variable, the relationship between the two other variables would disappear. With a moderator, the relationship will remain; only its direction and/or strength will change.

Similar to the outcomes, I had some ideas up front about possible moderating/-mediating variables. Because of the exploratory nature of this research I considered a number of additional variables, such as education. The first moderator I will consider is the one based on the evaluation principle “practice makes perfect.”

### Number of exercises hit the mark

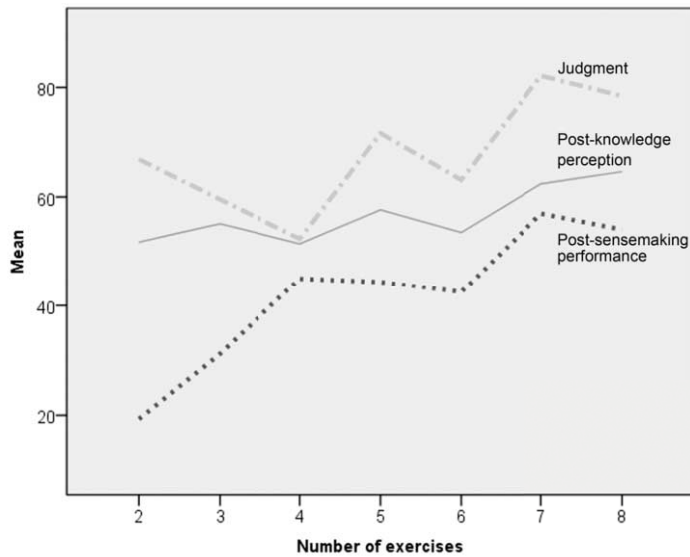
When setting up the training one of the concerns was the amount of exercises (Hypothesis 7). With too few exercises, the training might not reach its purpose. It is a training game and the material needs rehearsal. I could not do too many exercises on the other hand, because that would require too much time and effort on behalf of the participants. Eventually I decided on a total of eight exercises—two per week at home and one per meeting (Level 3). But did all those exercises matter?

To answer this, it would have been better to manipulate the number of exercises and then see what the effects are. Of the 130 participants who completed the post-sensemaking test 98 played all eight exercises. That is 75% of the sample and an important reason why we need to be cautious. Fortunately, the remaining participants varied in the number of exercises they played. This means we are able to get some idea on the influence on the number of exercises with regards to the three main outcome variables.

The results are displayed in Figure 11.3, which demonstrates a trend that the more exercises one plays, the higher the scores are on how they judged the training, perceived to have learned, and performed on the sensemaking test. Clearest are the results with the test. This is the most important indicator and here we see that playing more exercises matters. Correlations confirm this: the exercise number relates to judgment,  $r = .30$ ,  $p = .001$ ; to knowledge perception,  $r = .31$ ,  $p = .001$ ; and to the test,  $r = .50$ ,  $p < .001$ .<sup>2</sup> Practice does make perfect, as the patrollers frequently stated.

But how many exercises are really needed? This data suggests that all exercises helped and that if we extrapolate the data, playing even more exercises will lead to better results. However, in considering the appraisal component, the game’s use appeared to peak with the fourth home exercise (i.e., Exercise 5 in Fig. 11.3) and subsequently flattened or declined (Level 5), suggesting that for players the value of playing diminished.

<sup>2</sup> A MANOVA (using Pillai’s trace and the number of exercises as covariate) shows that the number of exercises matters too,  $V = 0.26$ ,  $F(3,112) = 12.8$ ,  $p < .001$ . This confirms that the number of exercises is predictive for the outcomes.



**Fig. 11.3** The average performance on the three outcome variables in percentages by the number of exercises played ( $N = 130$ )

The combination of these two results suggests that although for players it may have become a kind of “routine,” the extra exercises were still useful for developing a deeper understanding. But it also seems that if more exercises will be added, better results are not necessarily achieved. We can therefore reach the same conclusion as during the discussion (Level 9). Some participants said the number was excessive while others expressed interest in playing more, but the overall conclusion was that the number was ideal (Level 9), thus “hitting the mark.”

### The average game score is king

I decided to pick one indicator from playing the game: the average game score (shown as “game score” in Fig. 11.1).<sup>3</sup> The average game score is the average of the total scores of each exercise a participant played. I picked this variable because it is the highest aggregated quantification of how participants played the game.

I expected that the game scores would moderate the outcomes (Hypothesis 8), because the variable is a composite consisting of various game objectives and these game objectives align with the learning objectives. Some games have scores that are unrelated to the subject matter, unlike *Levee Patroller*. The game scores are only dependent on whether or not players find the failures and are weighted. This

<sup>3</sup> The final game score could also be taken as indicator. This score relates even a bit stronger to the post-sensemaking test,  $r = .75$ ,  $p < .001$ .

makes them different from how for example the sensemaking test is scored (although participants lost many points too if they incorrectly suggested a picture was a non-failure).

However, it turns out that how participants performed in the game strongly determined how they judged the game, perceived their own knowledge about levee inspection, and performed on the sensemaking test. What also became clear from discussing the other results and what can be seen in Figure 11.1 is that the game score acts as a mediating variable for many relationships, such as between judgment and post-knowledge perception. This suggests that participants' performance in the game appears to be a major success factor behind the training. The average game score is king and should be treated as such. The explanations and implications of this finding are that

- If participants perform poorly with the game, they will hold more negative attitudes toward it (and vice versa). This does not mean we should make it easier for participants to get high scores, but that we should find ways to help poor performers to improve.
- Either the scores influenced participants in how they rated their knowledge perception and/or playing the game made participants realize what they know. Either way, we see that perceptions were shaped by what happened in the game.
- What is assessed with the sensemaking test and what is taught in the game closely align to each other. It is a proof that the game teaches what it is supposed to and that the game scores are a good indicator for how people would perform in other circumstances.

This is a promising result. Just as flight simulator trainees gain confidence in real-life flying following success in simulations, it provides evidence that we can be confident about patrollers who have successfully finished the game. In addition, if learning objectives are aligned with the game's goals, the game scores are a better performance indicator than a self-assessment.

### **Computers skills are key**

I expected that the game scores would be moderated by the player's computer skills (Hypothesis 9.1) and game skills (Hypothesis 9.2). From Level 7 it became clear that computer skills matter. Just like the game scores they are a central hub within the network of relationships between variables. Participants with good computer skills *a)* play digital games, *b)* have commitment, *c)* are higher educated, *d)* are (relatively) young, and *e)* have higher motivation and expectations about the training.

As expected, higher scores are achieved by those with better computer skills (Fig. 11.1). Knowing that the game scores closely relate to what the game tries to achieve, we find that people with little computer skills are at a huge disadvantage.

The problem with computer skills might be a problem of the present and past, not of the future. Yet especially with this target group that consists of a great number of people who like to be outdoors and work with their hands, it can be expected that

the issue of computer literacy will continue to exist. Computer skills are integral for success and so this should be a concern.

We cannot draw the same conclusion for game skills. Although these do relate to the game scores, this moderation is mediated by the computer skills variable. With this target group playing games does not bring a direct extra advantage. This raises the issue if something like game skills or game literacy exists. Maybe they are just computer skills. However, one must note that if more experienced gamers would have been part of the sample, their game experience may have moderated the game scores.

To confirm this, I also looked into another variable: the dichotomous variable of whether or not participants ever played a First-Person-Shooter (FPS). Although univariate analysis shows that the FPS variable is a good predictor of the average game scores, when the computer skills variable is used as a covariate it turns out to be meaningless. In other words, playing FPS games is mediated by computer skills as well.

### **Other variables are not a barrier**

In discussing the judgment outcome I concluded that how participants judged the game seems to be dependent on their appreciation of the game. Some liked it more than others. Up front I expected this to happen. I thought specifically that participants with a positive game attitude would judge the game-based training higher (Hypothesis 10). In Level 7 we noticed that this is a true assumption. So people who like to play games or are open-minded about it tend to appreciate the training more.

Unlike game attitude I expected that motivation to learn the material of the training program (Hypothesis 11.1) and expectations about it (Hypothesis 11.2) moderate the results on all main outcomes. It first became clear that motivations and expectations can be reduced to the component called “success potential” (Level 7). This component variable concerns the potential for success based on the participants’ initial attitudes toward learning from this game-based training.

This component turns out to be not so influential. It relates to game attitudes, probably because participants who are fond of games have higher expectations about the training, and to computer skills, presumably because participants with better computer skills are more confident about succeeding in this training.

This lack of influence is unexpected. Motivation and expectations are often seen as crucial to the success of an educational or training program. A possible explanation is that the initial success potential is not of much relevance in a game-based training, because the game will trigger participants and thereby change their motivations and expectations. I observed this throughout the training. Participants started liking the game more, something which is also highlighted by the appraisal component.

The commitment variable also fosters this explanation. Before the training commitment to levee inspection is like computer skills a key variable. Those who are more committed to levee inspection are the ones who *a)* have better computer skills,

*b*) have higher motivation and expectations about the training, *c*) perceive to be more knowledgeable, and *d*) are more knowledgeable (as shown by the pre-sensemaking test). In short, it seems that the committed participants were bound to be a lot more successful with the game.

However, the results show otherwise. Commitment has some indirect relationships, but nowhere a strong direct relationship with the outcome variables are visible. This shows that this game as an intervention was able to grab the attention of those who felt less committed.

Further evidence that the initial positioning of participants may not matter concerns their pre-knowledge perception and pre-sensemaking performance. Pre-knowledge perception had no relationship with the average scores and no direct relationship with the post-sensemaking performance. The pre-sensemaking results did. Participants with a high pre-sensemaking performance tended to have high average game scores, indicating that knowledge provides players an advantage in the game.

However, this advantage does not result in getting even better scores on the post-sensemaking test. When controlling for the scores on the pre-sensemaking test, the relationship between the average game scores and post-sensemaking performance is hardly affected,  $pr = .64$ ,  $p < .001$ , and when we control for the scores on the post-sensemaking test to look into the relationship between the pre-sensemaking test scores and the game scores, the relationship is non-existent. That is why a dashed line is portrayed in Figure 11.1. It is a relationship that needs to be acknowledged, but its influence is in sustaining existing knowledge structures and not in providing an additional advantage.

All of this shows that commitment, motivation, expectations, and existing knowledge are not barriers or moderators for success such as computer skills. The game grabs each and every one of the participants—not only the highly committed or motivated. It also teaches something useful to all participants—irrespective of their existing knowledge. This makes the game a perfect instrument for a mixed group of participants.

### **The double-edged sword of education**

However, the previous conclusion is tempered by a number of caveats: first, the need for computer skills, and second, the level of education. Whereas the type of education had no influence before the training (Level 7), it had an influence on how participants performed in the game and on the test. Because education is linked to computer skills as well, I made sure to see this was not mediating the influence. It was not. Education exerts a unique contribution to the performances.

It is also a double-edged sword. Participants with a higher education performed better in the game and performing better in the game leads to better end results on the test. In addition, the level of education turns out to have a direct effect on the test results. This means that besides computer skills, the level of education matters. Reasons why education matters are possibly that

- Higher educated participants are better learners. Throughout an educational process they are able to absorb and understand the information much better compared to lower educated participants. They may also know how to perform better at tests.
- Some of the material and the game itself are quite complex. I noticed that participants had difficulty with the concept of failure mechanisms and also with the logic of the game. Higher educated participants had less trouble with this.
- The game has been developed with too much of an academic mindset. The game has been developed with a team of students, university faculty members, and experts with university degrees. Nobody was involved or consulted on how lower educated people learn.

This is a concern. With the use of these games, the differences become larger and not smaller. Segregation is a result between those who get it and those who do not. On the other hand, performance differences might be unavoidable within a target group with widely differing levels of education and it was clear that the majority was satisfied about the game and so they were satisfied about their own learning experience. The game is and can be valuable for everyone. These results only highlight that participants with a higher level of education derive greater benefits from it. Consequently, I recommend that developers take more care in understanding how less educated players are able to profit from a game experience.

### ***Differences among Sample and Others***

The final set of hypotheses concern looking into characteristics that could make a difference in the outcomes among subgroups of the participants and in comparison to others. Two clear examples are the type of patroller (volunteers vs. regular employee vs. expert employee) and its affiliation (Organization A vs. Organization B. vs. Organization C). I made a further comparison based on age, with students, super experts, and to another group of patrollers who did not participate with the game-based training.

#### **The influence of type and affiliation**

Before the training I was aware that different patroller types exist. I expected that those who are preoccupied with it professionally, the expert employees, would have better results on knowledge perception and sensemaking performance compared to volunteers and regular employees (Hypothesis 12.2). I did not expect differences between volunteers and regular employees (Hypothesis 12.1). Both are novices. I further expected that the learning gains would be higher for volunteers and regular employees, because the game would have little to teach the expert employees (Hypothesis 12.3).

Unfortunately, I recruited from Organizations A and B not so many employees, let alone expert employees (Level 4). About 73% of the sample are volunteers; 18% employees; and 9% expert employees. Another unfortunate distribution happened: Organization C had many more (expert) employees. About 62% of their participants were employees compared to 15% and 34% for Organizations A and B, respectively. So it seems that the participants' affiliation<sup>4</sup> is difficult to disentangle from the type of patroller they are. This is why I decided to consider both variables simultaneously. Table 11.2 and Figure 11.4 show the results.

**Table 11.2** Two-way ANOVA analyses with Type and Affiliation as fixed factors

	<i>F</i>	<i>df</i> <sub>factor</sub>	<i>df</i> <sub>error</sub>	<i>p</i>	$\omega^2$
Pre-knowledge, <i>N</i> = 136					
Type	11.2	2	128	< .001***	.071
Affiliation	0.80	2		.45	0
Affiliation * Type	4.81	3		.003**	.052
Post-knowledge, <i>N</i> = 123					
Type	8.96	2	115	< .001***	.065
Affiliation	0.49	2		.61	0
Affiliation * Type	2.73	3		.05	.028
Pre-sensemaking, <i>N</i> = 140					
Type	4.36	2	132	.015*	.023
Affiliation	7.98	2		.001**	.048
Affiliation * Type	4.50	3		.005**	.048
Post-sensemaking, <i>N</i> = 144					
Type	4.17	2	122	.018*	.025
Affiliation	6.94	2		.001**	0
Affiliation * Type	2	3		.92	.068
Judgment, <i>N</i> = 134					
Type	0.37	2	126	.69	0
Affiliation	3.12	2		.048*	.016
Affiliation * Type	0.89	3		.45	0

*Note.* In line with Field (2005) I use  $\omega^2$  as an effect size measure. Howell (2010) explains how it needs to be calculated (pp. 438–440). If the square root of  $\omega^2$  is taken, we get a value compared to *r*. If  $\omega^2$  is negative, one should report a zero (Meyers, Gamst, & Guarino, 2006, p. 299). It has been suggested that values of .01, .06, and .14 represent small, medium, and large effects, respectively (Kline, 1996).

\*\*\* *p* < .001; \*\* *p* < .01; \* *p* < .05 (two-sided).

Although at first I considered to lump regular and expert employees together, because of the unequal distribution, the results highlight that it is better to keep them separate. On pre-knowledge as well as post-knowledge perception the type of patroller has a significant main effect and the expert patrollers are largely responsible

<sup>4</sup> Affiliation could be considered a “random effect,” because many more organizations exist. It is not a fixed factor. However, for my purposes here it was better to find out what is causing the problems rather than generalizing the findings to all of the water authorities. If I use affiliation as a random effect, the results are harder to interpret and relate to my other findings. This is why I decided to make it fixed factor too.



for this. The Bonferroni post hoc tests revealed that experts have higher scores than regular employees and volunteers (both  $ps < .001$ ). Because the experts were concentrated in Organization C, this affiliation performs better than others (with A,  $p = .003$ ; with B,  $p = .026$ ) and we see a significant main effect of the interaction between the type of patroller and its affiliation. Simple effects analysis confirms these thoughts. The type of patroller only makes a difference at organizations with expert employees.<sup>5</sup>

With the post-knowledge perception, the experts continue to have higher scores than the regular employees and volunteers (both  $ps < .001$ ). This time no clear differences are found between the organizations (with the Bonferroni post hoc tests). Only if I would have considered one-tailed tests, a difference is noticeable between Organizations A and C. The simple effects analysis shows that within Organization C the type of patroller remains of influence, but in Organization B differences disappeared.

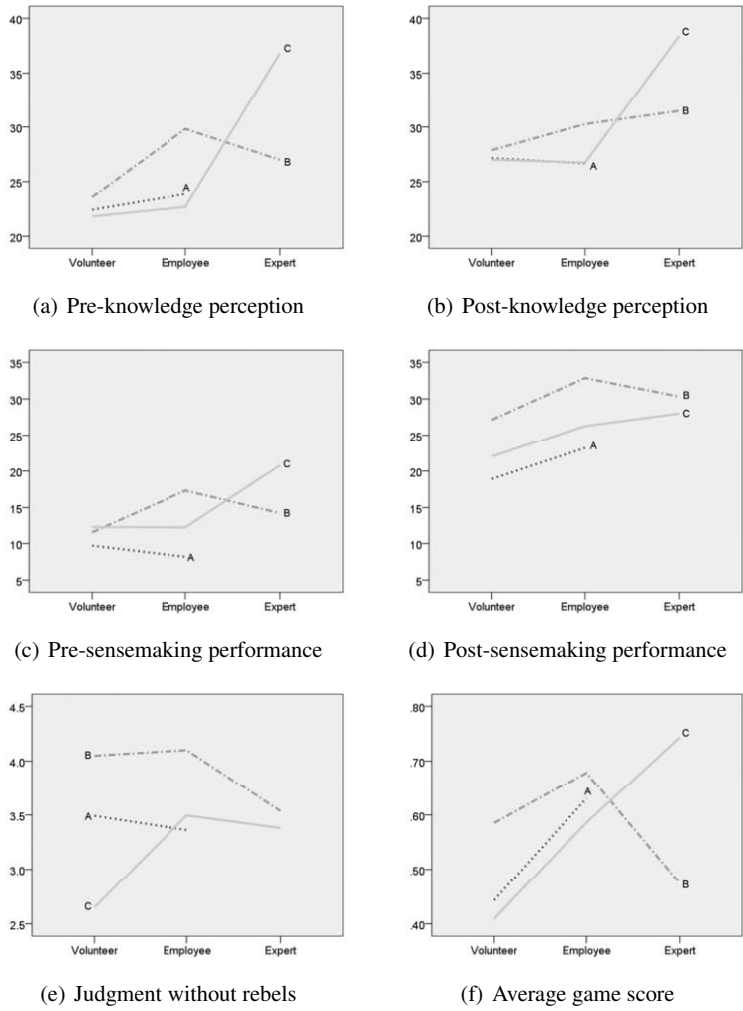
In the sensemaking tests appear a change of influence: not the Type variable but the Affiliation variable is most influential. With the pre-sensemaking performance experts again have higher scores compared to regular employees and volunteers ( $p = .006$ ). With the Affiliation variable we find that Organization C beats B ( $M_C = 15.0$ ,  $SD_C = 5.52$ ;  $M_B = 12.2$ ,  $SD_B = 4.87$ ;  $p = .010$ ) and Organization B, on its turn, beats A ( $M_A = 9.37$ ,  $SD_A = 4.07$ ;  $p < .001$ ). These relations are not completely unexpected based on the descriptions of the participating organizations (Level 4). Organization C has the many experts and Organization B has been training her members the years preceding the training, whereas none of the members at Organization A had participated in an event for years. It might be that with scoring their perception, the patrollers at Organization B underestimated their knowledge compared to those at Organization A. The simple effects analysis is consistent with the results on the pre-knowledge perception. Only at Organization A, where no expert employees participated, no differences are noticeable between the patroller types.

The post-sensemaking performances reveal interesting results. Organization C ( $M_C = 25.4$ ,  $SD_C = 6.87$ ) with all its expert employees is not the best performing organization anymore. This became Organization B ( $M_B = 27.7$ ,  $SD_B = 7.45$ )—where especially her regular employees performed well. Organization A—and this should be no surprise by now—performed worst ( $M_A = 19.4$ ,  $SD_A = 8.49$ ). According to the post hoc tests, performance between Organizations B and C is similar (unless I would have considered an one-tailed test—a planned contrast test shows that B performs better,  $p = .038$ ) and that both outperform Organization A (B with  $p < .001$ ; C with  $p = .008$ ).

Regarding the patroller types, we find no clear differences (with the Bonferroni post hoc tests). The variable still has a main effect on post-sensemaking performance and this is because if I would have considered one-tailed tests, differences would appear between volunteers and experts but also between volunteers and employees. Between employees and experts we can be assured no differences exist.

<sup>5</sup> Simple effects analysis considers the impact of one independent variable on another when considering a dependent variable. Here I continuously considered whether the type of patroller is of influence at each organization for the main outcomes.

Another confirmation of the decreased influence of the Type variable is that the simple effects analysis makes clear no differences exist between patroller types at each organization.



**Fig. 11.4** Comparisons of the Affiliation (A vs. B vs. C) and Type (Volunteer vs. Regular Employee vs. Expert Employee) variables on the outcomes and game score

Consistent with the post-sensemaking performances we find with the Bonferroni post hoc tests that Organization B judged the game-based training more positively than others (with A,  $p = .018$  ; with C,  $p = .002$ ). They did better and liked it better, which is a relationship that is mediated by the game scores as we have seen before. Even after removing the rebel and co-rebel who both judged the training very neg-

actively, Organization B remains more positive.<sup>6</sup> However, this removal does make the main effect of Affiliation insignificant ( $p = .084$ ).

If we look at the gains in perception and sensemaking performance, t-tests comparing the experts on the one hand and the employees and volunteers on the other show that the experts changed less ( $M = 2.74$ ,  $SD = 3.25$ ) than the other types ( $M = 4.13$ ,  $SD = 5.13$ ), but this is not significant. With the sensemaking performance the experts also changed less ( $M = 8.10$ ,  $SD = 6.05$ ) than the other types ( $M = 13.3$ ,  $SD = 7.46$ ). This time it is significant,  $t(123) = 2.12$ ,  $p = .036$ ,  $r = .19$ . This may explain partially why Organization B performed better at the post-sensemaking test: the experts at Organization C did not improve on their sensemaking so much.

To explain why the experts did not improve so much I looked into the game scores, but these show that the experts at Organization C actually score on average the highest (Fig. 11.4). In general the volunteers had lower average scores ( $M = 52$ ,  $SD = 23$ ) compared to the employees ( $M = 62$ ,  $SD = 16$ ) and experts ( $M = 65$ ,  $SD = 22$ ),  $t(85) = 3.27$ ,  $p = .002$ ,  $r = .33$ . With this in mind, a possible explanation for the post-sensemaking performances is that volunteers and employees were more willing and open to accept and use the game's terms and content.

The unequal and small sizes of the groups warrant caution in interpreting the validity of the data. However, these results suggest that

- Employees and volunteers have a similar performance (Hypothesis 12.1). However, the results highlight that the employees had a tendency to perform better. They received significantly higher game scores and achieved a higher yet insignificant higher post-sensemaking performance.
- Expert employees were indeed experts (Hypothesis 12.2). They perceived to know more before and after the training, received the highest game scores in the game, and also performed better with the pre-sensemaking test.
- Playing the game helped to level the various participants on the post-sensemaking test. Differences became less noticeable. In fact, employees tended to perform better than the experts. Experts may have been less willing and open to the game.
- Learning gains were greater for employees and volunteers (Hypothesis 12.3), and more so for sensemaking performance than for perception.
- Previous training and events make a difference. Organizations B and C, which had organized several events performed better on the pre-sensemaking test than Organization A.
- My changes to the training after Organization A had an impact. Although Organization A may have had the least prepared participants, the differences are too large to have just been a cause by this. This confirms especially the importance of the start-meeting (Level 9).
- The process of the training seems influential too. At Organization A the training did not proceed as smooth, because it was the first time it was organized.

---

<sup>6</sup> In Figure 11.4 I removed the rebel and co-rebel, because their inclusion biases how other volunteers perceived the training, which is a perception more aligned with volunteers at the other two organizations.

At Organization C we had a false start. The training was made compulsory and planned during a vacation time. Except for some administrative issues, at Organization B the training went very well and we see that there the training was judged best.

One explanation for some of the above-mentioned results is that the employees and experts were younger than the volunteers,  $t(143) = -3.03, p = .003, r = .25$ . And age, as we will see, makes a difference.

### Older participants are hit twice

I was aware that the average age among patrollers was relatively old. I expected that age would have an influential role with player performance because of computer and/or game skills. I specifically reasoned that younger participants would perform better. To verify this, I compared the patroller game performances with those of students (Hypothesis 13.1) and compared the performance of the “younger” participants with the older ones (Hypothesis 13.2). Based on the sample data I considered anyone younger than 40 years as belonging to the group of younger participants.

It turns out that the youngest group of participants has higher means on post-knowledge perception,  $t(116) = -2.62, p = .010, r = .24$ , on game attitude,  $t(140) = -2.94, p = .010, r = .24$ , and on judgment,  $t(126) = -3.69, p = .010, r = .31$ . On pre-knowledge perception and success potential the means were equal. When we consider the game scores in addition to this, it becomes clear that the younger participant group achieved indeed higher scores,  $t(64) = -7.09, p < .001, r = .66$ . We know now of the critical role of the game scores when it concerns sensemaking performance and so it is should not be a surprise that whereas the younger participants had an equal pre-sensemaking performance, their post-sensemaking performance was greater,  $t(124) = -2.69, p = .008, r = .24$ .

Comparable to the youngest group of participants, the students have higher means on game attitude,  $t(191) = 7.67, p < .001, r = .49$ , and achieved higher scores on the start-exercise,  $t(56) = 6.75, p < .001, r = .67$ . A number of students played more than one exercise and filled out the post-sensemaking test. From their average performance we find an indication that students get more out of the game with less effort. Their performance was at least up to par with the patrollers (Level 10).

If we consider the age variable in general and relate this to the other variables we find something remarkable. In the previous model the age of participants was negatively related to judgment (Fig. 7.1). This was not mediated by the computer skills variable with which age strongly relates to. Another factor was at play, and I reasoned older participants could possibly be more critical of new media products, such as games.

The new model provides clarity (Fig. 11.1). The relationship with how the game is judged is mediated by the average game scores. It appears that age strongly relates to this variable and not to judgment. I thought this must be a mistake, because computer skills relates strongly to the game scores and also to age. But even when

controlling for computer skills, age remains having this strong relationship with the game scores,  $pr = -.41$ ,  $p < .001$ .

Thus, the previous suggestion about criticism seems incorrect. It is related to playing the game and not to how it is judged, but it is unrelated to computer skills and playing games. Although it is true that older participants tend to have lesser computer skills and play fewer games, this relationship is mediated by computer skills.

What could possibly explain this relationship is that participants need to learn how to use this game, which includes but goes beyond learning how to use the controls. Playing a game such as *Levee Patroller* requires to learn how to “read” the environment. For older participants, it requires a way of working and thinking that they are not accustomed to, compounding their limited computer skills. In addition, the design of games for older adults requires different requirements for reasons other than computer skills, which have not been taken into account in the design of the game (Gamberini et al., 2006; IJsselsteijn, Nap, de Kort, & Poels, 2007).

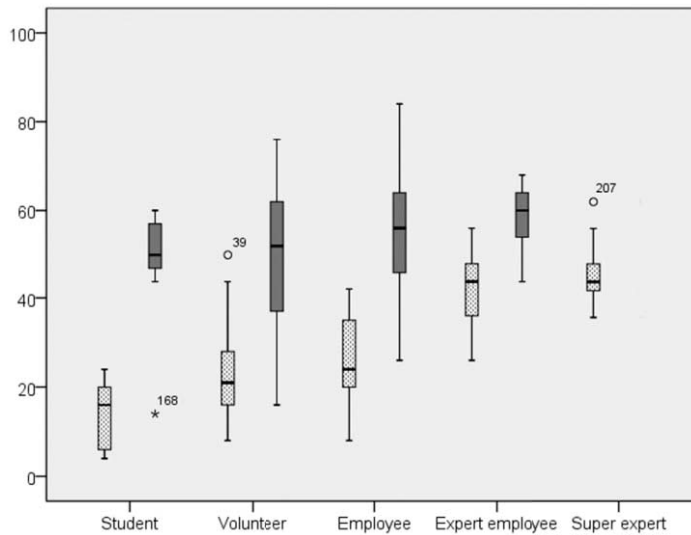
### **Patrollers become professional Hans Brinkers**

Students as well as super experts engaged with the sensemaking test too. In this manner I was able to benchmark the results with the patrollers. The student results would tell me to what extent patrollers know more than novices. I expected they would (Hypothesis 14.1). I included the super experts to benchmark the progress after the training. If the patrollers approximate the performance level of the super experts they can be considered “professional Hans Brinkers.” Compared to the super experts I expected that the patrollers would perform less at first (Hypothesis 14.2) and—being optimistic—I hypothesized that playing the game would make them approximate their performance level (Hypothesis 14.3).

I may have underestimated the potential power of the game, because after the training the employees even outperformed the super experts,  $t(45) = 3.45$ ,  $p = .001$ ,  $r = .46$ . The patrollers in general performed equal to the super experts. Before the training the super experts clearly outperformed the patrollers. The patrollers, on their turn, performed better than the students. This confirms all of the hypotheses. Figure 11.5 shows how the students, volunteers, employees, and (super) experts performed among each other on the pre- and post-test. The one super expert outlier (#207) was the only one able to compete with the top 50% of the virtually trained employees.

### **Conscious incompetent patrollers**

A final comparison was made with the use of field exercises (Level 10). I compared a Game Group that participated with the game-based training and a Control Group who did not. I did not consider their sensemaking performance, but I did look into communication and confidence among others. About these two outcomes in particu-



**Fig. 11.5** Boxplots of the percentual scores on the pre-test (colored in light grey and with pattern) and post-test (colored in dark gray) by the students, volunteers, employees, expert employees, and super experts

lar I expected that the Game Group would score higher. With many hours of training I assumed the Game Group should have more confidence in dealing with failures (Hypothesis 15.1). Because I assumed the game would give participants a shared mental model of failures and a shared vocabulary I hypothesized that they would communicate better—with each other and the Action Center (Hypothesis 15.2).

Unlike my positive findings about communication with the sensemaking test, with the field exercises no differences were found between the groups' perceptions about their communication within their team and with the Action Center. However, I did observe communication issues between the Game Group and the Action Center, because the Game Group participants used words that the Action Center was not familiar with. This highlights that if others in the organization are not familiar with the game's vocabulary, the communication is worse off. This also highlights that the Game Group continued to use the game's vocabulary, something I noticed during the post-interviews as well. This means these findings provide supportive evidence for vocabulary use (Hypothesis 5), but we cannot conclude that the Game Group communicated better.

In addition, surprisingly the Game Group felt less confident about the field exercise than the Control Group (i.e., less prepared and less knowledgeable). Although other variables may have been influential here, what may have happened is that the Game Group members were "conscious incompetent" about their knowledge and skills and the Control group "unconscious incompetent." Knowledge and skills

have to be refreshed—something the participants repeatedly indicated during the discussions—and the Game Group maybe felt like they should have had a refreshment a half year after the game-based training. Unlike the Control Group, they were aware that they needed to know more.

## Providing Additional Perspectives

The consideration of the working hypotheses helped to clarify the game's effectiveness and what factors contributed to it. Such clarification was achieved with the quantitative approach (QUAN). However, although clarifying and providing for evidence in a field eager to have some is useful, this approach does not provide the understanding that is needed to push the field forward. Knowing that *Levee Patroller* works is great as a showcase, but how will others learn from this? And how can we improve from what is achieved with this game? This required an alternative approach, one that is exploratory and that opens up the black-box of what happened during the training/evaluation. To achieve this is why I made use of a qualitative approach (QUAL) in parallel to the quantitative one.

With the qualitative approach my aim was to provide for a thick description of how participants experienced the training/evaluation and played the game. This would lead to a detailed understanding of what exactly happened throughout and shed light onto the results attained with the quantitative approach. It would also enhance transfer of this case to another (Firestone, 1993). This makes *Levee Patroller* more than a showcase. It becomes a case to learn from—a possible best practice.

In addition to the descriptions provided in the previous levels, the QUAL approach provides for a number of additional perspectives onto the QUAN results just described.

### *The Glass Is Half Full*

From the QUAN approach we determined that participants improved in terms of knowledge perception and sensemaking performance, suggesting they learned from the game. We further came to find out that game scores are an important contributing factor: the better players are in the game, the more they seem to learn from it.

The results of the QUAL approach indicate that no conceivable learning process happened in the game, that player behavior is hard to change, and that for some failures/objectives much more improvement is possible. So it seems the glass is half full: we received positive effects but the game could have a much greater impact.

### **No conceivable learning process**

What we expect to happen in a game is a regular learning process, one that expects learners to become better and better over time. Whereas they may first fail on something, they will do it right the next time. They will learn from their mistakes and move on. However, when we neglect the game scores but focus on the failure correctness scores instead, which show how well players performed in reporting a failure once they found this, we find that

- Performance is good.
- Improvement is incremental and subtle.
- Performance was good from the start.
- Performance improved game-wise and not failure-wise.

Players hardly engaged into a learning process as expected. Their performance was good, even from the start. They further improved only incrementally, highlighting that little learning took place in how to report a failure. This is surprising and may suggest that the results that were achieved with the game-based training could be so much better if the design of the game were to be improved. But even if better results could be achieved, this performance in the game remains a puzzlement, because it does not explain why the training had a positive impact on the training outcomes and that for some it had a higher impact than for others. Several related and supporting hypotheses could explain this:

1. Making sense of virtual failures is easy.
2. The reporting system in the game helps players in making sense of virtual failures.
3. Players learned from the mere exposure to the material.
4. Players who performed better in the game and on the tests have more elaborate mental models. This allows them to deal with failures quickly in the game, giving them more time to deal with others, and their elaborate mental models help to better make sense of failures when no support is available.

Of course, many players did not label the game as “easy,” but that relates to the controls, not to the actual content. If we consider for example the amount of non-failures, we see that players found it easy to distinguish failures from non-failures. The idea of mere exposure concurs with the notion of the game being seen as a “repetition instrument” (Level 9). By repeating confrontations with failures, elaborate mental models are constructed that players can use in the real world.

Nevertheless, the impact could have been greater if players did have a strong learning process in the game related to the content of the game. The next insight confirms this sentiment.



## Hard to change

Players were consistent and persistent in what they did in the game. What players reported the first time around they adhered to, even if feedback instructed otherwise. The following insights support this:

- Many players did not know how to use the feedback menu at the end of the sessions. An interface problem also led many participants to skip on this feedback menu. Although this is definitely not true for every participant, I observed a good number of participants who were not so interested in reading this end-of-exercise feedback. They seemed to be relieved or glad it was over and wanted to move on with something else. Once players are disengaged with the game, they must have a strong motivation to read the feedback carefully. A better approach might be to somehow integrate the feedback as part of the exercise.
- Some participants were satisfied with achieving a sufficient score (i.e., 55%). They did not feel any incentives to try harder in improving their reporting skills. Although this might be a matter of personal motivation, with the right game design mechanisms such participants may be stimulated to try harder.
- Participants disagreed with how the game viewed the failures. It is positive that participants are critical of what the game offers, but a too critical approach will inhibit the learning process (i.e., suffering from virtualphobia). What may help here is to embed the game with other instructional material. This will make the game content more believable to those who have a natural tendency to distrust anything virtual.

For each of the insights I provided a possible solution. Many more are imaginable. The point is that if player behavior can be positively influenced, this may lead to a greater impact (with the assumption that what the game teaches is correct). What is most critical, however, is that the “first blow is half the battle.” The first moment participants make sense of an object is extremely important. They made their choices and constructed or adapted their mental model related to the failure. The next time they will see this failure they will use their earlier choices and this model in making sense out of it.

This idea is consistent with ideas about sensemaking (Weick, 1995). This especially happens with unknowns—when people have no idea what they are dealing with. Then they start to make sense. For example, the first time a person goes to his new office he has to make sense of his traveling route. He or she will make a conscious effort and will make many decisions in choosing a route. After a route is picked, traveling to work becomes a routine. People do not think about it anymore. This appears to be the same with dealing with virtual failures and participants seemed to agree with this. They indicated that to them inspecting failures has become a routine.

All of this means that designers need to think about how they will facilitate the right construction of mental models the first time players encounter a phenomenon and that players adapt their behavior based on feedback in the game. If they get it right, outcomes will likely be greater.

### **Not lump everything together**

Upon dissecting the various learning objectives and considering the performance on the pictures on the sensemaking test and on the failures in the game something becomes very clear: performance differs and this performance is consistent between the various analyses, suggesting that the performance is inherent to a learning objective or failure. For example, at both the sensemaking test and in the game the watery slope was the poorest performed on failure and stone damage the best. In addition, performance on reporting was poor everywhere and for diagnosing a dramatic improvement was noted. From these observations two possible conclusions can be drawn:

1. The game is more effective in some failures/objectives than others. As a consequence, its effectiveness could be improved by considering why certain elements are less successful than others.
2. Some failures/objectives are more difficult than others and, therefore, have been less successful in teaching than others. As a consequence, more support should be provided for teaching these “difficult” failures/objectives.

As a support for the first conclusion I found patterns that suggest that players improved especially on items that are explicitly part of the game. With this I mean that the game’s content is connected meaningfully to its mechanisms and visualizations. For example, although the stone damage might be an easy failure, I frequently heard players say that they had to deal with this failure first. Otherwise the water level would rise and it would be difficult to find this failure. This game mechanism and implementation forced players to pay attention to this particular failure and this focus may have resulted in an improved performance compared to other failures. I further observed that some labels were easier to adapt to and that certain reporting elements were remembered better, which provides for support of the second conclusion.

The worth of each result needs to be contested too. Validation of the results provided the insight that the results on especially the diagnosing learning objective are inflated. But it is most important to not lump everything together and conclude that the game was effective. The game was arguably less good in teaching about assessing and reporting failures and the watery slope failure is one that deserves attention in how it has been implemented.

### ***Turned into Reflective Practitioners***

Although the QUAL findings suggest more improvement is possible, they also provide for further evidence that the game helped to turn participants into “professional Hans Brinkers.” The responses on the game questionnaire suggest that participants engaged in much meta-cognitive thinking about levee inspection and on the field exercise it was noticeable that the Game Group was able to focus on the substance

as well as form of the training, unlike the Control Group who only made remarks about the equipment and logistics of the training. I noticed this during the discussions too. Participants were not only reflecting about the game, but actually about the practice of inspecting levees:

*DB3-#1:* Or you have to go with three people [to effectively inspect a levee].

*DB3-#2:* In fact, if you want to properly inspect a levee you should walk two or three times. At the outer slope you have to look for those pitching stones but you also have to look there at the bottom of the ditch.

*DB3-#3:* To see those pitching stones, you will immediately lose your pal.

*DB3-#4:* Is he not leashed?

*DB3-#1:* That is what I mean. If you walk with three people, you can inspect the whole width of a levee and such a levee is quite wide.

*DB3-#5:* You now have to walk back and forth.

*DB3-#6:* Well, I would walk differently next time.

*DB3-#7:* Agreed.

*DB3-#7:* So what is the ideal way of inspecting?

In other words, participants started to think and reflect more about their practice, thereby becoming *reflective practitioners* (Schön, 1983, 1987). The game gave them the capacity to reflect on action and engage in a process of continuous learning, which is one of the defining characteristics of a professional practice.

## ***Gaming is Fragile***

Designing and implementing a game needs to be done with much care. From the QUAL approach it becomes clear why: gaming is dependent on the real world and on the interaction between players and the game. Therefore, we have to deal with two translations, one from the real world to the game and then one from the game to the player. This makes gaming fragile and two of the insights confirm this fragility.

### **If reality is broken, so is the game**

Many participants seem to blame the game environment for not being what the real world is like, which is why in Level 5 I wrote that “Reality is not broken, it is much better.” Many players, especially those suffering from virtualphobia, made various assertions that reality is better than how things work in the game. Although they might be right about a number of aspects, what struck me is that it seems easy but incorrectly so to blame the game. Often what is in the game is taken from reality. If this reality is broken, so is the game.

A relevant example concerns assessing. Many players complained about the game’s system. They did not agree with it. But it seems nobody agrees with it. Even the super experts disagreed among each other about how to assess a failure. If no consensus exist in reality, how could the game provide for a solution?

Another example are some of the ambiguities in reporting, such as when the water velocity is considered fast or slow. No rules of thumb are available and patrollers need to report this. Here too players complained that the game is not clear enough. A notable number of the “errors” made in the game are a result of these types of ambiguity. This has two related consequences:

1. If reality is broken, teaching with a game is hard.
2. The game could do better, but this requires clarity (if at all possible) from the real world.

It is questionable whether clarity will be achieved at all and so this necessitates a stronger embedding with other instructional material as well, including an elaborate debriefing on the game experience. But what this insight especially tells us is that we cannot expect a game to provide for miracles if we do not know the answers in the real world.

### **A game is a sensitive medium**

It was further noticeable that players experienced many frustrations throughout the game, from not finding failures to dealing with a mouse pointer. Although some of these frustrations may have been a result of a lack in computer skills by the participants, they played a central role in how players experienced the game. The majority of the gameplay responses concerned frustrations. As I concluded in Level 5, from this we can learn that a game is a sensitive medium and that maybe more so than with other instructional materials more care is needed in making sure that it has no errors and that its use is intuitive.

Special care should especially be taken when inserting humorous elements. Players took the game very serious and wanted to be treated seriously in return. A number of fun elements were not appreciated.

### ***Still for Some and Not for Others***

The results from the QUAL approach suggest that we need to be careful in drawing strong conclusions based on the QUAN results. I already argued that because education plays an important role, it does not mean that the game is not valuable for lower educated participants. One insight, this time pertaining to age, also confirms that we should not exclude people because of the QUAN results. Another insight suggests the opposite. Although certain people may have achieved the desired results, they prefer to learn by other means than playing a game.

### **Also for this generation**

Based on the QUAN approach, one may conclude this is a tool for especially the future generation of patrollers, because the current one is with an average of 47.6 years relatively old and the findings confirm this. However, if we consider the “outrage” I received when I posed this during the discussion, then the answer was clear-cut: this was also for them.

Of course, people may not want to admit this, because they are too prideful, or they like to get involved with the newest technologies too. But still, if we consider the clear majority of people who were satisfied about the training, we can conclude that although certain people may not benefit from it as much as others, it could still be used for them. It requires some effort, but as we have seen, the majority is willing to invest this into it and it still pays off. And how else are they going to get this practice?

I think the answers should be looked for into how we could help these participants that are at a disadvantage rather than concluding it may not be for them. Because of the experimental setting the difficulty was fixed, but we could imagine changing the difficulty depending on the skills of the player. Another possibility is to discover what may be so difficult for some and not for others and make a difference here too.

For sure I encountered some extreme participants who are unwilling to learn or simply cannot learn it and participants that for other reasons, such as cybersickness, could not participate with this training. This makes it most certainly a restrictive tool, more so for this generation than others. But it does not mean that this is a tool for the youth only.

### **Not everyone’s cup of tea**

I noticed throughout the training that learning by playing is not everyone’s cup of tea. Some of this might be attributed to learning styles. The comments made by those who did not like how they learned with this game seemed to point in the direction of a dislike of trial-and-error learning or discovery learning. They preferred to first learn everything from books or other sources and then apply this somewhere.

The game’s design was also not up to everyone’s taste. Some participants did not like the freedom they received in the game. They rather received clear instructions. This preference might relate to the level of education. Then we have participants who simply dislike computers or are wary of them (i.e., virtualphobia). The first left the training quickly; the second were just critical about it.

### **Reflecting on the Puzzle**

Some reflection is warranted pertaining to the results and the research approach. In this research a very special target group was approached about a very special topic

with a very unique game. Therefore, any generalization should be made with care. One can theorize about the target group in two opposite directions:

1. This is a very difficult target group. They lack computer skills, have little experience with playing games, and are relatively old. Achieving this success with this “extreme” target group suggests even better results are more likely to be achieved with other target groups.
2. The success achieved with this target group was because of their lack in computer skills, little experience with playing games, and the little attention patrollers receive in general. Playing the game was a completely new experience to them and this raised their curiosity and interest to participate. In addition, the special care and attention given may have led to a *Hawthorne effect*: any other three-week activity could have led to positive results too.

Regarding the subject of levee inspection, this practice does share similarities to others. Fire fighters, police officers, emergency personnel, plant operators, and safety inspectors among many others have to make sense of risks too. Important differences do exist, such as that those practitioners are likely to have more experience with the risks they are involved with. Based on the insights with expert employees with levee inspection, we can theorize that such more experienced practitioners are more critical about games, retrieve less value from them, and are less susceptible to their content. Games of much higher quality seem needed for those target groups.

In addition, these other practitioners may have a much stronger organizational affiliation. Many of the patrollers are volunteers and most of the employees are infrequently preoccupied with the inspection organization. This means that the “organizational baggage” (i.e., organizational culture, vocabulary, standards, and procedures) participants bring to the table is little. With other target groups this baggage might be considerably more and this probably influences the outcomes. This would necessitate to consider the organization much more in-depth than what has been done here.

Existing organizational baggage may also limit a game’s impact. Although levee inspection exists for centuries, little is still known about levee failures and many procedures and structures do not exist. For example, the pictures used in the sense-making test take up most of the failure visualizations that exist and so the game’s failure visualizations contribute significantly to the field of levee inspection. For organizations that drive on a large body of knowledge and use tested procedures, the contribution of a game will not be as influential.

Finally, the type of game limits generalization. *Levee Patroller* is a singleplayer 3D first-person game (in addition to its design and the training setup). With other game types different factors might be of influence and importance. For example, if a game has game scores that do not reflect its learning objectives, the average game score is likely not as critical as in this study.

In terms of the research approach pursued, the following needs to be mentioned:

- The pre-interviews could have possibly contaminated results on other methods because they could be viewed as an intervention. Unlike the other participants,

the pre-interviewees already thought about failures and failure mechanisms and discussed this with me. I checked to see if the pre-interviewees achieved different results on the knowledge perception and sensemaking performance outcomes and they did not. However, some contamination may have still occurred.

- The post-sensemaking performance might suffer from a *test-retest bias*. The participants may have learned how to fill out the test the first time and may have been more economical with their words because of this. Considering the three weeks lapse in between both tests I thought this bias would be minimal.
- The sensemaking test in general should be subjected to scrutiny, especially because it is based on the coding of just a single person. However, the interviews, the strong correlation with the game scores, and the results with students and super experts validate its results.

I realize that many more analyses are possible with the retrieved data and that the results can be approached from many more perspectives. But for the purposes of this investigation the most important answers were found, which I will summarize right now.

## Lessons Learned

Did the game work? We can now confidently say it did. As a matter of fact, the game-based training made an impact on aspects that it did not aim for, considering the number of secondary outcomes it affected. Also, the success of the training in terms of participation was beyond expectations. But most importantly, it increased participants' inspection knowledge and skills. In sum, these are the major findings for concluding that the game can be considered effective:

- Four-fifth of the participants played almost everything.
- Participants indicated to have learned from it on all matters that are important to the inspection.
- Participants became able to make sense of failures—whether virtual or real—more accurately and with fewer words.
- Participants were more knowledgeable than students before the training. After the training they approximated the level of sensemaking by (super) experts.
- Participants used far more the same vocabulary and among participants the variation in answers decreased.
- Participants became more aware about the practice of levee inspection and the need for training. They also gained more self-confidence about inspecting failures and became motivated to learn more.
- Participants judged the game positively and found the game useful, also a half year after the training ended, and indicated that the training setup was perfect as is.
- Participants started to think and reflect more about their practice—they became reflective practitioners.

It also became clear that results varied widely. After dissecting the results, it appeared that a complex network of factors play a role in the game's effectiveness. Surprisingly, it is not influenced by a positive game attitude, commitment to inspection, motivation to learn, and high expectations about the training. These factors determined the appraisal of the training but not its outcomes. This suggests that the game was able to enthuse participants about levee inspection—except for a few rebels and strongly virtualphobic participants.

Factors that were of importance are computer skills, age, and education, which are factors that influence each other but also have each a unique contribution to the outcomes. Highly skilled and educated younger participants were able to retrieve more out of the training than others. This suggests that the game would be more successful with a target group with these characteristics. The study with students confirmed this. With much less training they were able to attain similar results.

However, the most important factor concerns how the game is played. Of course, how the game is played is partially determined by other factors, such as computer skills, age, and education, but these factors do not explain for all of its variance. In examining the factors, the average game score was used and this turned out to have strong relationships with all of the outcomes of the training, indicating it is a crucial factor: better players learn more. What helps in achieving a higher score is playing more. The number of exercises played had an influence on the game score and consequently on the outcomes.

Other factors concern participants' initial expertise, the organization the participant is affiliated with, and the training setup. Expert participants performed well but were less susceptible in reconstructing their knowledge based on the virtual experiences. Differences were observed between the participating organizations too. Some of these differences could be attributed to how the training was set up. The "big pond" at one of the organizations was considerably less effective, leading its affiliated members to perform less than the members of the two other organizations.

Many more factors might be of influence and these could provide for different conclusions. If I only considered the self-assessment results, the importance of education was underestimated and of judgment overestimated. Therefore, we need to be prudent about self-assessment results and about concluding causal relationships in the use of games. Another reason to be prudent is that the game might still be valuable for those who seem to benefit less from playing the game.

Exploration into how players actually played the game learned us that the game maybe reached half of its potential. Participants may have learned from some of the feedback, but generally they became better in playing the game and not in making sense of virtual risks. This tells us that in terms of effective design, much is still to be gained and especially because how players perform in the game is most crucial.

In improving the design, the fragility of the medium needs to be taken into account. Many issues people had with the game were because of a lack of answers in the real world. Also, players experienced many frustrations throughout the game and reducing these would likely make a major impact on the experience.

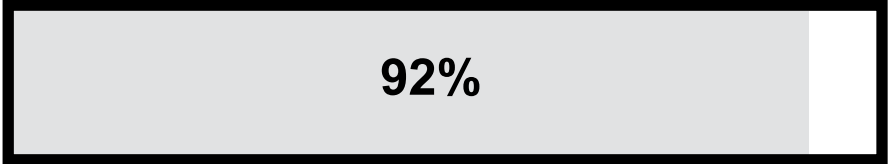


## Level 12

# The Future of Game-Based Training

*I would have never guessed that gaming would help me with something!—Participant at a session at Organization A*

*You learn more from this in three weeks than being with the levee inspection for 30 years—Participant #6*



92%

This is the end of another story of *Levee Patroller*. The first dealt with the design and was told to learn about the design of games with a serious purpose (Harteveld, 2011). This story is about its evaluation and is told to investigate the use and effectiveness of digital games to train practitioners.

Game-based training has apparent potential, but so far adoption remains limited and especially within the domain of safety and crisis response. However, gaming is promising in this domain in particular. A game allows players to deal with risks in a (relatively) safe and flexible environment and use that experience when it really matters. In other words: a game helps players to make sense of real risks by making sense of virtual ones first.

This book focused on the most pressing issue: does it actually work? The aim was a comprehensive and rigorous examination of how *Levee Patroller* was used to increase our understanding of game-based training. A three-week training with *Levee Patroller* was set up at three organizations and a total of 147 participants were involved. The training was evaluated by combining a quasi-experimental design with mixed methods research. The results provides for clear evidence that games can be used to help practitioners make sense of risks. But they also suggest that much improvement is possible, which highlights that important work is to be done in the nearby future—with *Levee Patroller* but with game-based training in general. This level is devoted to discussing this future.

The goals of this level are to describe

- The idea of *sensegaming*;

- The top ten lessons learned for future training/evaluations;
- Recommendations for future research and practice; and
- The future of *Levee Patroller*.

## Toward Sensegaming

The main hypothesis of the investigation into game-based training with *Levee Patroller* was to consider if participants will have increased their knowledge and skills and are able to communicate better with each other. It was further argued for that the driving force behind the acquisition of knowledge and skills concerns sensemaking. Sensemaking is a process that creates order from chaos, happens when challenged, leads to the (re-)construction of knowledge, and does not occur in a vacuum.

When sensemaking is provided within the medium of a game, the game provides a structure to create order, enables the (re-)construction of knowledge by discovery and trial-and-error learning, and challenges the player with game elements that need to be made sense of. The game experiences and results may not be completely similar because players differ from one another. Players' history, culture, identity, and other factors play a role in how knowledge is (re)constructed.

On all outcomes the game had a significant effect and this made it an effective tool for what it is designed for. As we have seen the game made even an impact beyond for what it is designed for. Several secondary outcomes were also affected. To this end the game can be considered a success and therefore the driving force behind it, which concerns its sensemaking process.

However, further exploration into how players actually played the game learned us that the game maybe reached half of its potential. Participants may have learned by being engaged with the subject matter, but generally they became better in playing the game and not in making sense of virtual risks.

These positive results confirm that games are a potential powerful tool to enable players to make sense of phenomena. They also tell us that much more work is to be done to increase its potential. One such step is to explore the concept of sense-making further with regards to games. This connection between sensemaking and gaming is what I call *sensegaming*:

Sensemaking + gaming = sensegaming

This sensegaming can be defined as a process by which players give meaning to a virtual experience, which is carefully designed, with an unknown, rare, or poorly understood phenomenon in order to give meaning to a real experience about this phenomenon. The applicability of sensegaming seems natural to the domain of safety and crisis response, but in general one can think of the following:

- *Gaming sense of the rare and hard to experience*: Similar to levee failures other phenomena may exist that are rare and hard to experience.

- *Gaming sense of the hard to see*: Certain phenomena are hard to see. One can think of physics or phenomena that happen in our galaxy. By playing a game players receive the opportunity to understand these phenomena.
- *Gaming sense of the future*: Games can be applied to explore what the future is like. This use has been practiced for decades (Duke & Geurts, 2004), but by thinking of it in terms of sensegaming it may help in improving its design and understanding its implications.

This book has provided evidence for the potential of sensegaming. Future work will need to provide directions into how it can be improved and extended beyond what has been presented here.

## Advice for Future Training/Evaluations

In addition to exploring the connection between sensemaking and games, this book made especially a contribution in how to set up a training/evaluation. I have a top ten list of advice for designing a game-based evaluation as well as training.

### *Advice for Evaluation*

In creating the evaluation with *Levee Patroller*, I made use of the ten evaluation principles. These principles can be applied for other game evaluations too.

1. *Rome was not built in a day*: Although gaming has a long and rich history, game-based training has not been comprehensively studied until recently. We cannot expect that game-based training works most optimally right away. An understanding of what works and what does not is more fruitful than the evidence itself.
2. *No comparison of apples and oranges*: Although comparisons can be useful, they are not always possible or sensible. In addition, game-based training deserves further investigation in its own right.
3. *More than the tip of the iceberg*: A comprehensive examination needs to go far beyond the standard questionnaires usually associated with game-based evaluations. Proof that it works requires a rigorous evaluation.
4. *The proof of the pudding is in the eating*: Players learn from playing the game, but often the gameplay is treated as a black-box. By not considering the gameplay much valuable information is lost.
5. *The icing and the cake*: In evaluating a game not only obvious design elements should be considered, such as its graphics and controls. The complete design needs to be taken into account. This requires a full understanding of the design by evaluators.

6. *Ain't nothing like the real thing*: If a game is designed for a specific target group, then it should be played with that target group. Playing the game with, for example, students may lead to different outcomes.
7. *Practice makes perfect*: Training requires practice—also with digital games. This means a game needs to be played more than once to see its effects and facilitators/evaluators need to consider this in their setup.
8. *Big fish in a small pond*: The impact of a game involves much more than the quality of the game, although that is a major part (hence it is the big fish). Facilitation, documentation, and other contextual variables matter too.
9. *See the big picture*: Evaluation needs to be comprehensive, but it is impossible to take every aspect into account and to consider those that are in a detailed manner. It is a choice of breadth over depth. This also means the research should not focus all its attention on one or two aspects.
10. *It takes two to tango*: It takes two to tango, but it takes three to design a meaningful game. It also takes three to evaluate a game. Game evaluation needs to occur in an interdisciplinary fashion by taking multiple perspectives into account.

### ***Advice for Training***

From the evaluation experience with *Levee Patroller*, we can also learn a laundry list of practical advice for designing a game-based training. These are the main recommendations:

1. *For all the world and his wife*: Not everybody may be as literate in computers and games, but games can be used successfully for everyone. Exceptions exist, because rebels and co-rebels are everywhere. Count on about 5% dropouts with a successful training.
2. *There is no place like home*: Everybody experienced playing at home as a blessing. Providing a distance training with a digital game is perfectly possible.
3. *Order! Order!*: But such a distance training needs to be very structured. The participants need to receive incentives to stay involved. Also, do not expect that participants will continue out of their own or remain involved afterward.
4. *First blow is half the battle*: The first moment participants make sense of an object is extremely important. They will stick to their guns after that, because whatever information is available or feedback is provided, they will continue to make sense of that object as they did the first time around.
5. *From “learn to” to “learn from”*: Participants need to first learn how to play the game before they learn from the game. A good introduction is essential to make participants enjoy the game and get them quickly to learning from the game.
6. *Make them hand and glove*: Try to link the learning objectives with the game scores. The game scores will provide a good estimate of how well players have learned from the game. This suggestion probably does not work with all types of games and learning objectives, but it is worth a try.

7. *Make a big pond*: The game is a big fish and it needs to swim in a big pond. Make sure that is there by embedding the game with other educational material and opportunities.
8. *Dealing with the spice of life*: Although variety is the spice of life, with game-based training this needs to be taken into account. One should especially take care of the variety in computer skills and education.
9. *No funny business*: The players take the game very seriously and so should the designer. Do not think that humorous elements will make the game better. It may also not be necessary because practitioners become already engaged just by the game's topic.
10. *Little things that matter*: Be wary of errors and misconceptions by players—how little they may seem. They have an important influence on the user experience. Be especially wary of automatically crawling mouse pointers.

The main challenge for using game-based training is to figure out how the game could contribute to existing activities and how it could be embedded into the organization. In each profession, this may require a different approach and lead to different results.

## Recommendations for Future Research and Practice

In addition to my main recommendation described earlier, which is to explore the connection between sensemaking and games, based on the training/evaluation I have several other recommendations for future research and practice, from increasing the possibilities of using game data to considering game design patterns for educational games.

### *Game Analytics for Serious Games*

One of the unique contributions of this book concerns the consideration of game data. This consideration relates to an up-and-coming area called *game analytics* (Isbister & Schaffer, 2008; Moura, Seif El-Nasr, & Shaw, 2011; Seif El-Nasr, Drachen, & Canossa, 2013; Kim et al., 2008; Medler & Magerko, 2011). This refers to the practice of gathering, analyzing, and disseminating data collected from games. Although it is done above all to collect data on how players actually play games, this data collection could be anything that happens in and around the game which can be logged.

What makes this perfectly possible is that many games are played over the Internet. Even game consoles and portable game devices have an Internet connection. Game companies are increasingly utilizing this possibility to understand their players and improve their games. User information is a critical success factor and a major

asset in one of the fastest growing and largest entertainment industries. That is why game industry–academic relationships are developed beyond technology and education and into game usability and analysis (Lameman et al., 2010). It is also why game companies are starting to hire data analysts.



**Fig. 12.1** The walking routes by players in one of the three regions. Created by Almar Joling

The trouble with game data is what to do with it. Massive amounts of data are retrieved and it is difficult to analyze and visualize this data with regular analytical tools. Heatmaps are an example of a popular visualization technique (cf., Drachen & Canossa, 2011). With color coding it is visualized in what areas of a level most players have died or in what areas players spent most of their time.

For analyzing the game data for *Levee Patroller*, a logging tool was developed. Unfortunately, this tool was too restricted and cumbersome for my purposes and so I decided to look at the raw data instead and transform this into a dataset manually—albeit this was time-consuming and cumbersome too and much more error prone. This led to an initial database of 1.474 variables and then I only looked at the variables that were most relevant to me. The data is by far not as massive as some of the data from popular entertainment games, but it is large enough for various interesting analyses beyond what I have presented here. For instance, I could also have considered when, how, and where players walked in the regions. The logs tracked every step players made and based on this it is possible to visualize their walking

routes among others (Fig. 12.1). Another possibility would be to explore the various player types (Canossa & Drachen, 2009).

I expect that the consideration of game data will make many advances in the next upcoming years and future research should be directed in applying those advances to the area of serious games (cf., Nacke, Drachen, & Göbel, 2010). Future research should also be directed at the development of game data methods and tools that are specifically developed for serious games. Such games have different demands on what data is collected and how.

Game data matters and will make a difference. This is not only true for the entertainment industry. It will most likely define the future of game-based training as well.

### ***Game as Research Method***

The generation of data by playing games makes games a potentially powerful research tool. This potential has been described earlier by Meijer (2009). He developed two low-tech games, one for hypothesis generation and another for hypothesis testing, and concludes that “gaming simulation is an excellent additional research method for controlled analysis” and that “future research could use gaming simulation as the research method of choice” (p. 171). He did this with paper-and-pencil material. Imagine what would be possible with digital games.

I imagined these possibilities, because frequently the thought occurred to me that I could have reframed the research with *Levee Patroller* by not considering the game as research object, but by using it as a means-to-an-end, as a research method. I could have used the game to look into decision-making or into sensemaking in crisis situations to name a few out of many possibilities. What decisions do people make and why?

When proposing this idea, the first concern is validity: does the game give an accurate portrayal of the actual world? Can we generalize what happens in a game-world? This book makes clear that a close link exists between the virtual environment and the actual environment. Others have pointed out this close link as well (Blascovich & Bailenson, 2011).

Games enable simulation of situations that rarely occur and which are difficult to study, such as a flooding. Although the game environment might be artificial and limited, it at least give us a possibility to study individual and organizational behavior in a certain context. Because crisis situations are hard to study in reality, I foresee a promising future for games as research method in the domain of safety and crisis response. A few researchers are already experimenting with this idea (Mendonça & Fiedrich, 2006; Pfaff, 2012; van Ruijven, 2008).

However, the potential of games as a research method extends far beyond this domain alone. One can think of its use in psychology, marketing, and public policy.

### ***Game as Assessment Tool***

Something else the game data clarified is that it is a useful predictor of actual performance. The game performance strongly related to the test performance. This stresses yet another potential use of games: assessment. Similar to that of research I have addressed this potential before (Harteveld, 2011). What I mentioned then is that this use is tricky. Games stop being a safe environment and suddenly becomes very serious.

This is an issue, because with *Levee Patroller* I was not assessing the participants and some participants already complained it was too much of an examination (Level 9). They said that they were not enjoying the game because of it. In this case I felt it was a matter of perception and of having experience with games, but undoubtedly the issue becomes greater when patrollers are being certified after successfully finishing the game. Future research should enlighten us on this particular issue—on when and how participants experience problems and what should be done.

I also mentioned that assessment does not necessarily need to be directed at people. Anything can be assessed. With this in mind I could have reframed my research into another direction. My research could have involved how to improve failure reporting. In that case I could have experimented with different reporting procedures and assessed their effectiveness. The current reporting procedure seems to be very helpful and I could have found clear evidence by contrasting this with another one. By changing terms and adjusting the procedure I could have also improved the reporting procedure experimentally. With a game we do not have to wait for a crisis to learn how to improve the organization.

Likewise, the use of games for assessment extends further than the domain of safety and crisis response. A much recent example is *Houthoff Buruma The Game*. This game was developed for the Dutch law firm Houthoff Buruma to find top talent among newly graduated jurists. Players are immersed into a case and have to distinguish themselves not only by professional qualities, but also through creativity, solution mindedness, stress resistance, and social skills.

### ***Game Design Patterns for Educational Games***

The first principle in this book is that “Rome was not built in a day.” We are just starting to look critically at what games can do and how we can improve them. The results in this book highlight this. We have tried to create a successful game by balancing the worlds of Reality, Meaning, and Play and yet we see much room for improvement.

To start, the results on certain failures and learning objectives were disappointing. Although these particular failures and objectives might be more difficult compared to others, I suspect that this poor performance is a result of insufficiently making clear to players how it should be done. The fact that players rarely changed their



initial choices confirms this. This suggests that we have to improve the feedback in the game. The question is how.

Another problem is that players should be made more situationally aware. Few report all signals and at the end categorical thinking or a tunnel vision is lurking around the corner. Players acquire fixed expectations. It is a positive development that they have an idea, but they should remain open to new information without being entrapped in old categories. Again the question is how.

These problems are specific to this game, but the answers might be applicable to many games. They may even be solved already by others in their games. Existing or non-existing, what is missing is an investigation of best practices of game mechanics for learning in games. Although much has been said about learning in and from games (e.g., Gee, 2003; Squire, 2011; de Freitas & Maharg, 2011), much less attention has been paid about what designers should do. If feedback needs to be implemented, what are the options? What are the advantages and disadvantages of the options? Under what circumstances does an option work and when not?

What I am requesting are *design patterns* for learning in and from games. The idea of design patterns comes from Alexander, Ishikawa, and Silverstein (1977) who wrote about architecture and urban design and defined a pattern as something that:

...describes a problem that occurs over and over again in our environment, and then describes the core of the solution to that problem in such a way that you can use this solution a million times over, without ever doing it the same way twice (p. x).

The idea about design patterns has been picked up and used for games in general (Björk & Holopainen, 2005; Schell, 2008). Besides a few number of initial attempts (Ecker, Müller, & Zylka, 2011; Kelle, Klemke, & Specht, 2011), the idea to look for game design patterns for educational games remains a suggestion like I am making here (Kiili, 2010).

## The Future of *Levee Patroller*

The training/evaluation with *Levee Patroller* took place in 2010 and we are now a couple of years farther down the road. I will conclude this book with discussing what happened after the training/evaluation.

I started this book by stating that games have a major potential within the domain of safety and crisis response, but that so far only a number of prototypes have been developed. Many of these prototypes are not further developed and for various reasons. They fell between the cracks. Unlike these games *Levee Patroller* was fully developed in 2006 and ready to be used. But what is the situation with the game in 2012? Did it fall between the cracks too?

After the training we developed, among other things, an administrator tool which allows the creation of player groups and tracking game performances. It is similar to the original logging tool, except that this one is Internet-based, updates automatically, and is free of bugs. Now anybody can administer the training. Because of

the limited funding these new developments progress slowly, but the game and its surrounding material continue to be improved. It is still alive and kicking.

Research investigations have continued as well. The latest concerns the combination of Geographical Information Systems (GIS) and a Procedural Content Generation (PCG) tool<sup>1</sup> to build new regions quickly and based on data of actual levees and their surroundings. The tool reads the GIS data and then algorithmically rather than manually generates a new region in the game in about ten minutes. Some manual tweaks are still needed, but this approach undoubtedly generates a new region much faster than doing everything manually. The question remains to what extent it is playable. That is a challenge that PCG scholars are investigating (Smith, Gan, Othenin-Girard, & Whitehead, 2011).

The game has remained part of the levee inspection course. In the past couple of years, this has been taught across the border, with trainees in South America and Asia trying to find failures in a Dutch landscape too. This is positive, but this course does not make use of the game as it should be used, and beyond this integration with the course not much has happened. The game is still not widely applied by the water authorities and only one more license has been sold.

The game-based training did not change this. The participating organizations were satisfied about the training, but they have no idea of how they will do it themselves. They prefer to outsource it and then it becomes quite costly—especially if hundreds of patrollers need to be trained. Although costs could be reduced by ignoring the end-meeting, a game-based training still requires far more attention, administration, and resources compared to a simple instruction evening. However, the authorities remain interested in it.

Therefore, *Levee Patroller* has not fallen between the cracks yet. It is not used as it should be used, but the future holds promise. Despite this promise, I am skeptical that it will ever be applied on a large scale, considering the very slow progress we have made so far and the difficulty of attracting additional investors.

Time will tell how *Levee Patroller* will develop. Even if it falls between the cracks the stories about its design and evaluation will remain. We can learn from these in building a strong knowledge base in game-based training.

---

<sup>1</sup> This tool is called *SketchaWorld* and has been developed by Smelik (2011).

# References

- Aarseth, E. J. (2001). Computer game studies, year one. *Game Studies*, 1(1). Retrieved from <http://www.gamestudies.org/0101/editorial.html>
- Abt, C. C. (1970). *Serious games*. New York, NY: Viking.
- Aldrich, C. (2004). *Simulations and the future of learning: An innovative (and perhaps revolutionary) approach to e-learning*. San Francisco, CA: Pfeiffer.
- Aldrich, C. (2009). *The complete guide to simulations and serious games: How the most valuable content will be created in the age beyond Gutenberg to Google*. San Francisco, CA: Pfeiffer.
- Ale, B. J. (2009). *Risk: An introduction. The concepts of risk, danger and chance*. New York, NY: Routledge.
- Alexander, C., Ishikawa, S., & Silverstein, M. (1977). *A pattern language: Towns, buildings, construction*. New York, NY: Oxford University Press.
- Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.
- Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Anderson, J. R. (1995). *Cognitive psychology and its implications*. New York, NY: Worth Publishers.
- Arnab, S., Dunwell, I., & Debattista, K. (Eds.). (2012). *Serious games for healthcare: Applications and implications*. Hershey, PA: Information Science Reference. doi: 10.4018/978-1-4666-1903-6
- Asher, J. W. (1976). *Educational research and evaluation methods*. Boston, MA: Little Brown, and Company.
- Babbie, E. (1989). *The practice of social research* (5th ed.). Belmont, CA: Wadsworth.
- Baldwin, T. T., & Ford, J. K. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology*, 41(1), 63–105. doi: 10.1111/j.1744-6570.1988.tb00632.x
- Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice-Hall.
- Baranowski, T., Buday, R., Thompson, D. I., & Baranowski, J. (2008). Playing for real: Video games and stories for health-related behavior change. *American Journal of Preventive Medicine*, 34(1), 74–82. doi: 10.1016/j.amepre.2007.09.027
- Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, 128(4), 612–637. doi: 10.1037/0033-2909.128.4.612
- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173–1182. doi: 10.1037/0022-3514.51.6.1173
- Beal, S. A., & Christ, R. E. (2004). *Training effectiveness evaluation of the Full Spectrum Command game* (Tech. Rep. No. 1140). Fort Benning, GA: U.S. Army Research Institute for the Behavioral and Social Sciences.

- Beale, I. L., Kato, P. M., Marin-Bowling, V. M., Guthrie, N., & Cole, S. W. (2007). Improvement in cancer-related knowledge following use of a psychoeducational video game for adolescents and young adults with cancer. *Journal of Adolescent Health, 41*(3), 263–270. doi: 10.1016/j.jadohealth.2007.04.006
- Beck, J. C., & Wade, M. (2004). *Got game: How the gamer generation is reshaping business forever*. Boston, MA: Harvard Business School Press.
- Becker, K. (2008). *The invention of good games: Understanding learning design in commercial videogames*. Unpublished doctoral dissertation, University of Calgary, Calgary, AB.
- Bekebrede, G. (2010). *Experiencing complexity: A game-based approach for understanding infrastructure systems*. Delft, the Netherlands: Next Generation Infrastructures Foundation.
- Bekebrede, G., Warmelink, H. J. G., & Mayer, I. S. (2011). Reviewing the need for gaming in education to accommodate the net generation. *Computers & Education, 57*(2), 1521–1529. doi: 10.1016/j.compedu.2011.02.010
- Bergeron, B. P. (2006). *Developing serious games*. Hingham, MA: Charles River Media.
- Bickman, L. (1987). The functions of program theory. *New Directions for Program Evaluation, 1987*(33), 5–18. doi: 10.1002/ev.1443
- Björk, S., & Holopainen, J. (2005). *Patterns in game design*. Hingham, MA: Charles River Media.
- Blascovich, J., & Bailenson, J. (2011). *Infinite reality: Avatars, eternal life, new worlds, and the dawn of the virtual revolution*. New York, NY: HarperCollins.
- Bogost, I. (2007). *Persuasive games: The expressive power of videogames*. Cambridge, MA: The MIT Press.
- Bogost, I. (2011). *How to do things with videogames*. Minneapolis, MN: University of Minnesota Press.
- Boin, A., 't Hart, P., Stern, E., & Sundelius, B. (2005). *The politics of crisis management: Public leadership under pressure*. Cambridge, UK: Cambridge University Press.
- Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. Cambridge, MA: The MIT Press.
- Brown, K. G., Sitzmann, T., & Bauer, K. N. (2010). Self-assessment one more time: With gratitude and an eye toward the future. *Academy of Management Learning & Education, 9*(2), 348–352.
- Brown, S. J., Lieberman, D. A., Gemeny, B. A., Fan, Y. C., Wilson, D. M., & Pasta, D. J. (1997). Educational video game for juvenile diabetes: Results of a controlled trial. *Informatics for Health and Social Care, 22*(1), 77–89. doi: 10.3109/14639239709089835
- Cailliois, R. (1958/1961). *Man, play and games* (M. Barash, Trans.). Champaign, IL: University of Illinois Press.
- Calleja, G. (2011). *In-game: From immersion to incorporation*. Cambridge, MA: MIT Press.
- Cameron, B., & Dwyer, F. (2005). The effect of online gaming, cognition and feedback type in facilitating delayed achievement of different learning objectives. *Journal of Interactive Learning Research, 16*(3), 243–258. Retrieved from <http://www.editlib.org/p/5896>
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.
- Canossa, A., & Drachen, A. (2009). Play-personas: Behaviours and belief systems in user-centred game design. In T. Gross et al. (Eds.), *Human-Computer Interaction—INTERACT 2009* (Vol. 5727, pp. 510–523). Berlin, Germany: Springer. doi: 10.1007/978-3-642-03658-3\_55
- Carlile, P. R. (2002). A pragmatic view of knowledge and boundaries: Boundary objects in new product development. *Organization Science, 13*(4), 442–455. doi: 10.1287/orsc.13.4.442.2953
- Castronova, E. (2005). *Synthetic worlds: The business and culture of online games*. Chicago, IL: University of Chicago Press.
- Chatham, R. E. (2007, July). Games for training. *Communications of the ACM, 50*(7), 36–43. doi: 10.1145/1272516.1272537
- Chen, H. T. (1990). *Theory-driven evaluations*. Newsbury Park, CA: Sage.
- Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*(2), 121–152.

- Clark, R. E. (2007). Learning from serious games? Arguments, evidence, and research suggestions. *Educational Technology*, 47, 56–59.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Cohen, L., Manion, L., Morrison, K., & Morrison, K. R. B. (2007). *Research methods in education*. New York, NY: Routledge.
- Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., & Boyle, J. M. (2012). A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education*, 59(2), 661–686. doi: 10.1016/j.compedu.2012.03.004
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Chicago, IL: Rand McNally.
- Cooper, C., & Block, R. (2006). *Disaster: Hurricane Katrina and the failure of homeland security*. New York, NY: Times Books.
- Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., . . . Foldit players (2010). Predicting protein structures with a multiplayer online game. *Nature*, 446, 756–760. doi: 10.1038/nature09304
- Crandall, B., Klein, G., & Hoffman, R. R. (2006). *Working minds: A practitioner's guide to cognitive task analysis*. Cambridge, MA: The MIT Press.
- Creswell, J., & Clark, V. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.
- Crookall, D. (1995). *Debriefing: The key to learning from simulation/games*. Thousand Oaks, CA: Sage.
- Csikszentmihalyi, M. (1991). *Flow: The psychology of optimal experience*. New York, NY: Harper Perennial.
- de Freitas, S. (2006). *Learning in immersive worlds: A review of game-based learning* (Tech. Rep.). Bristol, UK: Joint Information Systems Committee.
- de Freitas, S., & Maharg, P. (2011). *Digital games and learning*. London, UK: Continuum Press.
- de Freitas, S., & Oliver, M. (2006). How can exploratory learning with games and simulations within the curriculum be most effectively evaluated? *Computers & Education*, 46(3), 249–264. doi: 10.1016/j.compedu.2005.11.007
- de Graeff, J. J., van der Heide, O., Mouwen, J. M. A. M., & van der Wal, J. T. (1987). *Het waterschap in kort bestek* [The water authority in short]. The Hague, the Netherlands: VUGA.
- Dervin, B. (1998). Sense-making theory and practice: An overview of user interests in knowledge seeking and use. *Journal of Knowledge Management*, 2(2), 36–46. doi: 10.1108/13673279810249369
- Dervin, B., Foreman-Wernet, L., & Lauterbach, E. (2003). *Sense-making methodology reader: selected writings of Brenda Dervin*. Cresskill, NJ: Hampton Press.
- Dervin, B., & Naumer, C. M. (2009). Sense-making. In M. J. Bates & M. N. Maack (Eds.), *Encyclopedia of library and information sciences* (3rd ed., pp. 4696–4707). Boca Raton, FL: Taylor & Francis. doi: 10.1081/E-ELIS3-120043227
- Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: defining “gamification”. In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments* (pp. 9–15). New York, NY, USA: ACM. doi: 10.1145/2181037.2181040
- Dewey, J. (1938). *Experience and education*. New York, NY: MacMillan Publishing Company.
- Dibbell, J. (2011). Serious games. *Technology Review*, 114(1), 74–76.
- Dienes, Z., & Perner, J. (1999). A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences*, 22(5), 735–808. doi: 10.1017/S0140525X99452185
- Dodgson, M., Gann, D. M., & Salter, A. (2007). “In case of fire, please use the elevator”: Simulation technology and organization in fire engineering. *Organization Science*, 18(5), 849–864. doi: 10.1287/orsc.1070.0287
- Drachen, A., & Canossa, A. (2011). Evaluating motion: Spatial user behaviour in virtual environments. *International Journal of Arts and Technology*, 4(3), 294–314. doi: 10.1504/

- IJART.2011.041483
- Duke, R. D., & Geurts, J. (2004). *Policy games for strategic management: Pathways into the unknown*. Amsterdam, the Netherlands: Dutch University Press.
- Dunne, J. R., & McDonald, C. L. (2010). Pulse!!: A model for research and development of virtual-reality learning in military medical education and training. *Military Medicine*, 175(1), 25–27.
- Dyckmeester, B. (1940). *Het waterschap: Als voorbeeld van een typisch Nederlandse corporatie* [The water board: An example of a typical Dutch corporation]. n.p.
- Dykstra, E. (2009). *Katrina: Orkaan in Nederland?* [Katrina: Hurricane in the Netherlands?]. Alphen aan den Rijn, the Netherlands: Kluwer.
- E-Semble. (2010). *XVR: Virtual reality training for safety and security* [Brochure]. Delft, the Netherlands: E-Semble.
- Ecker, M., Müller, W., & Zylka, J. (2011). Game-based learning design patterns: An approach to support the development of “better” educational games. In P. Felicia (Ed.), *Handbook of research on improving learning and motivation through educational games: Multidisciplinary approaches* (pp. 137–152). Hershey, PA: Information Science Reference. doi: 10.4018/978-1-60960-495-0.ch007
- Eco, U. (1998). *Serendipities: Language and lunacy* (W. Weaver, Trans.). San Diego, CA: Harcourt.
- Ederly, D., & Mollick, E. (2009). *Changing the game: How video games are transforming the future of business*. Upper Saddle River, NJ: FT Press.
- Egenfeldt-Nielsen, S. (2006). Overview of research on the educational use of video games. *Digital Kompetanse*, 1(3), 184–213.
- Egenfeldt-Nielsen, S. (2007). *Beyond edutainment: The educational potential of computer games*. London, UK: Continuum Press.
- Egenfeldt-Nielsen, S., Smith, J. H., & Tosca, S. P. (2008). *Understanding video games: The essential introduction*. New York, NY: Routledge.
- Endsley, M. R., & Garland, D. J. (Eds.). (2000). *Situation awareness: Analysis and measurement*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Federation of American Scientists. (2006). *Summit on educational games: Harnessing the power of video games for learning* (Tech. Rep.). Washington, DC: Federation of American Scientists.
- Feinstein, A. H., & Cannon, H. M. (2001). Fidelity, verifiability, and validity of simulation: Constructs for evaluation. *Developments in Business Simulation and Experiential Learning*, 28, 57–67.
- Field, A. (2005). *Discovering statistics using SPSS* (2nd ed.). London, UK: Sage. Paperback.
- Field, A., & Hole, G. (2003). *How to design and report experiments*. London, UK: Sage.
- Firestone, W. A. (1993). Alternative arguments for generalizing from data as applied to qualitative research. *Educational Researcher*, 22(4), 16–23. doi: 10.3102/0013189X022004016
- Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley.
- Fjermestad, J., & Hiltz, S. R. (1998). An assessment of group support systems experimental research: methodology and results. *Journal of Management Information Systems*, 15, 7–149.
- Forthofer, M. S. (2003). Status of mixed methods in the health sciences. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 527–540). Thousand Oaks, CA: Sage.
- Frank, A. (2007). Balancing three different foci in the design of serious games: Engagement, training objective and context. In D. Thomas & R. L. Appelman (Eds.), *Conference Proceedings of DiGRA 2007: Situated Play* (pp. 567–574). Tokyo, Japan: University of Tokyo.
- Galanes, G. J., & Adams, K. (2011). *Effective group discussion: Theory and practice* (14th ed.). New York, NY: McGraw-Hill Education.
- Gamberini, L., Alcaniz, M., Barresi, G., Fabregat, M., Ibanez, F., & Prontu, L. (2006). Cognition, technology and games for the elderly: An introduction to ELDERGAMES Project.

- PsychNology Journal*, 4(3), 285–308.
- Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, motivation, and learning: A research and practice model. *Simulation & Gaming*, 33(4), 441–467. doi: 10.1177/1046878102238607
- Gee, J. (2003). *What video games have to teach us about learning and literacy*. New York, NY: Palgrave Macmillan.
- Geertz, C. (1973). *The interpretation of cultures: Selected essays*. New York, NY: Basic Books.
- Gioia, D. A., & Chittipeddi, K. (1991). Sensemaking and sensegiving in strategic change initiation. *Strategic Management Journal*, 12(6), 433–448. doi: 10.1002/smj.4250120604
- Girard, C., Ecalte, J., & Magnan, A. (2012). Serious games as new educational tools: How effective are they? A meta-analysis of recent studies. *Journal of Computer Assisted Learning*. doi: 10.1111/j.1365-2729.2012.00489.x
- Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago, IL: Aldine Publishing Company.
- Harteveld, C. (2009). Making sense of studying games: Using sensemaking as a perspective for game research. In G. Kin & Y. Cai (Eds.), *Learn to game, game to learn: Proceedings of the 40th ISAGA Conference*. Singapore: National University of Singapore.
- Harteveld, C. (2011). *Triadic game design: Balancing reality, meaning and play*. London, UK: Springer.
- Harteveld, C., Guimarães, R., Mayer, I. S., & Bidarra, R. (2010). Balancing play, meaning and reality: The design philosophy of Levee Patroller. *Simulation & Gaming*, 41(3), 316–340. doi: 10.1177/1046878108331237
- Harz, C. R., & Stern, P. A. (2008). Serious games for first responders: Improving design and usage with social learning theory. *Dissertations Abstracts International Section A: Humanities and Social Sciences*, 69(7–A), 2681.
- Haskell, R. (1998). *Reengineering corporate training: Intellectual capital and transfer of learning*. Charlotte, NC: Information Age Publishing.
- Hays, R. T. (2005). *The effectiveness of instructional games: A literature review and discussion* (Tech. Rep. Nos. 2005–004). Orlando, FL: Naval Air Warfare Center Training Systems Division.
- Hevner, A. R. (2007). A three cycle view of design science research. *Scandinavian Journal of Information Systems*, 19(2), 87–92.
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105.
- Howell, D. (2010). *Statistical methods for psychology* (7th ed.). Belmont, CA: Wadsworth.
- Huizinga, J. (1938/1955). *Homo ludens: A study of the play-element in culture* (R. F. C. Hull, Trans.). Boston, MA: Beacon Press.
- Hussain, T., Feurzeig, W., Cannon-Bowers, J., Coleman, S., Koenig, A., Lee, J., ... Wainess, R. (2010). Development of game-based training systems: Lessons learned in an interdisciplinary field in the making. In J. Cannon-Bowers & C. Bowers (Eds.), *Serious game design and development: Technologies for training and learning* (pp. 47–80). Hershey, PA: Information Science Reference. doi: 10.4018/978-1-61520-739-8.ch004
- IJsselstein, W., Nap, H. H., de Kort, Y., & Poels, K. (2007). Digital game design for elderly users. In *Proceedings of the 2007 conference on future play* (pp. 17–22). New York, NY, USA: ACM. doi: 10.1145/1328202.1328206
- Isbister, K., & Schaffer, N. (Eds.). (2008). *Game usability: Advice from the experts for advancing the player experience*. Burlington, MA: Morgan Kaufmann Publishers.
- Jennett, C., Cox, A. L., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., & Walton, A. (2008). Measuring and defining the experience of immersion in games. *International Journal of Human-Computer Studies*, 66(9), 641–661. doi: 10.1016/j.ijhcs.2008.04.004
- Jensen, E. (2009). Sensemaking in military planning: A methodological study of command teams. *Cognition, Technology & Work*, 11, 103–118. doi: 10.1007/s10111-007-0084-x
- Johnson, R., & Onwuegbuzie, A. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, 33(7), 14–26. doi: 10.3102/0013189X033007014

- Johnson, R. B., & Turner, L. A. (2003). Data collection strategies in mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 297–319). Thousand Oaks, CA: Sage.
- Johnson, S. (2005). *Everything bad is good for you: How today's popular culture is actually making us smarter*. New York, NY: Riverhead Books.
- Jones, S. (2008). *The meaning of video games: Gaming and textual strategies*. New York, NY: Routledge.
- Juul, J. (2005). *Half-real: Video games between real rules and fictional worlds*. Cambridge, MA: The MIT Press.
- Kapp, K. (2012). *The gamification of learning and instruction: Game-based methods and strategies for training and education*. San Francisco, CA: John Wiley & Sons.
- Kato, P. M. (2010). Video games in health care: Closing the gap. *Review of General Psychology*, 14(2), 113–121. doi: 10.1037/a0019441
- Kato, P. M., Cole, S. W., Bradlyn, A. S., & Pollock, B. H. (2008). A video game improves behavioral outcomes in adolescents and young adults with cancer: A randomized trial. *Pediatrics*, 122(2), 305–317. doi: 10.1542/peds.2007-3134
- Ke, F. (2009). A qualitative meta-analysis of computer games as learning tools. In R. E. Ferdig (Ed.), *Handbook of research on effective electronic gaming in education* (Vol. 1, pp. 1–32). Hershey, PA: Information Science Reference.
- Ke, F., & Grabowski, B. (2007). Gameplaying for maths learning: Cooperative or not? *British Journal of Educational Technology*, 38(2), 249–259. doi: 10.1111/j.1467-8535.2006.00593.x
- Kelle, S., Klemke, R., & Specht, M. (2011). Design patterns for learning games. *International Journal of Technology Enhanced Learning*, 3(6), 555–569. doi: 10.1504/IJTEL.2011.045452
- Kienhuis, J. H. M., Westerwoudt, T. W., van der Wal, J. T., & Berge, A. P. (1993). *Het waterschap van oud naar nieuw* [The water authority from old to new]. Delft, the Netherlands: Hoogheemraadschap van Delfland.
- Kiili, K. (2010). Call for learning-game design patterns. In F. Edvardsen & H. Kulle (Eds.), *Educational games: Design, learning and applications* (pp. 299–311). Hauppauge, NY: Nova Publishers.
- Kim, J. H., Gunn, D. V., Schuh, E., Phillips, B., Pagulayan, R. J., & Wixon, D. (2008). Tracking real-time user experience (true): a comprehensive instrumentation solution for complex systems. In *Proceedings of the 26th annual SIGCHI Conference on Human factors in Computing Systems* (pp. 443–452). New York, NY, USA: ACM. doi: 10.1145/1357054.1357126
- Kirriemuir, J., & McFarlane, A. (2004). *Literature review in games and learning* (Tech. Rep. No. 8). Bristol, UK: Futurelab.
- Kitchin, R. (1994). Cognitive maps: What are they and why study them? *Journal of Environmental Psychology*, 14(1), 1–19. doi: 10.1016/S0272-4944(05)80194-X
- Kitzinger, J. (1995). Qualitative research: Introducing focus groups. *BMJ*, 311(7000), 299–302. doi: 10.1136/bmj.311.7000.299
- Klabbers, J. H. G. (2006). *The magic circle: Principles of gaming and simulation*. Rotterdam, the Netherlands: Sense Publishers.
- Klasson, C. R. (1964). Business gaming: A progress report. *Academy of Management Journal*, 7(3), 175–188.
- Klein, G., Moon, B., & Hoffman, R. R. (2006a). Making sense of sensemaking 1: Alternative perspectives. *IEEE Intelligent Systems*, 21(4), 70–73. doi: 10.1109/MIS.2006.75
- Klein, G., Moon, B., & Hoffman, R. R. (2006b). Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent Systems*, 21(5), 88–92. doi: 10.1109/MIS.2006.100
- Kline, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, 56(5), 746–759. doi: 10.1177/0013164496056005002
- Knight, J. F., Carley, S., Tregunna, B., Jarvis, S., Smithies, R., de Freitas, S., ... Mackway-Jones, K. (2010). Serious gaming technology in major incident triage training: A pragmatic con-



- trolled trial. *Resuscitation*, 81(9), 1175–1179. doi: 10.1016/j.resuscitation.2010.03.042
- Knowles, M. S., Elwood F. Holton, I., & Swanson, R. A. (1998). *The adult learner* (5th ed.). Houston, TX: Gulf Publishing Company.
- Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development*. Upper Saddle River, NJ: Prentice Hall.
- Koster, R. (2005). *A theory of fun for game design*. Scottsdale, AZ: Paraglyph Press.
- Kraiger, K., Ford, K. J., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology*, 78(2), 311–328. doi: 10.1037/0021-9010.78.2.311
- Kriz, W. C., & Hense, J. U. (2006). Theory-oriented evaluation for the design of and research in gaming and simulation. *Simulation & Gaming*, 37(2), 268–283. doi: 10.1177/1046878106287950
- Kuit, M. (2002). *Strategic behavior and regulatory styles in the netherlands energy industry*. Delft, the Netherlands: Eburon.
- Kurenov, S. N., Cance, W. W., Noel, B., & Mozingo, D. W. (2009). Game-based mass casualty burn training. In J. Westwood et al. (Eds.), *Medicine meets virtual reality 17* (pp. 142–144). Amsterdam, the Netherlands: IOS Press. doi: 10.3233/978-1-58603-964-6-142
- Lameman, B. A., Seif El-Nasr, M., Drachen, A., Foster, W., Moura, D., & Aghabeigi, B. (2010). User studies: A strategy towards a successful industry-academic relationship. In *Proceedings of the International Academic Conference on the Future of Game Design and Technology* (pp. 134–142). New York, NY, USA: ACM. doi: 10.1145/1920778.1920798
- Laurel, B. (2001). *Utopian entrepreneur*. Cambridge, MA: The MIT Press.
- Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge, MA: Cambridge University Press.
- Lazzaro, N. (2008). The four fun keys. In K. Isbister & N. Schaffer (Eds.), *Game usability: Advice from the experts for advancing the player experience* (pp. 317–343). Burlington, MA: Morgan Kaufmann.
- Lee, A. T. (2005). *Flight simulation: Virtual environments in aviation*. Hampshire, UK: Ashgate.
- Leemkuil, H. H. (2006). *Is it all in the game? Learner support in an educational knowledge management simulation game*. Unpublished doctoral dissertation, University of Twente, Enschede, the Netherlands.
- Leemkuil, H. H., de Jong, T., & Ootes, S. (2000). *Review of educational use of games and simulations* (Tech. Rep. No. IST-1999-13078). Enschede, the Netherlands: University of Twente.
- Ma, Y., Williams, D., Prejean, L., & Richard, C. (2007). A research agenda for developing and implementing educational computer games. *British Journal of Educational Technology*, 38(3), 513–518. doi: 10.1111/j.1467-8535.2007.00714.x
- Malone, T. W. (1981). Toward a theory of intrinsically motivating instruction. *Cognitive Science*, 5(4), 333–369. doi: 10.1016/S0364-0213(81)80017-1
- Malone, T. W., & Lepper, M. (1987a). Intrinsic motivation and instructional effectiveness in computer-based education. In R. Snow & M. Farr (Eds.), *Aptitude learning and instruction* (pp. 152–188). London, UK: Lawrence Erlbaum Associates.
- Malone, T. W., & Lepper, M. (1987b). Making learning fun: A taxonomy of intrinsic motivation for learning. In R. Snow & M. Farr (Eds.), *Aptitude learning and instruction* (pp. 223–253). London, UK: Lawrence Erlbaum Associates.
- Marsh, T., Nickole, L. Z., Klopfer, E., Xuejin, C., Osterweil, S., & Haas, J. (2011). Fun and learning: Blending design and development dimensions in serious games through narrative and characters. In M. Ma, A. Oikonomou, & L. C. Jain (Eds.), *Serious games and edutainment applications* (pp. 273–288). London, UK: Springer. doi: 10.1007/978-1-4471-2161-9
- Mayer, I. S. (2009). The gaming of policy and the politics of gaming: A review. *Simulation & Gaming*, 40(6), 825–862. doi: 10.1177/1046878109346456
- Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, 38(1), 43–52. doi: 10.1207/S15326985EP3801\_6

- Mäyrä, F. (2008). *An introduction to theory and game studies: Games in culture*. London, UK: Sage.
- McDonald, C. L. (2010). Avatars and diagnosis: Delivering medical curricula in virtual space. In J. Cannon-Bowers & C. Bowers (Eds.), *Serious game design and development: Technologies for training and learning* (p. 233-245). Hershey, PA: Information Science Reference. doi: 10.4018/978-1-61520-739-8.ch012
- McGonigal, J. (2011). *Reality is broken: Why games make us better and how they can change the world*. London, UK: Jonathan Cape.
- McGrath, D., & Hill, D. (2004). UnrealTriage: A game-based simulation for emergency response. In *Proceedings of the Huntsville Simulation Conference*. Huntsville, AL.
- Medler, B., & Magerko, B. (2011). Analytics of play: Using information visualization and game-play practices for visualizing video game data. *Parsons Journal for Information Mapping*, 3(1), 1–12. Retrieved from <http://piim.newschool.edu/journal/issues/2011/01/>
- Meijer, S. A. (2009). *The organisation of transactions: Studying supply networks using gaming simulation*. Wageningen, the Netherlands: Wageningen Academic Publishers.
- Meister, J. (Ed.). (2002). *Pillars of e-learning success*. New York, NY: Corporate University Exchange.
- Mendonça, D., & Fiedrich, F. (2006). Training for improvisation in emergency management: Opportunities and limits for information technology. *International Journal of Emergency Management*, 3(4), 348–363. doi: 10.1504/IJEM.2006.011301
- Merriam, S. B., & Caffarella, R. S. (1999). *Learning in adulthood: A comprehensive guide* (2nd ed.). San Francisco, CA: Jossey-Bass.
- Meyers, L., Gamst, G., & Guarino, A. (2006). *Applied multivariate research: Design and interpretation*. Thousand Oaks, CA: Sage.
- Michael, D., & Chen, S. (2006). *Serious games: Games that educate, train, and inform*. Boston, MA: Thomson Course Technology PTR.
- Millsap, R. E., & Maydeu-Olivares, A. (Eds.). (2009). *The Sage handbook of quantitative methods in psychology*. London, UK: Sage.
- Ministry of Transport, Public Works and Water Management. (2006). *Spatial planning key decision 'room for the river': Investing in the safety and vitality of the Dutch river basin region* (Tech. Rep.). The Hague, the Netherlands: Ministry of Transport, Public Works and Water Management.
- Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record*, 108(6), 1017–1054.
- Moskal, B. M. (2010). Self-assessments: What are their valid uses? *Academy of Management Learning & Education*, 9(2), 314–320.
- Moura, D., Seif El-Nasr, M., & Shaw, C. D. (2011). Visualizing and understanding players' behavior in video games: Discovering patterns and supporting aggregation and comparison. In *Proceedings of the 2011 ACM SIGGRAPH Symposium on Video Games* (pp. 11–15). New York, NY, USA: ACM. doi: 10.1145/2018556.2018559
- Muller, D., Judd, C. M., & Yzerbyt, V. Y. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology*, 89(6), 852–863. doi: 10.1037/0022-3514.89.6.852
- Murray, J. (1997). *Hamlet on the holodeck*. New York, NY: The Free Press.
- Nacke, L. E., Drachen, A., & Göbel, S. (2010). Methods for evaluating gameplay experience in a serious gaming context. *International Journal of Computer Science in Sport*, 9(2). Retrieved from <http://www.iacss.org/index.php?id=96>
- Narayanasamy, V., Wong, K. W., Fung, C. C., & Rai, S. (2006). Distinguishing games and simulation games from simulators. *Computers in Entertainment*, 4(2). doi: 10.1145/1129006.1129021
- Oblinger, D. (2004). The next generation of educational engagement. *Journal of Interactive Media in Education*, 2004(8). Retrieved from <http://www.jime.open.ac.uk/article/2004-8-oblinger/198>

- O'Neil, H. F., Wainess, R., & Baker, E. L. (2005). Classification of learning outcomes: evidence from the computer games literature. *The Curriculum Journal*, 16(4), 455–474. doi: 10.1080/09585170500384529
- Orlikowski, W. J., & Gash, D. C. (1994). Technological frames: Making sense of information technology in organizations. *ACM Transactions on Information Systems*, 12(2), 174–207. doi: 10.1145/196734.196745
- Orvis, K. A., Horn, D. B., & Belanich, J. (2008). The roles of task difficulty and prior videogame experience on performance and motivation in instructional videogames. *Computers in Human Behavior*, 24(5), 2415–2433. doi: 10.1016/j.chb.2008.02.016
- Peng, W., Lin, J.-H., & Crouse, J. (2011). Is playing exergames really exercising? A meta-analysis of energy expenditure in active video games. *Cyberpsychology behavior and social networking*, 14(11), 681–688. doi: 10.1089/cyber.2010.0578
- Peters, V., Vissers, G., & Heijne, G. (1998). The validity of games. *Simulation & Gaming*, 29(1), 20–30. doi: 10.1177/1046878198291003
- Peterson, R. A. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of Consumer Research*, 28(3), 450–461. doi: 10.1086/323732
- Pfaff, M. S. (2012). Negative affect reduces team awareness: The effects of mood and stress on computer-mediated team communication. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 54(4), 560–571. doi: 10.1177/0018720811432307
- Pieper, J. (1948/1998). *Leisure: the basis of culture* (G. Malsbary, Trans.). South Bend, IN: St. Augustine's Press.
- Pillemer, K. (2011). *30 lessons for living: Tried and true advice from the wisest Americans*. New York, NY: Hudson Street Press.
- Pinsonneault, A., & Kraemer, K. L. (1990). The effects of electronic meetings on group processes and outcomes: An assessment of the empirical research. *European Journal of Operational Research*, 46(2), 143–161. doi: 10.1016/0377-2217(90)90128-X
- Pivec, M., & Pivec, P. (2008). *Games in schools* (Tech. Rep.). Brussels, Belgium: Interactive Software Federation of Europe.
- Plass, J., Moreno, R., & Brünken, R. (Eds.). (2010). *Cognitive load theory*. New York, NY: Cambridge University Press.
- Pollak, O. (1943). Conservatism in later maturity and old. *American Sociological Review*, 8(2), 175–179.
- Postigo, H. (2007). Of mods and modders: chasing down the value of fan-based digital game modifications. *Games and Culture*, 2(4), 300–313. doi: 10.1177/1555412007307955
- Prensky, M. (2001). *Digital game-based learning*. New York, NY: McGraw-Hill.
- PwC Entertainment & Media. (2012). *Global entertainment and media outlook: 2012–2016* (Tech. Rep.). New York, NY: PricewaterhouseCoopers.
- Raser, J. (1969). *Simulations and society: An exploration of scientific gaming*. Boston, MA: Allyn and Bacon.
- Ratan, R., & Ritterfeld, U. (2009). Classifying serious games. In U. Ritterfeld, M. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 10–24). New York, NY: Routledge.
- Reeves, B., & Read, J. L. (2009). *Total engagement: Using games and virtual worlds to change the way people work and businesses compete*. Boston, MA: Harvard Business Press.
- Rieber, L. P. (1996). Seriously considering play: Designing interactive learning environments based on the blending of microworlds, simulations, and games. *Educational Technology Research & Development*, 44(2), 45–58. doi: 10.1007/BF02300540
- Rijkswaterstaat. (2011). *Water management in the Netherlands* (Tech. Rep.). The Hague, the Netherlands: Author.
- Rolfe, J., & Staples, K. (1988). *Flight simulation*. Cambridge, UK: Cambridge University Press.
- Rollings, A., & Adams, E. (2003). *Andrew Rollings and Ernest Adams on game design*. Indianapolis, IN: New Riders Publishing.

- Russell, D. M., Stefik, M. J., Pirolli, P., & Card, S. K. (1993). The cost structure of sensemaking. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (pp. 269–276). New York, NY, USA: ACM. doi: 10.1145/169059.169209
- Salen, K., & Zimmerman, E. (2004). *Rules of play: game design fundamentals*. Cambridge, MA: MIT Press.
- Sawyer, B. (2002). *Serious games: Improving public policy through game-based learning and simulation* (Tech. Rep.). Washington, DC: Woodrow Wilson International Center for Scholars.
- Sawyer, B., & Smith, P. (2008, February 18). *Serious games taxonomy*. Serious Games Summit at the Game Developers Conference, San Francisco, CA.
- Scannell, E. E., & Les Donaldson, E. S. (2000). *Human resource development: The new trainer's guide* (3rd ed.). New York, NY: Perseus Publishing.
- Schank, R. (1997). *Virtual learning: A revolutionary approach to building a highly-skilled workforce*. New York, NY: McGraw-Hill.
- Schell, J. (2008). *The art of game design: A book of lenses*. Burlington, MA: Morgan Kaufmann.
- Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. New York, NY: Basic Books.
- Schön, D. A. (1987). *Educating the reflective practitioner: Toward a new design for teaching and learning in the professions*. San Francisco, CA: Jossey-Bass Publishers.
- Seif El-Nasr, M., Drachen, A., & Canossa, A. (Eds.). (2013). *Game analytics: Maximizing the value of player data*. London, UK: Springer.
- Shaffer, D. W. (2006). *How computer games help children learn*. New York, NY: Palgrave Macmillan.
- Simon, H. A. (1969). *The sciences of the artificial* (3rd ed.). Cambridge, MA: The MIT Press.
- Sitzmann, T. (2011). A meta-analytic examination of the instructional effectiveness of computer-based simulation games. *Personnel Psychology*, 64(2), 489–528. doi: 10.1111/j.1744-6570.2011.01190.x
- Sitzmann, T., Brown, K. G., Casper, W. J., Ely, K., & Zimmerman, R. D. (2008). A review and meta-analysis of the nomological network of trainee reactions. *Journal of Applied Psychology*, 93(2), 280–295. doi: 10.1037/0021-9010.93.2.280
- Sitzmann, T., Ely, K., Brown, K. G., & Bauer, K. N. (2010). Self-assessment of knowledge: A cognitive learning or affective measure? *Academy of Management Learning & Education*, 9(2), 169–191.
- Slager, C. (Ed.). (2003). *De ramp [The disaster]*. Amsterdam, the Netherlands: Atlas.
- Smelik, R. M. (2011). *A declarative approach to procedural generation of virtual worlds*. Unpublished doctoral dissertation, Delft University of Technology, Delft, the Netherlands.
- Smith, G., Gan, E., Othenin-Girard, A., & Whitehead, J. (2011). PCG-based game design: Enabling new play experiences through procedural content generation. In *Proceedings of the 2nd International Workshop on Procedural Content Generation in Games* (pp. 7:1–7:4). New York, NY, USA: ACM. doi: 10.1145/2000919.2000926
- Smith, R. (2010). The long history of gaming in military training. *Simulation & Gaming*, 40(1), 6–19. doi: 10.1177/1046878109334330
- Spicer, D. P. (1998). Linking mental models and cognitive maps as an aid to organisational learning. *Career Development International*, 3(3), 125–132. doi: 10.1108/13620439810211126
- Squire, K. D. (2004). *Replaying history: Learning world history through playing Civilization III*. Unpublished dissertation, Indiana University, Bloomington, IN.
- Squire, K. D. (2007). Games, learning, and society: Building a field. *Educational Technology*, 47(5), 51–54.
- Squire, K. D. (2011). *Video games and learning: Teaching and participatory culture in the digital age*. New York, NY: Teachers College Press.
- Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's museum of vertebrate zoology, 1907–39. *Social Studies of Science*, 19(3), 387–420. doi: 10.1177/030631289019003001
- Stone, R. (2005). Serious gaming. *Defence Management Journal*, 31, 142–144.

- Stone, R. (2009). Serious games: Virtual reality's second coming? *Virtual Reality*, 13(1), 1–2. doi: 10.1007/s10055-008-0109-7
- Surface, E. A., Dierdorff, E. C., & Watson, A. M. (2007). *Special operations language training software measurement of effectiveness study: Tactical Iraqi study final report* (Tech. Rep. No. 2007010602). Raleigh, NC: Surface, Ward & Associates.
- Sweller, J., van Merriënboer, J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10, 251–296. doi: 10.1023/A:1022193728205
- Tashakkori, A., & Teddlie, C. (Eds.). (2003). *Handbook of mixed methods in social and behavioral research*. Thousand Oaks, CA: Sage.
- Tate, R., Haratatos, J., & Cole, S. (2009). Hopelab's approach to Re-Mission. *International Journal of Learning and Media*, 1(1), 29–35. doi: 10.1162/ijlm.2009.0003
- Taylor, J., & Van Every, E. (2000). *The emergent organization: Communication as its site and surface*. Mahwah, NJ: Lawrence Erlbaum Associates.
- ten Brinke, W. B. M., & Bannink, B. A. (2004). *Risico's in bedijekte termen: Een thematische evaluatie van het nederlandse veiligheidsbeleid tegen overstromen* [Risks in levee terms: A thematic evaluation of the the dutch safety policy against flooding] (Tech. Rep. No. 500799002). Bilthoven, the Netherlands: RIVM.
- Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. *Psychological Review*, 8, 247–261.
- Tobias, S., & Fletcher, J. D. (Eds.). (2011). *Computer games and instruction*. Charlotte, NC: Information Age.
- Tobias, S., & Fletcher, J. D. (2012). Reflections on “a review of trends in serious gaming”. *Review of Educational Research*, 82(2), 233–237. doi: 10.3102/0034654312450190
- Tolone, W. (2009). Making sense of complex systems through integrated modeling and simulation. In Z. Ras & W. Ribarsky (Eds.), *Advances in information and intelligent systems* (Vol. 251, pp. 21–40). Heidelberg, Germany: Springer.
- Trimension. (2001). *Calamiteitenmap: Beschrijving calamiteitenzorgsysteem* [Calamity dossier: Description calamity care system] (Tech. Rep.). The Hague, the Netherlands: Unie van Waterschappen.
- Vale, M. (2001). *The princely court: Medieval courts and culture in North-West Europe, 1270-1380*. New York, NY: Oxford University Press.
- van der Meer, F.-B. (1983). *Organisatie als spel: Social simulatie als methode in onderzoek naar organiseren* [Organization as game: Social simulation as method in a research into organizing]. Unpublished doctoral dissertation, University of Twente, Enschede, the Netherlands.
- van Bueren, E., Mayer, I. S., Hartevelde, C., & Scalzo, R. (2009). Van tekentafel naar bestuurlijke implementatie: Gamen met bestuurders in de rechtspraak en het Openbaar Ministerie [From the judiciary designers table to administrative implementation: Gaming with professionals in the judiciary and the Public Prosecution Office]. *Bestuurskunde*, 18(3), 47–59.
- Vance, S. C., & Gray, C. F. (1967). Use of performance evaluation model for research in business gaming. *The Academy of Management Journal*, 10(1), 27–37.
- van der Spek, E. D. (2011). *Experiments in serious game design: A cognitive approach*. Unpublished doctoral dissertation, Utrecht University, Utrecht, the Netherlands.
- van der Spek, E. D., Wouters, P., & van Oostendorp, H. (2011). Code red: Triage or cognition-based design rules enhancing decisionmaking training in a game environment. *British Journal of Educational Technology*, 42(3), 441–455. doi: 10.1111/j.1467-8535.2009.01021.x
- van Duin, M. J., & Hendriks, A. M. T. (1995). *Watersnood in 1995: Een terugblik voor de toekomst* [Water disaster in 1995: A review for the future] (Tech. Rep. No. 1). Arnhem, the Netherlands: NIBRA.
- Van Eck, R. (2006). Digital game-based learning: It is not just the digital natives who are restless. *EDUCAUSE*, 41(2), 16–30.
- van Ruijven, T. (2008). Serious games as experiments for emergency management research: A review. In M. A. Santas, L. Sousa, & E. Portela (Eds.), *Proceedings of the 8th International Conference on Information Systems for Crisis Repsonse and Management*. Lisbon, Portugal.

- Vogel, J. J., Vogel, D. S., Cannon-Bowers, J., Bowers, C. A., Muse, K., & Wright, M. (2006). Computer gaming and interactive simulations for learning: A meta-analysis. *Journal of Educational Computing Research*, 34(3), 229–243. doi: 10.2190/FLHV-K4WA-WPVQ-H0YM
- Von Ahn, L. (2006). Games with a purpose. *Computer*, 39(6), 92–94. doi: 10.1109/MC.2006.196
- Von Ahn, L., & Dabbish, L. (2004). Labeling images with a computer game. In E. Dykstra-Erickson & M. Tscheligi (Eds.), *ACM CHI 2004 Conference on Human Factors in Computing Systems* (pp. 319–326). Vienna, Austria: ACM Press. doi: 10.1145/985692.985733
- Vroom, V. H. (1964). *Work and motivation*. New York, NY: Wiley.
- Warmelink, H. J. G., Meijer, S. A., Mayer, I. S., & Verbraeck, A. (2009). Introducing serious gaming in a multinational: experiences with the supervisor serious game for HSE training. In G. Y. Kin, Y. Cai, & Y. Gee (Eds.), *Learn to game, game to learn: Proceedings of the 40th ISAGA Conference* (pp. 41–57). Singapore: National University Singapore.
- Washbush, J., & Gosen, J. (2001). An exploration of game-derived learning in total enterprise simulations. *Simulation & Gaming*, 32(3), 281–296. doi: 10.1177/104687810103200301
- Weick, K. E. (1993). The collapse of sensemaking in organizations: The Mann Gulch Disaster. *Administrative Science Quarterly*, 38(4). doi: 10.2307/2393339
- Weick, K. E. (1995). *Sensemaking in organizations*. Thousand Oaks, CA: Sage.
- Weick, K. E., & Sutcliffe, K. M. (2001). *Managing the unexpected: Assuring high performance in an age of complexity*. San Francisco, CA: Jossey-Bass.
- Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the process of sensemaking. *Organization Science*, 16(4), 409–421. doi: 10.1287/orsc.1050.0133
- Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge, MA: Cambridge University Press.
- Whitney, W. D., & Smith, B. E. (1901). *The century dictionary and cyclopedia: A work of universal reference in all departments of knowledge with a new atlas of the world* (Vol. 11). New York, NY: The Century Company.
- Wilson, K. A., Bedwell, W. L., Lazzara, E. H., Salas, E., Burke, C. S., Estock, J. L., ... Conkey, C. (2009). Relationships between game attributes and learning outcomes. *Simulation & Gaming*, 40(2), 217–266. doi: 10.1177/1046878108321866
- Winn, B. M. (2009). The design, play, and experience framework. In R. E. Ferdig (Ed.), *Handbook of research on effective electronic gaming in education* (Vol. III, pp. 1010–1024). Hershey, PA: Information Science Reference. doi: 10.4018/978-1-59904-808-6.ch058
- Winn, B. M., & Heeter, C. (2006). Resolving conflicts in educational game design through playtesting. *Innovate Journal of Online Education*, 3(2). Retrieved from <http://innovateonline.info/index.php?view=article&id=392>
- Wouters, P., van der Spek, E. D., & van Oostendorp, H. (2009). Current practices in serious game research: A review from a learning outcomes perspective. In T. M. Connolly, M. Stansfield, & L. Boyle (Eds.), *Games-based learning advancements for multi-sensory human computer interfaces: Techniques and effective practices* (p. 232–250). Hershey, PA: Information Science Reference. doi: 10.4018/978-1-60566-360-9.ch014
- Yaman, M., Nerdel, C., & Bayrhuber, H. (2008). The effects of instructional support and learner interests when learning using computer simulations. *Computers & Education*, 51(4), 1784–1794. doi: 10.1016/j.compedu.2008.05.009
- Young, M. F., Slota, S., Cutter, A. B., Jalette, G., Mullin, G., Lai, B., ... Yukhymenko, M. (2012). Our princess is in another castle: A review of trends in serious gaming for education. *Review of Educational Research*, 82(1), 61–89. doi: 10.3102/0034654312436980
- Zagal, J. P. (2010). *Ludoliteracy: Defining, understanding, and supporting games education*. Pittsburgh, PA: ETC Press.
- Zichermann, G., & Cunningham, C. (2011). *Gamification by design: Implementing game mechanics in web and mobile apps*. Sebastopol, CA: O'Reilly Media.
- Zyda, M., Hiles, J., Mayberry, A., Wardynski, C., Capps, M. V., Osborn, B., ... Davis, M. J. (2003). Entertainment R&D for defense. *IEEE Computer Graphics and Applications*, 23(1), 28–36. doi: 10.1109/MCG.2003.1159611

# Summary

## Making sense of virtual risks

A quasi-experimental investigation into game-based training

### Introduction

Along with the rise of digital games over the past decades came an increased interest for using games beyond entertainment. So it happened that games have appeared to make children knowledgeable about their disease as well as games that help to make computers smarter. Although a few successes are known, much research—in particular about educational games—seems to suggest little evidence for games’ advantages. Such evidence makes clear that we need to speak of “the rise of a *potential* powerful tool.” Gaming has potential, theoretically and based on some of the “hints” from literature, but we need to figure out how to utilize and proof that potential.

The existing literature suggests that more studies are needed that investigate the *effective* design and use of games. Although extant research has taught us a number of lessons and the field seemed to have learned from the mistakes from the past, we still have no clear idea of how games are produced that provide reliably prespecified objectives.

In addition, the field is especially in dire need of comprehensive and *rigorous* studies, that is, studies that go beyond anecdotal, descriptive, or judgmental evidence and that do not suffer from methodological flaws. Such studies require *innovation* too, because we may not be able to make use of games effectively without it and/or capture what impact they have.

To contribute to this emerging field, the case of *Levee Patroller* was investigated. This unique game was developed in 2006 and its name refers to the game’s target group. Levee patrollers are considered the “eyes and ears” of the Dutch water authorities, which are organizations that are responsible for the water quality, quantity, and safety in the Netherlands. They inspect levees, the artificial and natural barriers.

ers that protect a region from flooding, and report any risks they encounter. Much similar to the actual practice, in the game players have to find all virtual failures in a region and report these. If they do not find the failures in time or report them incorrectly, it could result in a levee breach that floods the whole virtual region.

This case was investigated for two reasons. First, it may be a unique game, it is not alone. Many similar game-like digital technologies have also been developed in the past decade, such as for training first responder response to hazardous materials and triaging patients during a crisis. These technologies have in common that they are situated in the same domain, that of *safety and crisis response*. They further attempt to bring forth the same value and by similar means. They aim for *knowledge* about risks and achieve this by means of *sensemaking*, which is roughly defined as a process by which people give meaning to phenomena. Finally, they even use a similar type of game genre. Each can be considered a 3D simulation. Therefore, investigating *Levee Patroller* helps to shed light on a particular specialization within the emerging use of games for serious purposes: the use of games to make sense of risks. Second, although *Levee Patroller* might be unique, it provides for a unique opportunity to contribute to maturing the field. Little is known about the use of games in the domain of safety and crisis response, a domain for which gaming has an incredible potential. In addition, unlike the known closely related technologies, this game has been fully developed to facilitate many hours of training and it found an actual application, as five water authorities participated in its development and wanted to build a curriculum around it.

The objectives behind the investigation was two-fold. The first objective relates to the dire need for evidence about the effectiveness of games. This objective was to design and implement an innovative game-based training intervention and evaluate its effectiveness in a comprehensive and rigorous manner. The following questions are associated with this objective:

1. What is the effectiveness of the training with *Levee Patroller*?
2. What factors contribute to its effectiveness?

Because so little is known about game-based training and in particular regarding the domain of safety and crisis response, the second objective was to develop a substantiated understanding of what makes a game successful in training practitioners to make sense of risks. Such understanding would be developed by considering the following questions:

1. How do participants experience the game-based training?
2. How do participants play the game?

However, what was really aimed for in this investigation was to establish a *thick description* of a game-based training. The study did not only aim for measuring the results, but also for providing a context from how these results were established. Because a mix of methods and methodologies were used to get this description, we could speak of establishing a “thicker description.”

For designing and implementing the training, ten evaluation principles were kept in mind. These principles are based on the state of the field and on how it could



move forward. Among others, the principles advocate a focus on the game itself, a consideration of the gameplay, a need for continued practice, an evaluation with its actual target group, and a setup that considers more than just the game. These evaluation principles declared the focus, scope, and assumptions behind the study.

### Learning objectives

The game was developed in an attempt at the Dutch water authorities to professionalize their members, including the patrollers, for dealing with flood risks. Another reason for its development is that the patrollers have to deal with rare but disastrous failures. Levee failures hardly occur and it is difficult to get any practical experience. In fact, despite of its “virtuality,” the game provides the only means to get experience in finding and reporting levee failures.

In developing the game, five learning objectives were identified: observing, reporting, assessing, diagnosing, and taking measures. These objectives indicate that the game involves knowledge: knowledge about recognizing failures and how to deal with them. In the end, the game is especially about teaching an ability to make sense of (virtual) risks, which involves technical skills called *sensemaking skills*.

The game achieves its objectives by engaging players in a process with many challenges that have to be made sense of and by influencing players’ meaning construction by steering into a preferred direction. The game provides a structure of what failures exist and how they need to be recognized and dealt with. Despite of this steering, the construction may still widely differ, because sensemaking processes do not occur in a vacuum. Players’ history, culture, identity, and other factors play a role in how knowledge is (re-)constructed.

The game-enabled sensemaking process may have an impact on *communication* too. The game was not developed for this impact, but arguably the game provides for a *common vocabulary* and *shared experience* that will make communication easier between the different actors involved in the inspection process. Speaking the same language and having a similar experience to draw upon will more likely lead to sharing the same meaning.

In short, these concepts relate to each other as follows: the game enables sense-making; this sensemaking leads to knowledge; knowledge enables the development of (technical) skills; and acquiring knowledge and skills from a similar sensemaking process may have an impact on communication.

### The training/evaluation

Back in 2010 it was not known if the game worked, which is due to the mere fact that it was barely used and most certainly not as it should have been used. To fulfill this gap a *training/evaluation* was set up. It is deliberately called a training/evaluation because it concerns a design of a training with *Levee Patroller* as well as a design of its evaluation. Both designs are tied together and have slightly different purposes,

which led to some inherent tensions in the design and execution. With the training the purpose was to improve the trainees; with the evaluation the purpose was to get “objective” results about how the game improved the trainees.

The evaluation was grounded in *mixed methods research* and *quasi-experimental design*. Its unit of analysis concerned the (individual) players and the main outcomes that were considered are *judgments*, *knowledge perception*, and *sensemaking performance*. Judgments are about how players appreciated the game/training; knowledge perception refers to players’ self-assessment on the various learning objectives; and sensemaking performance is about how practitioners are able to deal with failures. Communication, operationalized by vocabulary usage, word count, and dispersion of responses, and affective learning outcomes, such as perceptions regarding levee inspection and a possible heightened awareness, were considered secondary outcomes. The following mix of qualitative and quantitative methods were used to measure and validate the outcomes and the likely moderating variables:

*Pre- and post-questionnaire* Before and after the training participants made a self-assessment of their knowledge and attitudes toward levee inspection. The pre-questionnaire was further used to gather contextual variables, such as age and game attitude, and the post-questionnaire to determine how participants judged the training.

*Pre- and post-sensemaking test* To determine the sensemaking performance, participants needed to make sense of virtual and real failure pictures before and after the training. To refrain from making any sense for them and to see an impact on communication, participants needed to answer open questions.

*Game questionnaire* After every exercise participants had to answer a small questionnaire based on a number of closed and open questions. This was used to understand how participants experienced a particular exercise and see how their experience with the game might change over time.

*Game data* Each exercise resulted in game data. This game data consists of quantitative data and qualitative data of how the participant played an exercise. With this data a reconstruction was made of how participants made sense of virtual failures.

*Pre- and post-interviews* Before and after the training a number of selected participants were interviewed (20 in total) with the purpose to get to know who these patrollers really are, test patrollers’ knowledge in alternative ways, and validate the sensemaking test.

*Discussion* At the end of the training a discussion was organized to discuss the effectiveness, suitability, and future of the game-based training with *Levee Patroller*. A number of statements were used to guide the discussion.

*Students* A part of the training was implemented with student-participants to a) assess how knowledgeable patrollers are compared to novices at the start of the training; and b) see how patrollers play the game compared to computer-skilled people.

*Super experts* The sensemaking test was also completed by super experts who are specialists in levee inspection. This enables one to see how patrollers perform compared to these super experts.

*Field exercise* Instigated by one of the participating organizations, a half year after the training a group who received the game-based training (Game Group) was compared with a group who did not during a field exercise (Control Group). Perceptions and communication were the focus of this comparison.

Based on the evaluation principles and strategy, cognitive load theory, and some common sense ideas about willingness and commitment, a structured three-week training was developed with *a)* a special research version of *Levee Patroller* that includes eight exercises, three regions, full responsibilities, and an increasing difficulty; *b)* a start- and end-meeting on a workweek evening; *c)* weekly assignments with a number of exercises to complete at home; and *d)* a website and a manual as instructional support. The start-meeting was meant to prepare participants to play at home and the end-meeting was used for the discussion. At both meetings participants completed the questionnaires and tests.

### Setup and implementation

Three water authorities agreed to participate and the training/evaluation was implemented at each. One water authority saw the training as an opportunity to revamp its relationship with its patrollers. They had not organized much for years. The second authority was convinced about the game's usefulness but did not know how to implement a game-based training. For them the training was an opportunity to find out if this was a possible way. The third authority still had to be convinced and for this reason they also proposed to compare a Game Group with a Control Group during a field exercise.

The setup differed per authority, in terms of training administration, recruitment, location, support, compensation, and its premise. Especially its premise, whether it was voluntary or compulsory, made a difference. Participants—in particular the volunteers—disliked the fact that it was made compulsory.

The total number of participants came down to 147. These participants were relatively old ( $M = 47.6$ ;  $SD = 12.1$ ); were practically all male; had a mixed education and diverse occupations; had little failure and game experience; and had little computer skills.

Of this number 5% dropped out of the training and for various reasons—predominantly a dislike regarding the game. However, in general the game-based training can be conveyed as successful. A large majority (80%) played at least five out of six exercises at home, which is a participation rate beyond what was expected.

### Results

In addition to the unexpected positive participation rate, these are the major findings for concluding that the game can be considered effective:

- Participants indicated to have learned from it on all matters that are important to the inspection.
- Participants became able to make sense of failures—whether virtual or real—more accurately and with fewer words.
- Participants were more knowledgeable than students before the training. After the training they approximated the level of sensemaking by (super) experts.
- Participants used far more the same vocabulary and among participants the variation in answers decreased.
- Participants became more aware about the practice of levee inspection and the need for training. They also gained more self-confidence about inspecting failures and became motivated to learn more.
- Participants judged the game positively and found the game useful, also a half year after the training ended, and indicated that the training setup was perfect as is.
- Participants started to think and reflect more about their practice—they became *reflective practitioners*.

A complex network of factors play a role in the game's effectiveness. Surprisingly, it is not influenced by a positive game attitude, commitment to inspection, motivation to learn, and high expectations about the training. These factors determined the appraisal of the training but not its outcomes, suggesting that the game was able to enthuse participants about levee inspection.

Factors that were of importance are computer skills, age, and education. Highly skilled and educated younger participants were able to retrieve more out of the training than others. The study with students confirmed this. However, during the discussion the participants clearly stated that this game was also for their generation—despite of their lack in computer skills.

The most important factor concerns how the game is played. The average game score turned out to have strong relationships with all of the training outcomes: better players learn more. What helps in achieving a higher score is playing more. The number of exercises played had an influence on the game score and consequently on the outcomes.

Investigation of how players experienced the game revealed that they had to learn how to “read” the game. It took them some exercises to learn the controls and get used to the virtual environment. After acclimatizing to this new environment, participants find it more fun and realistic and also perceive to learn more.

This investigation also revealed that players have been frustrated throughout. From this we can learn that a game is a sensitive medium. Little things could disrupt or frustrate a player.

Exploration into how players actually played the game learned us that the game maybe reached half of its potential. Participants may have learned from some of the feedback, but generally they became better in playing the game and not in making sense of virtual risks. This tells us that in terms of effective design, much is still to be gained and especially because how players perform in the game is most crucial.

## Conclusion

The game-based training with *Levee Patroller* provides for clear evidence that games can be used to help practitioners make sense of risks. The evidence also suggests that it has an impact on the communication between practitioners. These positive results confirm that games are a potential powerful tool to enable players to make sense of phenomena, a usage which has been termed *sensegaming*. Future research should explore this further as well as the possibilities of games as research method and as assessment tool. Other recommendations are to consider the use of game analytics and to explore educational game design patterns.

# Samenvatting

## Het geven van betekenis aan virtuele risico's

Een quasi-experimenteel onderzoek naar een op een spel gebaseerde training

### Introductie

Met de opkomst van digitale spellen is er in de afgelopen decennium de interesse gewekt voor het gebruik van spellen voor meer dan alleen vermaak. Zo zijn spellen ontwikkeld om kinderen te laten leren over hun ziekte alsmede spellen om computers slimmer te maken. Hoewel een aantal successen bekend zijn, suggereert veel onderzoek, met name over educatieve spellen, weinig bewijs voor het nut van games. Dit gebrek aan bewijs maakt duidelijk dat we moeten spreken over de 'opkomst van een in *potentie* sterk middel'. Spellen hebben potentie, op basis van de theorie en gebaseerd op aanwijzingen in de literatuur, maar we moeten nog uitzoeken hoe we deze potentie inzetten en bewijzen.

De bestaande literatuur suggereert dat meer studies nodig zijn die het *effectief* ontwerp en gebruik van spellen bestuderen. Ondanks dat uit bestaand onderzoek diverse lessen kunnen worden getrokken en het vakgebied van de fouten in het verleden heeft geleerd, hebben we nog steeds geen duidelijk idee van hoe spellen worden ontworpen die op betrouwbare wijze van te voren vastgestelde leerdoelen weten te bereiken.

Daarnaast heeft het vakgebied dringend behoefte aan uitvoerige en *rigoreuze* onderzoeken, dat wil zeggen onderzoeken die voorbijgaan aan het anekdotische, beschrijvende en opiniërende bewijs en die niet lijden aan methodologische zwakheden. Zulke onderzoeken hebben ook *innovatie* nodig, want zonder kunnen we wellicht niet effectief gebruikmaken van spellen en/of meten wat voor impact ze hebben.

Om aan dit opkomende gebied bij te dragen is de casus van *Dijk Patrouille* onderzocht. Dit unieke spel was ontwikkeld in 2006 en de naam refereert aan de doelgroep. Patrouille lopers (of dijkwachters) worden gezien als de ‘ogen en oren’ van de Nederlandse Waterschappen. Deze organisaties zijn verantwoordelijk voor de waterkwaliteit, waterkwantiteit en waterveiligheid in Nederland. De lopers inspecteren de dijken, wat kunstmatige en natuurlijke barrières zijn die een regio tegen een overstroming beschermen, en rapporteren elk risico die zij tegenkomen. Vergelijkbaar met de praktijk moeten spelers in het spel alle virtuele schadebeelden in een regio vinden en deze rapporteren. Als zij de schadebeelden niet op tijd vinden of incorrect rapporten, dan kan dit resulteren in een dijkdoorbraak dat het volledige virtuele gebied laat overstromen.

De casus was om twee redenen onderzocht. Ten eerste, hoewel het een uniek spel betreft, is het niet de enige in zijn soort. Veel vergelijkbare spelachtige digitale technologieën zijn in het afgelopen decennium ook ontwikkeld, zoals voor het trainen van eerste respons hulp bij gevaarlijke stoffen en voor een triage van patiënten gedurende een crisis. Deze technologieën hebben gemeen dat ze zich in hetzelfde domein bevinden, namelijk *veiligheid en crisis respons*. Ze streven verder hetzelfde nut na en doen dat op dezelfde manier. Ze richten zich op *kennis* over risico’s en verwezenlijken dit door middel van *betekenisverlening*, dat grofweg gedefinieerd is als een proces waarbij mensen betekenis geven aan verschijnselen. Ze gebruiken zelfs hetzelfde speltype. Iedere technologie kan gekenmerkt worden als een 3D simulatie. Kortom, het onderzoeken van *Dijk Patrouille* helpt om licht te werpen op een specifieke specialisatie binnen de opkomst van spellen voor serieuze doelen: het gebruik van spellen om betekenis te geven aan risico’s.

Ten tweede, hoewel *Dijk Patrouille* wellicht uniek is, geeft het ook een unieke mogelijkheid om bij te dragen aan het vakgebied. Weinig is bekend over het gebruik van spellen in het domein van veiligheid en crisis respons, wat een domein is waar spellen veel potentie hebben. Anders dan de bekende vergelijkbare technologieën is dit spel volledig ontwikkeld om vele trainingsuren mogelijk te maken en heeft het een daadwerkelijke toepassing gevonden. Vijf Waterschappen hebben deelgenomen in de ontwikkeling van het spel en wilden een curriculum rondom het spel bouwen.

De doelstellingen achter het onderzoek waren tweeledig. De eerste doelstelling is gerelateerd aan de dringende behoefte voor meer bewijs over de effectiviteit van spellen. Deze doelstelling was om een innovatieve op een spel gebaseerde training te ontwerpen en implementeren en om vervolgens de effectiviteit daarvan op een uitvoerige en rigoreuze wijze te evalueren. De volgende vragen horen bij deze doelstelling:

1. Wat is de effectiviteit van de training met *Dijk Patrouille*?
2. Welke factoren dragen bij aan de effectiviteit?

Omdat zo weinig bekend is over de inzet van spellen om mensen te trainen, en in het bijzonder met betrekking tot veiligheid en crisis respons, was de tweede doelstelling om een op feiten gevormde begrip te vormen over wat een spel succesvol maakt in het trainen van vakmensen om betekenis te geven aan risico’s. Zulk begrip zou bereikt kunnen worden aan de hand van de volgende vragen:

1. Hoe ervaren participanten een op een spel gebaseerde training?
2. Hoe spelen de participanten het spel?

Maar wat in dit onderzoek echt beoogd werd is om een ‘dikke beschrijving’ (thick description) te bewerkstelligen van een op een spel gebaseerde training. Het onderzoek had niet alleen ten doel om resultaten te meten, maar vooral ook om de context weer te geven van waaruit deze resultaten bereikt werden. Omdat een mix van methoden en methodologiën gebruikt zijn voor het maken van deze beschrijving, kunnen we ook spreken van het bewerkstelligen van een ‘dikkere beschrijving’.

Voor het ontwerpen en implementeren van de training zijn tien evaluatieprincipes in acht genomen. Deze principes zijn gebaseerd op de stand van zaken in het vakgebied en hoe er progressie geboekt kan worden. De principes pleiten onder andere voor een focus op het spel zelf, een beschouwing van hoe er gespeeld wordt, een noodzaak voor oefenen, een evaluatie met de echte doelgroep en een opzet dat meer is dan alleen het spel. Deze evaluatieprincipes hebben de focus, scope en assumpties van het onderzoek bepaald.

## Leerdoelen

Het spel was ontwikkeld in een poging van de Nederlandse Waterschappen om haar leden te professionaliseren, inclusief de dijkwachters, voor het omgaan met overstromingsrisico's. Een andere reden voor de ontwikkeling is dat dijkwachters te maken hebben met zeldzame maar gevaarlijke schadebeelden. Schadebeelden aan dijken komen nauwelijks voor en het is hierdoor moeilijk om praktijkervaring op te doen. Ondanks de 'virtualiteit' biedt het spel in feite de enige mogelijkheid om ervaring op te doen met het vinden en rapporteren van dit soort schadebeelden.

Tijdens de ontwikkeling zijn vijf leerdoelen geïdentificeerd: observeren, rapporteren, inschatting maken, diagnosticeren, en het nemen van maatregelen. Deze leerdoelen wijzen er op dat het spel om kennis gaat: kennis over het herkennen van schadebeelden en hoe hier mee omgegaan moet worden. Uiteindelijk dient het spel vooral de vaardigheid van het geven van betekenis aan (virtuele) risico's aan te leren, wat in het onderzoek *betekenisverlening vaardigheden* wordt genoemd.

Het spel bereikt deze doelen door spelers te engageren in een proces met veel uitdagingen waar betekenis aan gegeven dient te worden en door de constructie van betekenis te beïnvloeden door spelers in een bepaalde richting te sturen. Het spel beschikt namelijk over een bepaalde structuur van welke schadebeelden bestaan en hoe deze moeten worden herkend en behandeld dienen te worden. Ondanks het richtinggeven kunnen de constructies toch nog steeds flink verschillen, omdat betekenisverleningsprocessen nu eenmaal niet in een vacuüm plaatsvinden. De geschiedenis, cultuur en identiteit van de speler spelen naast andere factoren een rol in hoe kennis wordt ge(re)construeerd.

Het betekenisverleningsproces gedreven door het spel heeft wellicht ook een impact op *communicatie*. Het spel was niet ontwikkeld voor dit doel, maar het valt te beargumenteren doordat het spel voor een gezamenlijk vocabulair en een gemeenschappelijke ervaring zorgt en dit maakt de communicatie tussen verschillende be-



trokken actoren in het inspectieproces gemakkelijker. Het spreken van dezelfde taal en het hebben van dezelfde ervaring zal er eerder toe leiden dat dezelfde betekenis gedeeld wordt.

In het kort relateren de concepten als volgt tot elkaar: het spel maakt betekenisverlening mogelijk; deze betekenisverlening leidt tot kennis; de kennis maakt het mogelijk om vaardigheden te ontwikkelen; en het verkrijgen van kennis en vaardigheden vanuit vergelijkbare betekenisverleningsprocessen zal een impact hebben op communicatie.

## De training/evaluatie

In 2010 was het onbekend of het spel werkte, wat vooral komt doordat het spel nauwelijks gebruikt werd en zeker niet hoe het gebruikt zou moeten worden. Om dit gat te vullen werd de *training/evaluatie* opgezet. Met opzet is dit een training/evaluatie genoemd, want het betreft een ontwerp van een training met *Dijk Patrouille* alsmede een ontwerp van de evaluatie. Beide ontwerpen zijn met elkaar verweven, maar hebben verschillende doelstellingen, wat tot diverse spanningen in het ontwerp en uitvoering heeft geleid. Met de training was de doelstelling om de participanten te verbeteren; met de evaluatie was de doelstelling om een 'objectieve' resultaten te bemachtigen over hoe het spel de participanten verbeterd heeft.

De evaluatie was gebaseerd op een combinatie van *mixed methods onderzoek* en een *quasi-experimenteel ontwerp*. De analyse-eenheid betrof de (individuele) spelers en de voornaamste uitkomsten die in acht genomen werden, zijn *oordelen*, *kennisperceptie* en *betekenisverleningsprestatie*. Oordelen gaan over hoe spelers het spel/de training waarderen; kennisperceptie refereert aan de zelfinschatting van spelers op de verschillende leerdoelen; en betekenisverleningsprestatie gaat over hoe participanten omgaan met de schadebeelden. Communicatie, wat geoperationaliseerd is door te kijken naar vocabulair gebruik, het aantal woorden en de verspreiding van gebruikte concepten, en affectieve leeruitkomsten, zoals percepties aangaande de dijkinspectie en een verhoogd bewustzijn, werden beschouwd als secundaire uitkomsten. De volgende mix van kwalitatieve en kwantitatieve methoden zijn gebruikt om de uitkomsten en de moderator variabelen te meten en te valideren:

*Voor- en achteraf vragenlijst* Voor- en achteraf maakten participanten een zelfinschatting over hun kennis en attitudes met betrekking tot de dijkinspectie. De vooraf vragenlijst werd verder gebruikt om achtergrondvariabelen te meten, zoals leeftijd en spelattitude, en de achteraf vragenlijst om te bepalen hoe participanten de training beoordeeld hadden.

*Voor- en achteraf betekenisverleningstest* Om de betekenisverleningsprestatie van participanten te beoordelen dienden zij betekenis te geven aan virtuele en echte schadebeeldenfoto's, zowel voor- als achteraf aan de training. Participanten dienden hierbij open vragen te beantwoorden, dit om te voorkomen dat er al betekenis gegeven was en om een impact op communicatie te kunnen zien.

*Spelvragenlijst* Na elke oefening moesten participanten een korte vragenlijst beantwoorden met een aantal gesloten en open vragen. Dit werd gebruikt om te be-

grijpen hoe participanten een bepaalde oefening hadden ervaren en om te zien hoe hun ervaring mogelijk verandert gedurende de training.

*Speldata* Elke oefening resulteerde in speldata. Deze speldata bestaat uit kwantitatieve en kwalitatieve data over hoe een participant een oefening gespeeld heeft. Met behulp van deze data is een reconstructie gemaakt van hoe participanten betekenis hebben gegeven aan de virtuele schadebeelden.

*Voor- en achteraf interviews* Voor- en achteraf zijn een aantal geselecteerde participanten geïnterviewd (20 in totaal) met het doel om er achter te komen wie deze dijkwachters nu daadwerkelijk zijn, de kennis van de dijkwachters op alternatieve manieren te testen en de betekenisverleningstest te valideren.

*Discussie* Aan het einde van de training werd een discussie georganiseerd om de effectiviteit, geschiktheid en de toekomst van de op een spel gebaseerde training met *Dijk Patrouille* te bepalen. Een aantal stellingen werden gebruikt om deze discussie te begeleiden.

*Studenten* Een deel van de training werd geïmplementeerd met studenten om *a)* te bepalen hoe veel kennis dijkwachters hebben ten opzichte van nieuwkomelingen aan de start van de training; en *b)* te zien hoe dijkwachters het spel spelen ten opzichte van mensen met veel computervaardigheden.

*Superexperts* De betekenisverleningstest werd ook ingevuld door superexperts. Dit zijn specialisten in dijkspectie. Hiermee werd het mogelijk om dijkwachters met superexperts te vergelijken.

*Veldoefening* Aangezet door één van de deelnemende organisaties is een half jaar na de training een groep die de op een spel gebaseerde training heeft gevolgd (Spelgroep) tijdens een veldoefening vergeleken met een groep die dat niet gedaan heeft (Controlegroep). Percepties en communicatie waren de focus van deze vergelijking.

Gebaseerd op de evaluatieprincipes en evaluatiestrategie, ‘cognitive load theory’ en een aantal ideeën op basis van gezond verstand over bereidheid en commitment is een gestructureerde drieweekse training ontwikkeld met *a)* een speciale onderzoeksversie van *Dijk Patrouille* met acht oefeningen, drie regio’s, volledige verantwoordelijkheden en een toenemende moeilijkheid; *b)* een start- en afsluitbijeenkomst op een doordeweekse avond; *c)* wekelijkse opdrachten met een aantal oefeningen die thuis gespeeld moeten worden; en *d)* een website en een handleiding als ondersteuning. De startbijeenkomst was bedoeld om participanten gereed te krijgen om thuis te kunnen spelen en de afsluitbijeenkomst was bedoeld voor de discussie. Op beide bijeenkomsten hebben de participanten de vragenlijsten en testen gemaakt.

## Opzet en implementatie

Drie Waterschappen gingen akkoord met deelname en de training/evaluatie werd bij ieder geïmplementeerd. Eén Waterschap zag de training als een mogelijkheid om nieuw leven te blazen in hun relatie met de dijkwachters. Zij hadden niet veel

georganiseerd in de jaren voorafgaand aan de training. Het tweede Waterschap was overtuigd van het nut van het spel, maar wist niet hoe het geïmplementeerd kon worden. Voor hun was de training een mogelijkheid om te kijken of dit toevallig de manier was. Het derde Waterschap moest nog steeds overtuigd worden van het nut en om die reden stelden zij voor om een vergelijking te maken tussen een Spelgroep en een Controlegroep tijdens een veldoefening.

De opzet verschilde per Waterschap in de administratie, werving, locatie, ondersteuning, compensatie, en de voorwaarde. Met name de voorwaarde, of het vrijwillig of verplicht was, maakte een verschil uit. Participanten, en dan vooral de vrijwilligers, konden het niet waarden dat de training verplicht was.

Het totaal aantal participanten kwam neer op 147. Deze participanten waren relatief oud ( $M = 47.6$ ;  $SD = 12.1$ ); vrijwel allemaal mannelijk; hadden verschillende opleidingsniveaus en diverse beroepen; hadden weinig ervaring met schadebeelden en het spelen van spellen; en hadden weinig computervaardigheden.

Van dit aantal verliet 5% de training en om verschillende redenen, voornamelijk vanwege een aversie ten opzichte van het spel. Desondanks kan de training als een succes gezien worden. Een overgrote meerderheid (80%) speelde minstens vijf van de zes oefeningen thuis en dat is een betere participatie dan wat vooraf voor mogelijk werd gehouden.

## Resultaten

Naast deze onverwacht positieve participatie zijn dit de belangrijkste bevindingen waarmee geconcludeerd kan worden dat het spel als effectief beschouwd kan worden:

- Participanten geven aan te hebben geleerd op alle aspecten die belangrijk zijn voor de inspectie.
- Participanten zijn in staat om met een betere accuraatheid en minder woorden betekenis te geven aan schadebeelden, of deze nu virtueel of echt zijn.
- Participanten hadden meer kennis dan studenten vooraf. Achteraf benaderden zij in het geven van betekenis aan risico's het niveau van (super) experts.
- Participanten gebruikten veel meer dezelfde vocabulair en tussen participanten verminderde de variatie in antwoorden.
- Participanten werden meer bewust over dijkinspectie en de noodzaak van training. Ze verkregen ook meer zelfvertrouwen in het inspecteren van schadebeelden en raakten gemotiveerd om meer te leren.
- Participanten beoordeelden het spel positief, vonden het nuttig, ook een half jaar na de training, en gaven aan dat de opzet zoals in de training perfect was.
- Participanten begonnen te denken en te reflecteren over de praktijk, wat een indicatie is dat ze *reflectieve professionals* geworden zijn.

Een complex netwerk van factoren speelt een rol in de effectiviteit van het spel. Het wordt verbazingwekkend niet bepaald door een positieve spelattitude, commitment aan de inspectie, motivatie om te leren en hoge verwachtingen over de training.

Deze factoren bepaalden de waardering van de training maar niet de uitkomsten. Dit suggereert dat het spel in staat was om participanten te enthousiasmeren over dijkinspectie.

Factoren die wel van belang waren, zijn computervaardigheden, leeftijd en educatie. Jonge deelnemers met flinke computervaardigheden en een hoog opleidingsniveau konden meer uit de training halen dan anderen. Het onderzoek met de studenten bevestigt dit. Tijdens de discussie stelden de participanten echter dat dit spel ook voor hun generatie was, en dat ondanks hun gebrek aan computervaardigheden.

De belangrijkste factor betreft hoe het spel gespeeld is. De gemiddelde spelscore bleek een sterke relatie te hebben met alle uitkomsten: betere spelers leren meer. Wat helpt in het bereiken van een hogere score is om meer te spelen. Het aantal gespeelde oefeningen had een invloed op de spelscore en dus ook op de uitkomsten.

Onderzoek naar hoe spelers het spel ervaren hadden, laat zien dat ze moesten leren om het spel te 'lezen'. Het duurde een aantal oefeningen voordat spelers de besturing geleerd hadden en aan de virtuele omgeving gewend waren. Na aan deze nieuwe omgeving gewend te zijn, vonden participanten het leuker, meer realistisch en gaven ze ook aan er meer van te leren.

Dit onderzoek laat ook zien dat spelers gedurende het spel gefrustreerd waren. Hieruit kunnen we afleiden dat een spel een sensitief medium is. Kleine zaken kunnen een speler storen of frustreren.

Verkenning van hoe spelers het spel gespeeld hebben, toont aan dat het spel wellicht slechts half van zijn potentie heeft bereikt. Participanten hebben mogelijk wat geleerd van de terugkoppeling, maar over het algemeen werden ze beter in het spelen van het spel, maar niet in het geven van betekenis aan virtuele risico's. Dit betekent dat als het om effectief ontwerp gaat, dat er nog steeds veel te winnen valt en zeker omdat het cruciaal is hoe spelers presteren in het spel.

## Conclusie

De op een spel gebaseerde training met *Dijk Patrouille* levert duidelijk bewijs op dat spellen gebruikt kunnen worden om vakmensen te helpen om betekenis te geven aan risico's. Het bewijs suggereert ook dat het een impact heeft op de communicatie tussen vakmensen. Deze positieve resultaten bevestigen dat spellen een in potentie sterk middel zijn om spelers te leren om betekenis te geven aan verschijnselen. Dit gebruik wordt *sensegaming* genoemd. Toekomstig onderzoek dient dit verder te verkennen alsmede de mogelijkheden voor spellen als onderzoeksmethode en als beoordelingsmethode. Andere aanbevelingen betreffen het gebruik van 'game analytics' en een verkenning van educatieve spelontwerp patronen.

# Curriculum Vitae

Casper Harteveld was born in Delft, the Netherlands on November 23, 1982. He completed his secondary education at Grotius College in Delft in 2000. He obtained a Bachelor's Degree (2006) and a Master's Degree (2007, cum laude) in "Systems Engineering, Policy Analysis and Management" (SEPAM) at Delft University of Technology. He further received a Bachelor's Degree (2007) in Psychology at Leiden University in Leiden, studied a semester abroad at Carnegie Mellon University in Pittsburgh, PA, and completed a summer internship at the United Nations headquarters in New York City. From 2007 till 2012 Casper worked as Ph.D. Researcher at the Faculty of Technology, Policy & Management at Delft University of Technology and was affiliated with the research institute Deltares and the TU Delft Centre for Serious Gaming.



His research interests involve the design and evaluation of games with a serious purpose. Casper was the lead designer of the 3D digital game *Levee Patroller*, which has been exhibited at the NEMO Science Museum in Amsterdam and the Delft Science Centre, and (co-)designer and advisor of countless other games—from simple board and card games to virtual worlds. He is the author of the book *Triadic Game Design* (published by Springer in 2011), wrote over 20 other publications, and was guest editor for *Simulation & Gaming*. For his work he was awarded the Young Talent Prize of Information Systems from The Royal Holland Society of Sciences and Humanities.

Casper is currently an Assistant Professor at the College of Arts, Media and Design at Northeastern University in Boston, MA.

This page intentionally left blank

**Deltares Select Series**  
**Volume 11**  
**ISSN 1877-5608 (print)**  
**ISSN 1879-8055 (online)**

Deltares is an independent Dutch research institute for water, subsurface and infrastructure issues. Its mission is to develop and apply top-level expertise on these issues for people, planet and prosperity. Worldwide, around 900 staff are working to find innovative solutions that make living in the world's deltas, coastal areas and river basins safe, clean and sustainable. Since 2009 Deltares has been publishing the Deltares Select Series. Alongside peer reviewed proceedings of symposia organized by Deltares, this series publishes dissertations, edited volumes and monographs by Deltares employees. The editor-in-chief of the series is Dr. Jaap Kwadijk, Chief Scientist of Deltares. For more information, see [www.deltares.nl](http://www.deltares.nl).

Previously published in this series:

- Volume 10. G. Hoffmans, The influence of Turbulence on Soil Erosion
- Volume 9. A. te Linde, Rhine at Risk? Impact of Climate Change on Low-Probability Floods in the Rhine Basin and the Effectiveness of Flood Management Measures
- Volume 8. I. Pothof, Co-current Air-Water Flow in Downward Sloping Pipes
- Volume 7. M.A. Van, E.J. den Haan and J.K. van Deen (Eds.), A Feeling for Soil and Water, a Tribute to Prof. Frans Barends
- Volume 6. C. Schipper, Assessment of Effects of Chemical Contaminants in Dredged Material on Marine Ecosystems and Human Health
- Volume 5. M. Marchand, Modelling Coastal Vulnerability - Design and Evaluation of a Vulnerability Model for Tropical Storms and Floods
- Volume 4. S.A.M. Karstens, Bridging boundaries. Making Choices in Multi-actor Policy Analysis on Water Management
- Volume 3. J.S.M. van Thiel de Vries, Dune Erosion During Storm Surges
- Volume 2. P.M.S. Monteiro & M. Marchand (Eds.), Catchment2Coast, a Systems Approach to Coupled River-Coastal Ecosystem Science and Management
- Volume 1. F.J. Los, Eco-Hydrodynamic Modelling of Primary Production in Coastal Waters and Lakes Using BLOOM

This page intentionally left blank