An inquiry into gaming simulations for decision making
From design to knowledge management

Roungas, Bill

DOI
10.4233/uuid:4fcec9c1-a165-4429-ae43-1f75c0feb1a5

Publication date
2019

Document Version
Final published version

Citation (APA)
Roungas, B. (2019). *An inquiry into gaming simulations for decision making: From design to knowledge management*. [Dissertation (TU Delft), Delft University of Technology].
https://doi.org/10.4233/uuid:4fcec9c1-a165-4429-ae43-1f75c0feb1a5

AN INQUIRY INTO GAMING SIMULATIONS FOR DECISION MAKING

FROM DESIGN TO KNOWLEDGE MANAGEMENT

An Inquiry into Gaming Simulations for Decision Making

From Design to Knowledge Management

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology,
by the authority of the Rector Magnificus, prof. dr. ir. T.H.J.J. van der Hagen,
Chair of the Board for Doctorates,
to be defended in public on
Tuesday 4 June 2019 at 12:30 p.m.

By

Vasileios Roungas

Master of Science in Game and Media Technology,
Utrecht University, The Netherlands
and Master of Science in Business Mathematics,
Athens University of Economic and Business, Greece
born in Athens, Greece

This dissertation has been approved by the promotors

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. ir. A. Verbraeck,	Technische Universiteit Delft, promotor
Prof. dr. ir. S.A. Meijer,	KTH Royal Institute of Technology, Sweden, promotor

Onafhankelijke leden:	
Prof. mr. dr. J.A. de Bruijn	Technische Universiteit Delft
Prof. dr. R.M.P. Goverde	Technische Universiteit Delft
Prof. dr. J.T. Harviainen	University of Tampere, Finland
Prof. dr. ir. H. Kanegae	Ritsumeikan University, Japan
Prof. dr. A. Tsalgatidou	University of Athens, Greece

This research was fully funded by ProRail.

Keywords: Gaming simulations, Decision making, Complexity, Knowledge management, Game design, Game theory, Validation, Debriefing

Printed by: Pletsas, K., - Z. Kardari O.E.

Cover: Gaming Evolution by Dr. Konstantinos Roungas

ISBN 978-94-6384-044-6

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

Life must be lived like play.

Plato

Acknowledgements

The journey to achieve a Ph.D. is undoubtedly a long one, and in my case, it took even longer than usual. It has been almost two decades since I was first admitted to my undergraduate degree and up to today I have obtained one bachelor's degree, two master's degrees, and of course this Ph.D. During this quest, numerous people helped me most of whom have done it unconditionally. I would hence like to dedicate this small part of the thesis to thanking them.

I would first like to express my deep appreciation to my two supervisors, Professors Sebastiaan Meijer and Alexander Verbraeck. Sebas, as my daily supervisor, was not only tasked with guiding me towards a high-quality and timely fulfilment of my thesis but also with forging me to become an independent researcher upon obtaining the Ph.D. He was always present when I needed guidance but also gave me enough space to develop my own ideas and grow. On the other hand, Alexander, being my promotor, was at all times aware and fully informed of the state of things, through frequently held meetings, which would always left me dazzling on the amount of information that can be suppressed in one hour. The debates with, and overall guidance of both Sebas and Alexander restrained me from losing sight during the first chaotic years of the Ph.D. and eventually enabled me to publish more than 10 papers and complete my thesis on time.

The project I was a part of these 4 years was fully funded by ProRail, where many people helped me, some more than others, to gather crucial pieces of knowledge that eventually led in what I believe to have been a mutually beneficial relationship. While I cannot thank individually in these few lines every person that offered their time and energy to me, I would specifically like to thank Jelle van Luipen for being my go-to person with regards to ProRail and for enabling me to understand a previously totally unknown sector to me. Our meetings as well as our informal chats gave me insights on the complexity of the railway sector and made me change the way I look at things when entering a train or even when checking the timetable. I would also like to thank Wilco Tielman for being more helpful that he can actually imagine. Wilco would always find a solution to my problem and always with a smile in his face.

A special thank you goes of course to Femke Bekius with whom I shared my office at Delft and our undefined working space at ProRail. I want to thank her not only for bearing with me, though I believe I was easygoing, but also for shaping my research through our numerous discussions. I only hope that I have returned that favour to her. I would also like to thank Julia Lo for broadening my horizon in regards to the human element involved in games and with whom I had exciting discussions in the practical as well as the philosophical aspects of knowledge. Another person that enabled me to gain insights in an area little known to me, especially when I was still in the beginning of my Ph.D., was Marieke de Wijse. After months of diving into the literature, my discussions with Marieke revealed to me various aspects of debriefing and the struggle between theory and practice that is quite present in this field. During the penultimate year of my Ph.D., I had the honour of supervising two master students, Rachele and Jessica, with whom I got my first taste on the excitement and the challenges of being a supervisor. Both of them were excellent during our cooperation and they produced high quality research, especially in a master level. Lastly, at least from the North European front, I would like to thank several of my colleagues at TU Delft, like Orcun, Qasim, Arman, and Andreas, to name only but a few, with whom I spent quite some time outside the TBM building (for obvious reasons) during which time I learnt quite a lot on web security, DDoS attacks, and the Ramadan.

Whilst not directly related with my years as a Ph.D. researcher, I would like to thank two people that showed me what truly means to help someone unconditionally and to whom I will be eternally grateful. Professor Anastasios Karaganis, my econometrics professor during my undergraduate days, was the first academic that inspired me to move beyond my lack of excitement of the first years as a bachelor student and pursue my first master degree in financial mathematics. On top of that, he even supported me when I decided to leave everything behind and pursue new studies and subsequently a new career in games. Since we first met, his door has always been open for help, a piece of advice, or just an informal chat. Almost a decade ago, when I decided to change my professional course, at moments this endeavour seemed and was overwhelming. Professor Aphrodite Tsalgatidou during those times offered me tremendous help and support, which greatly smoothed my transition from economics to computer science and from Greece to the Netherlands.

My decision to change the course of my life and move to the Netherlands was perceived by many acquaintances of mine as irrational and juvenile. Perhaps it was; I was close to 30 with a master degree ready to boost my career, but this was not what I wanted. It was Ioanna and Kanta who helped me enormously to clear the fog in my head and overcome the feeling of insecurity that accompany someone who (at least) seemingly has an early midlife crisis. Words cannot express my gratitude towards them and the same applies for my parents and best friends, Mary and Kostas. Their unconditional love and support, both financially and emotionally wise, have made every decision of mine a bit easier. Knowing that regardless the successes or failures, there are people behind you that love and support you no matter what, is a true feeling of freedom. This thesis and most probably the majority of my academic achievements and my overall well being would have been absolutely impossible without them. Last but not least, I would like to express my love and gratitude to my girlfriend Sany, who not only never nagged for having a distance of more than 2000 kilometres between us the majority of the time, but was also extremely supportive, caring, and my go-to person for all my proofreading needs.

Contents

1

INTRODUCTION

There is nothing permanent except change.

Heraclitus

IN an era where the complexity of all things around us has dramatically increased, grasping, let alone fully comprehending, that complexity seems to be quite a challenging task. As a result, tools which, a few decades ago, used to satisfactorily provide insights on the functioning of certain parts of the world, can no longer address the complexity-derived challenges. There is a need for more rigorous methods that would enable the understanding of real-world problems within such an ever-changing complex environment. Gaming simulations is a discipline/method that has the potential to capture the complexity surrounding us and provides for a platform to better understand it. The primary reason is that gaming simulations incorporate by definition the "perpetrator" of the aforementioned complexity, humans. Since human behaviour cannot be characterised as 100% rational, systems involving humans often tend to behave in a seemingly unpredictable way, or what is called bounded rationality (Simon, 1957), hence the complexity and the subsequent challenges.

This thesis intends to identify and address challenges throughout the whole lifecycle of gaming simulations. Gaming simulations lie in the intersection of the analytical and design sciences, hence, first and foremost, this thesis aims at acknowledging the struggle between these two scientific communities. It subsequently goes as far as to combine methods from both fields, with intent to improve the decision-making quality of gaming simulations.

In Section 1.1, the term *systems*, which is used extensively in this thesis, along with how it has evolved through time, is introduced. In Section 1.2, gaming simulations are

Parts of this chapter have been published in Roungas et al. (2018h) and Roungas et al. (2019a).

defined and characterised according to their different attributes. In Section 1.3, a presentation of ProRail, the primary case study, is given. Finally, Section 1.4 provides an overview of the structure of the thesis.

1.1. SYSTEMS

According to Merriam-Webster (2018), a system is a regularly interacting or interdependent group of units forming an integrated whole. Systems can be physical, like the animal kingdom or the human body, or man-made, like an organisation. This thesis focuses on man-made systems, i.e. organisations, and particularly on those aspects of systems that are the intertwine of humans and technologies.

During the first decades of the 20^{th} century, labour studies were concerned with the adaptation of humans to the organisational and technical framework of production (Ropohl, 1999), where studies like that of Mayo (1946) acknowledged the human-factor in industrial relations. However, it was not until the '50s, when in Tavistock Institute in London, the term *Socio-Technical Systems* (STS) was coined (Trist and Bamforth, 1951). STS was the culmination of research based on the idea that the separate approaches to the technical and the social systems of organisations could no longer suffice (Trist, 1981). STS created a paradigm shift in organisational studies, which until then would not consider the "people cost", by identifying the interdependencies between the technological artefacts and humans (Trist, 1981).

A few decades later, the need to capture the evolving nature of certain systems, like economies, and their mechanisms gave birth to the term *Complex Adaptive Systems* (CAS) (Holland, 1992). CAS are composed of multiple interacting parts, physical parts and agents, that evolve and adapt their behaviour over time based on new information. CAS are also characterised by the fact that the whole is more than the sum of the parts. In other words, the properties of a CAS is not just the sum of the properties of each physical system and agent belonging in it (Holland, 1995). Whilst it might initially seem overwhelming to study CAS, especially in a formal way, there are techniques that are capable of capturing this emerging complexity, like control theory and game theory (Holland, 2006). Game theory, in particular, is later analysed in this thesis as a powerful tool for formalising game design.

1.1.1. SYSTEMS' COMPLEXITY

The complexity of systems, and particularly the complexity of the decision-making process within those systems, can be explored through three levels: technical, actor, and context complexity, which are analysed below. Given these three levels, it becomes evident that complexity is not just the result of systems' increased size but is mainly caused by the numerous interdependencies among the different aspects of those systems. In turn, while these interdependencies are abstracted to a certain degree, they still bear a significant amount of complexity, which needs to be translated into game design choices. The result is artefacts, i.e. games for decision making, characterised by numerous and complex structures with limited knowledge on behalf of researchers and practitioners on how to understand and model them.

TECHNICAL COMPLEXITY

With regards to technical complexity, a system can be viewed along three levels:

- Functionally organised in *aspect systems* (Veeke et al., 2008), each of which defines the main responsibilities for the actors involved. In order for the system to function properly, these aspects need to be aligned; so, an entire-system approach is necessary for the analysis.
- Geographically organised, in which case the system is divided into regions, where its functions come together. These regions can be viewed as *subsystems* (Veeke et al., 2008), with all the advantages and limitations this entails.
- Organised by distinguishing between the operational and strategic levels. In order for the system to function properly, these levels need to be aligned as well.

The technical complexity of the system depends on the number of technical changes or uncertainties, but also on how the system is viewed. A change in one aspect of the system usually requires alignment with one or more of other aspects and subsequently for many subsystems to adapt, which in turn influences both the operational and strategic level.

ACTOR COMPLEXITY

Decision-making processes on complex systems usually involve multiple actors with different perspectives and interests. These actors are not necessarily hierarchically organised, but in any case they are mutually dependent upon one another. As a result, they form a network of interdependencies. In such a network, the course of the decision-making depends on the behaviour of, and interactions between these actors (de Bruijn and ten Heuvelhof, 2008). This results in an often messy, spaghetti-like structure. Moreover, the formal structures actors have to work in and with are often hierarchical, which might give some actors a special position, thus making this even more complex.

CONTEXT COMPLEXITY

During the process of decision-making, both the network of actors involved and the content of problems and solutions might change over time. This dynamic behaviour is for a large part the result of many interdependencies (e.g., a change in one regional subsystem has effect on the national system, a change of actor's A behaviour might impact the behaviour of actor B, etc.). Moreover, decision-making processes are always impacted by unforeseen external developments such as political decisions, media attention, and technical innovations (Bekius et al., 2018).

1.1.2. MOTIVATION ON STUDYING SYSTEMS

The core reason for which humans are interested in studying systems is their seemingly inherent need to understand and control these systems (Casti, 1986). It is indeed a natural urge, not just out of curiosity, to deeply comprehend how certain systems work, in order to be able to improve them or at the very least adjust our behaviour according to their boundaries. As a result, the increased complexity of modern systems, as it has been described so far in this first chapter, requires new methods which can capture and adequately abstract the different elements of systems that cause this inflated complexity. While it seems to be easier said than done, gaming simulations as a discipline appear to possess that toolbox that would allow to tackle and understand complexity.

1.2. GAMING SIMULATIONS

Gaming simulations are a subgroup of simulations, in which human participants take part. The term *gaming simulations* has been characterised in the past as cumbersome (Thavikulwat, 2004); this is perhaps why numerous terms have been used to describe them, like simulations, games, serious games, simulators, human-in-the-loop, to name a few. Hereinafter, gaming simulations are referred to as games.

Games are imitations of real-world systems (RS) designed to solve a problem; their primary purpose is not merely to entertain but to educate, train, steer decision-making processes, etc. (Crookall, 2010, Michael and Chen, 2005, Zyda, 2005). From Plato (1866) and Aristotle (1916), to Locke (1712) and Rousseau (1979), all the way to Dewey (1966) and Piaget (1952), philosophers and pedagogues considered play, and consequently games, to be an important part of education. Yet, it was only after the first half of the 20^{th} century that simulation-based games became more widespread (Abt, 1970, Huizinga, 1949) and even took a few more years for games to be used in areas other than learning and training (Duke, 1974). Nowadays, games are one of the simulation techniques with the highest stakeholder engagement rates (Jahangirian et al., 2010) and they have become a powerful tool for organisations, which need to identify solutions in uncharted areas (Duke and Geurts, 2004). As a result, research in games has dramatically increased over the last decade (Roungas et al., 2018d).

1.2.1. GAME CHARACTERISATION

Games span in various different domains; the generated knowledge from games can be used for different purposes and they can be addressed to a diversified audience (Roungas et al., 2018h). Hence, any analysis on games should first be based on a comprehensive characterisation. Characterisation of games can occur at different levels depending on the criteria applied each time. This thesis adopts three characterisations, primarily the characterisation proposed by Grogan and Meijer (2017), as shown in Table 1.1, and to a lesser degree the characterisations proposed by Peters and Vissers (2004), as shown in Table 1.2, and Morrison and Meliza (1999) depicted in Figure 1.1.

The characterisation proposed by Grogan and Meijer (2017) is based on two criteria: the type of knowledge generated by the game and the stakeholders who are the beneficiaries of this knowledge. With regards to the type of knowledge generated, the authors distinguish between the two following categories: i. Generalisable knowledge, meaning that the knowledge acquired during the game provides for broad insights beyond the scope of a particular game scenario, and ii. Contextual knowledge, meaning that the knowledge acquired during the game provides for deep insights closely related to a particular game scenario.

With regards to the beneficiary of the generated knowledge, the authors again distinguish between two categories: i. Participants, meaning that the beneficiaries of the knowledge acquired during the game are persons who play the game, and ii. Principals, meaning that the beneficiaries of the knowledge acquired during the game are any stakeholders other than the participants, like decision-makers, managers, researchers etc.

It should be noted that this categorisation is not absolute, in the sense that the knowledge type and beneficiary are not Boolean variables. Instead, the categorisation should be treated more like a continuum, where some games fall more in the Generalisable/ Participant quadrant, hence considered to be more focused on learning, some other games fall in the Contextual/Principal quadrant, hence considered to be more focused on design, and so forth.

Table 1.1: Canonical Applications of Gaming Methods (Grogan and Meijer, 2017).

Knowledge Type	Knowledge Beneficiary	
	Participant	Principal
Generalisable	Teaching Experiential learning Dangerous tasks	Research Hypothesis generation and testing Artefact assessment
Contextual	Policy Organisational learning Policy intervention	Design Interactive visualisation Collaborative design

The characterisation proposed by Peters and Vissers (2004) is also based on two criteria, one of which is also the knowledge beneficiary. The other criterion pertains to whether a game has a predefined outcome or, at the very least, a predefined list of possible outcomes as opposed to having an unknown outcome. The former are characterised as *Closed Simulations* while the latter as *Open Simulations*. While the characterisation by Peters and Vissers (2004) is not directly adopted in this thesis, the criteria of open and closed games is mentioned often, especially for policy and decision making games, which are characterised as open.

Table 1.2: Applications of Simulation Games (Peters and Vissers, 2004).

Performance Criteria Defined in Advance?	Who has to learn from the simulation game?	
	Participant	Others
Yes (Closed Simulation)	Training/education	Assessment
No (Open Simulation)	Development/exploration	Research

Finally, the characterisation proposed by Morrison and Meliza (1999) is again based on two criteria but none of them are related to knowledge. Instead, this approach is concerned with the realism of games. Based on this typology, two major aspects of games are acknowledged, the social and the physical realism, as shown in Figure 1.1. Social realism refers to the level of realism of the game with regards to human interaction. Physical realism refers to the level of realism of the game with regards to the equipment and the environment. Based on these two aspects, any game can be classified in three categories:

- Constructive Simulations, which rely on simulated operators, equipment and situations. They refer to a computerised program, since it would be impossible to simulate an operator, whose all aspects are simulated, otherwise.

- Virtual Simulations, which rely on real operators in simulated equipment and situations. They refer to an interactive process, which might be computerised though not necessarily. Its key difference from the first type is that some of the processes during the simulation are handled by real people, which adds the element of interactivity. In this thesis, unless stated otherwise, the term game should be considered synonym to the term Virtual Simulations.
- Live Simulations, which rely on real operators and equipment in simulated situations. They refer to a hybrid situation, where apart from the human interference, the game also includes real equipment.

In Figure 1.1, three additional terms are introduced:

- Reality. According to Oxford dictionary, reality is the conjectured state of things as they actually exist, rather than as they may appear or might be imagined. Given reality's broad definition and the fact that, regardless the beliefs of someone, it is impossible to fully measure reality, the term "Abstract of reality" is introduced.
- Abstract of reality. Abstract of reality is defined as this part of the world that can be potentially measured throughout a study. Most probably, throughout the course of a study, it will not be possible to fully measure everything that can potentially be measured. Therefore, the term "Construct of reality" is introduced.
- Construct of reality. Construct of reality is defined this part of the world that can be potentially measured and will be measured throughout a study. This construct of reality includes all the data that are collected and all the knowledge that is possessed and that is acquired, during the period of a study.

1.2.2. GAME FRAMING

Based on the characterisation proposed by Grogan and Meijer (2017), this thesis focuses on games that belong in the lower right quadrant, as shown in Table 1.1. In other words, games that generate contextual knowledge and for which the knowledge beneficiary is the principal. Similarly, according to the characterisation proposed by Peters and Vissers (2004), this thesis focuses again on games that belong in the lower right quadrant, as shown in Table 1.2, i.e. open simulations from which stakeholders other than the participants learn. Finally, based on the characterisation proposed by Morrison and Meliza (1999), this thesis focuses on games that lie at the intersection of virtual and live simulations.

While the term gaming simulations might be perceived to describe games that use a computerised simulation model in their core, that should not be considered an absolute definition (Klabbers, 2009). As a result, throughout the course of this project, research spanned from pure simulations, to computer-based games, all the way to non-digital board or paper-based games, as shown in Figure 1.2. The exploration of pure simulations sheds light on the different aspects of simulation models and particularly with regards to validation. On the other hand, non-computerised games are commonly used by organisations due to their low cost and configurability, and they still fall under the definition of gaming simulations.

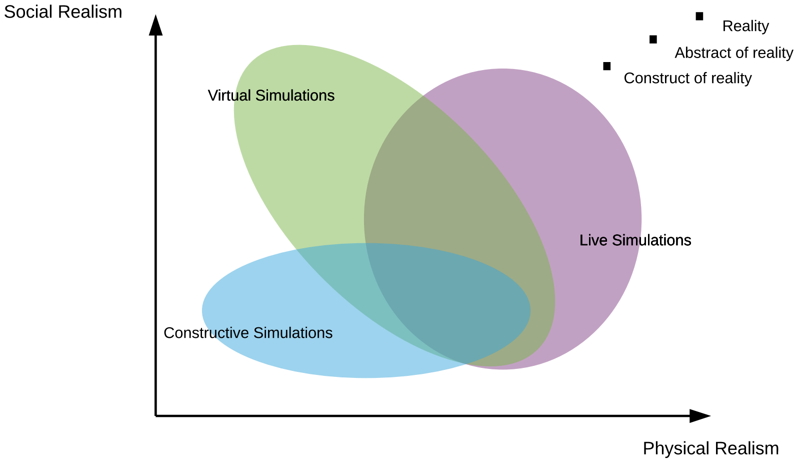


Figure 1.1: Typology of games and simulations.

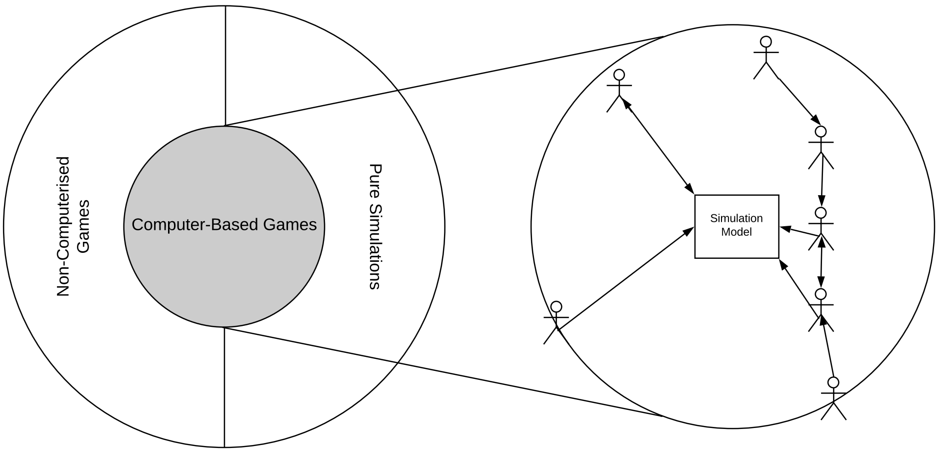


Figure 1.2: Illustration on the games and simulations explored in this thesis.

1.3. PRORAIL

ProRail, which is the primary case study in this thesis, is the Dutch government task organisation that takes care of maintenance and extensions of the national railway network infrastructure, of allocating rail capacity, and of traffic control. With regards to its processes, ProRail is a defacto engineering company, thus subject to the limitations of complex systems, as those are described in Section 1.1.1. As a result, the methods and conclusions of this thesis, while they could be generalised for any type of company, they are first and foremost focused on engineering organisations. ProRail should not be confused with the principal passenger railway operator in the Netherlands, which is Nederlandse Spoorwegen (NS), also known as Dutch Railways.

More than 15 years ago, ProRail started building simulations both internally and through the use of third-party packages, and has todate developed a wide range of simulations, extending from microscopic (Middelkoop and Loeve, 2006) and macroscopic (Middelkoop and Bouwman, 2001) to gaming simulations (Meijer, 2015, Middelkoop et al., 2012). In effect, ever since 2009 ProRail is using games to test future changes in the infrastructure, timetable, and processes, amongst other things (Meijer, 2012b). These games vary from testing with and training single train traffic operators to large scale multi-actor games that also involve operators from NS. ProRail's intention was, and still is, to develop a set of games as a means to test innovations in a safe environment (van den Hoogen, 2019).

In this thesis, several simulations and games from ProRail are used as case studies and for further analysis. In more detail, in Part I, two games from ProRail are used, along with a case from Sweden, with a view to developing and testing a framework for game design using game theory. Two microscopic simulations are analysed in Part II. Finally, in Part IV, three games are used to develop and test a knowledge management framework for games.

1.4. THESIS OVERVIEW

This thesis consists of 16 chapters. This chapter along with Chapter 2 and Chapter 3 serve as an introduction to the field of games, to the research gaps in this field, and to the research approach followed throughout the thesis respectively. Part I consists of Chapter 4 and Chapter 5 in which game theory is explored as a tool for game design. Part II consists of five chapters; in Chapter 7 and Chapter 8, a framework for selecting the most appropriate simulation validation and verification methods and statistical techniques is proposed. In Chapter 9 and Chapter 10, the implementation of an automated simulation validation process and the lessons learned from its application to the railways are described. In Chapter 6, the implications of the validity of games is explored. Part III consists of Chapter 11 and Chapter 12 in which methods and pitfalls associated with debriefing are identified, and Chapter 13 in which results from interviews with 21 facilitation and debriefing experts are presented. Part IV consists of Chapter 14 in which a knowledge management framework of games for decision making is proposed, and Chapter 15 in which the application of the framework in the railways is presented. Finally, in Chapter 16, concluding remarks are made.

2

Research Gaps in Games

Nothing exists except atoms and empty space;
everything else is just opinion.

Democritus

The limited adaptation of games in areas other than education has yielded numerous unaddressed problems with regards to their application and the applicability of the results derived from them. Several of these problems are associated with the design of games and the primary reason for that is that gaming as a discipline lies between the design and the analytical sciences. Moreover, the nature of these simulations, which involve human participation, significantly increases the associated cost. Subsequently, many game sessions are conducted with small sample sizes in terms of the number of participants, which has implications on the validity and the generalisability of the results. Finally, perhaps the least researched aspect of games is the overall management of the knowledge acquired through games. While this knowledge can create valuable insights, and at the same time is quite expensive to obtain, there is almost a complete lack of methodologies on how to manage it. Solving this can maximise the insights obtained from a game as well as allow the future use of the acquired knowledge for a small fraction of the cost.

2.1. Limitations in the Gaming Field

Games have several limitations, which are most likely a consequence of the small body of knowledge for games focused on applications other than learning and training (Duke and Geurts, 2004). These limitations are in close relationship with the different kinds of

Parts of this chapter have been published in Roungas et al. (2018h).

validity. In literature, various kinds of validity pertaining to games are identified. The most important ones are i. construct validity, ii. internal validity, iii. external validity, and iv. reliability for research and usage of the game insights and results (e.g. for policy making or decision making) (hereinafter referred to as game utilisation).

Construct validity is the degree to which a test measures what it is supposed to measure (Cronbach and Meehl, 1955). Consequently, for games, construct validity is the degree to which an artefact, i.e. a game, represents the construct from reality it is supposed to represent. With regards to the construct validity, games have been identified to be in the intersection of design and analytical sciences, which generates confusion, misunderstanding, and frustration between the two communities (design and analytical), due to the fact that they are working separately and that they have different success criteria. Another major limitation is the rare application of rigorous design methods in game design (Fumarola et al., 2012), which has two significant drawbacks. First, problematic areas in the real system that would benefit from gaming are not identified in a systematic way. Secondly, games in various occasions are not consistent with the real-world system they are suppose to imitate. Finally, an important game design decision that is often not in line with the purpose of the game is the abstraction level. Games span in various applications ranging from low-tech board games to high-tech simulation games (Meijer, 2015). Selecting the type of game, and consequently the abstraction level, depends on the objective of the game. Too much abstraction could lead to vague and unclear goals within the game. On the other hand, too little abstraction could result in strict rules that do not adhere to the real-world system under study and also it could not provide the flexibility to use games for future systems for which the rules are not known.

Internal validity refers to the degree to which the behaviour of the dependent variable is explained by the behaviour of the independent variable (Reis and Judd, 2014). For games, internal validity examines the degree to which the game is an adequate representative of the system it imitates. With regards to the internal validity of games, the greatest impediment is the informal way with which games are validated, e.g. face validation. Consequently, validation relies heavily upon experts' opinion and hence to their subjectivity. This limitation is related to the lack of design methods for games (mentioned in the previous paragraph) as well as to the usually low number of participants (mentioned in the next paragraph). These interconnections of the game characteristics not only reveal the complexity of games but also the need for methodologies that would first acknowledge and then model these interconnections.

External validity examines the degree to which results from a study can be valid and generalised for the population in general or even for different populations (Aronson et al., 2018). For games, external validity is closely related with the sample size of the game session (Grogan and Meijer, 2017) and debriefing (Seymour, 2012). The latter is responsible for transferring the acquired knowledge from the game session to reality, making it applicable in a real context, and thus generalisable. With regards to the external validity of games, the most important aspects are the sample size (number of participants) and debriefing. A small sample size is easy to obtain but has limited possibilities for analytical conclusions thus limited possibilities for generalising the observations from the game. A large sample size, while it solves the analytical problem and the generalisability of the results, it is usually expensive to obtain and difficult to coordinate. While debriefing has been identified to be the most important feature of games (Crookall, 2010), very few research articles report results on the effectiveness of debriefing (Dufrene and Young, 2014), and even less for games focused in areas other than education.

Game utilisation refers to the minimisation of errors and bias of a study, by ensuring that results are replicable (Yin, 2009), as well as the utilisation of the results and insights from a game in reality. In games, where the uncertainty introduced by humans eliminates the possibility for absolutely reproducible results (Hofmann, 2015), game utilisation is related with the way the knowledge created within the game is captured, documented, managed, and reused. With regards to the game utilisation, it has been strongly suggested that, even though it may be complex, the collection of empirical data from game sessions can create a body of knowledge that facilitates the researchers' understanding of systems' behaviour (Hofstede and Meijer, 2007). Despite that, the research on methodologies on management and reuse of knowledge derived from games is severely limited (Roungas et al., 2018d). The only exception has been the military, where large scale strategic games have been widely used (Herz and Macedonia, 2002), and methodologies for managing and reusing knowledge have been proposed and tested through the years. Nevertheless, the majority of those methodologies are domain specific and consequently the state of the art is specialised to the military; thus, most of the associated outcomes cannot be generalised and used in other domains.

Based on the literature reviewed and the definitions of the different kinds of validity, four main areas in game research can be identified:

1. Game design, which relates to the construct validity of games.
2. Game validation, as defined in simulations (Balci, 1998), which relates to the internal validity of games.
3. Game sessions, which relates to the external validity of games.
4. Knowledge management, which relates to the game utilisation.

Potential directions for improvements are discussed in Section 2.2 based on which, in Chapter 3, the actual contributions of this thesis are presented.

2.2. Potential Directions for Improvements

In this section, potential directions for improvements, based on the limitations presented in Section 2.1, are explored. Methodologies and work in-progress that can help mitigate or even fully address the limitations are identified. These methodologies concern only approaches focused in applications other than education and training. The four subsections are based on the four main areas of games, identified in Section 2.1.

2.2.1. Design

Game design as a discipline lies in the intersection of the design and analytical sciences with all the limitations that this entails, as identified in Section 2.1. Therefore, bridging the gap between those two communities will result in more robust, generalisable, and context rich theories.

One of the first attempts towards bridging the design and analytical aspects of games was made by Duke (1980), in which he identified 9 steps for game design: 1. Developing written specifications for game design, 2. Developing a comprehensive schematic representation of the real-world problem, 3. Selecting components of the real-world problem to be gamed, 4. Identifying the game components, 5. Providing content for the game components, 6. Providing the game mechanics, 7. Implementing and testing the game, 8. Evaluating the game, and 9. Field testing the game.

Point 1 refers to what is widely known as *Requirements*. In layman's terms, requirements are a list of needs communicated by the client that define the objective of the game (Bethke, 2003). Potential directions for improvements in this area could be to:

- Develop theories that would unify all aspects of games by incorporating notions from the traditional requirement engineering field (Goguen, 1994) as well as the social sciences, in the form of human factors and mental models (Lo et al., 2014), also known as emotional requirements (Callele et al., 2006).

- Enable seamless and effective communication between the stakeholders like game designers, decision makers etc. Harteveld (2011). A way would be to use gaming as it was proposed by Raghothama et al. (2017) in a participatory manner.

- Develop domain specific languages by using well-established methods, like the Unified Model Language (UML), that would allow to more effectively elicit the game requirements Bethke (2003).

- Develop tools for more effectively documenting the requirements and the game design choices throughout the whole design lifecycle Roungas and Dalpiaz (2016).

Point 2 refers to the identification of the problem in the real-world that needs to be gamified. This point relates to the limitation that problematic areas in the real system are not identified in a systematic way. Potential directions for improvements in this area could be to:

- Develop methodologies for more effectively understanding and analysing STS and CAS. Particularly, formal methods, like game theory, can be used to pinpoint the problematic areas of multi-actor systems (Vorobeychik and Wellman, 2009).

- Acknowledge and research the behaviour of actors within the scope of bounded rationality that seems to characterise many STS and CAS (Simon, 1972).

Point 3 refers to those parts of the real system that will be chosen to be part of the game; hence it heavily relies on point 2. Reality cannot be gamified in its full complexity, thus abstractions should be made. Hence, this point concerns the abstraction level of the game. A potential direction for improvement in this area requires the understanding of the advantages and disadvantages of the different levels of abstractions (Meijer, 2012a). In turn, given the different characteristics of games and the real system under study (game objective, intended audience, etc.), designers will be able to choose the level of abstraction that would both maximise the outcome of the game and keep the development cost in manageable levels.

Points 4, 5, and 6 are each the natural next step of the previous point. In point 4 the particular game components are chosen; in point 5, subject matter experts (SMEs) cooperate with game designers to provide the appropriate content; in point 6, game designers choose the most effective game techniques given the problem at hand. A potential direction for improvement for point 6 could be to build a lexicon of game techniques (game mechanics) and use case studies to understand under which circumstances certain techniques should be preferred than others (Sweeney and Meadows, 2010).

Points 7, 8, and 9 concern the validation of the game and are discussed in the following section.

2.2.2. VALIDATION

Similar to game design, game validation has to combine methods from the design and the analytical sciences. A validation study should balance the attention between the *Simulation* and the *Game Layers*. First and foremost, validation should not be performed just once after the implementation of the game; it should, instead, be carried out in an iterative manner not just during the implementation (Point 7) but throughout the whole lifecycle of the game (Balci, 1998). While Duke (1980) defines game evaluation (point 8) as the client's in-house validation once the final product has been delivered to them, this study refers to game evaluation as the complete validation and verification (V&V) study of a game (Balci, 2004). Similarly, the authors consider point 9, i.e. field testing, to be also part of the V&V study. Potential directions for improvements in this area could be to:

- Incorporate, given the numerous V&V methods (Balci, 1998), the appropriate ones for the *Simulation Layer* (Roungas et al., 2017), and whenever possible automate them (Roungas et al., 2018g).
- Develop methodologies that would allow the formalisation of the validation of the *Game Layer*, similar to what it was proposed in Section 2.2.1 for the game design.

It should be noted that while the game's sample size, i.e. the number of participants, and debriefing heavily influence the validity of a game, they are part of the game session, thus they are analysed separately in the following section.

2.2.3. GAME SESSIONS

Even the most realistic games are still an abstracted version of reality and as such the knowledge from a game session should effectively be transferred to, and generalised for, reality. Hence, the sample size and debriefing are of utmost importance for the analysis of the game outcome. Potential directions for improvements with regards to the sample size could be to:

- Verify, before characterising the sample size as small, that it is indeed small (Lenth, 2001) and that it actually compromises validity.
- Incorporate, if the sample is small, the appropriate V&V methods and statistical techniques (Roungas et al., 2018f).

- For games in which participants do not need to physically be present, use technology to reach a greater audience (Katsaliaki and Mustafee, 2012).

In this respect, potential directions for improvements with regards to debriefing could be to:

- Develop methodologies for formalising debriefing, which in turn can enable the facilitators to choose the most beneficial methods for debriefing given the game and the real system at hand.

- Build a standard for best practices in debriefing (Roungas et al., 2018a).

2.2.4. Knowledge Management

A knowledge management system (KMS) can been used for one or more purposes, like own-project improvement, cross-project improvement, organisational culture improvement, and network improvement. Currently, there are two KM strategies widely used in organisations: i. Personalization and ii. Codification (Markus, 2001). The former is more suitable for those professionals who inquire on experts' opinion but do not want to acquire their knowledge, and thus most likely prefer to consult an expert in a one-to-one conversation. The latter is more suitable for those professionals who want to learn from past projects and apply this knowledge in the future, and thus more likely prefer a documented and detailed record of these past projects. Potential directions for improvements in this area could be to:

- Establish strict and precise protocols for root cause analysis, in case problems or accidents occur throughout the lifecycle of a game or due to decisions made based on a game (Latino et al., 2016).

- Develop methodologies for acquiring, indexing, and reusing knowledge in and around games that can potentially benefit an organisation in the future (Roungas et al., 2018b).

- Build a repository of game artefacts, from game-related materials (e.g. game engines, boards, cards etc.) to any analysis pertaining to a game (e.g. reports, data etc.) (Roungas et al., 2018d).

- Build a network that would allow professionals within an organisation to easily identify the most appropriate experts based on the problem at hand (Ahuja et al., 2012).

2.3. Research Goals and Questions

The directions for improvement, presented in the previous section, are the first step towards defining the research scope of this thesis. The approach of first presenting the bigger image and then narrowing this image down to an attainable set of potential improvements was adopted, in order to again show the richness and complexity of the gaming field. In this section, the research goals and subsequently the research questions of this thesis are divided in two levels. On the first level, the thesis intends to either propose new methods or promote a new way of thinking on how to approach problems in each of the four main areas of games. On the second and higher level, the thesis intertwines these four areas with four research themes, in order to identify how these areas can advance the gaming field. Finally, the thesis aims to draw conclusions which would be agnostic with regards to the domain of the main case study used hereinafter, i.e. transportation/logistics.

2.3.1. FIRST LEVEL

On the first level, the thesis aims at contributing independently on the four main areas of games. In more detail:

In Part I, game theory is used to enable the formalisation of game design as well as the structured abstraction of the system under study. This part aims to answer two research questions:

1. What aspects from game theoretical analysis can be translated to game design, and in what way?
2. To what extent can the design of a meaningful game be determined from game theoretical analysis?

In Part II, the validation of pure simulations and the implications in games is explored. This part aims to answer four research questions:

1. How can the selection of validation and verification methods and applicable statistical techniques given the simulation and the real-world system at hand be optimised as to be more time efficient and rigorous?
2. How can the operational validation of a simulation model be automated or semi-automated, in order to reduce the time, cost, and human error associated with it?
3. Which factors play a critical role in the success of a simulation validation study?
4. Which are the implications with regards to the validity of games and which lessons can we learn from pure simulations?

In Part III, a methodology for extracting the tacit knowledge of experts with regards to how game sessions can be improved and the results from applying this methodology are presented. This part aims to answer one research question:

1. How can game sessions in general and debriefing in particular be performed and analysed in a rigorous scientific way?

In Part IV, an ontological framework for the knowledge management of games is proposed. This part aims to answer one research question:

1. How can knowledge from games be acquired and managed in order to reduce the risks and costs associated with games?

Upon answering these research questions, this thesis intends to also provide an understanding of the particularities of each area and how these areas intertwine and influence each other. Therefore, before moving to a higher level, in the conclusions, this thesis synthesises the individual conclusions from all four areas and answers one research question:

1. How do the four areas intertwine and influence one another?

2.3.2. SECOND LEVEL

Apart from the individual conclusions in the four areas of games, this thesis also draws conclusions in a higher level. These higher level contributions are based on three research themes in juxtaposition with the four areas of games. Therefore, in this second level, the thesis attempts to answer three research questions:

1. How do the analytical and design sciences facilitate, or even inhibit, improvements in each of the four areas?
2. How does complexity, as described in Section 1.1, present itself in the four areas and thus how can it be managed?
3. How is the quality of decisions made through games influenced by the four areas?

2.4. THE NEXT STEP

In this chapter, the research gaps and potential directions for improving them were identified, leading up to defining the research questions of this thesis. In the next chapter, Chapter 3, the research agenda is set and the methods with which this thesis aims to answer the research questions are briefly presented.

3

Research Approach

One thing I know, that I know nothing.
This is the source of my wisdom.

Socrates

Games require a multidisciplinary approach and more importantly a multidimensional methodological approach. Games cover areas from the pure analytical sciences (e.g. mathematics, data analytics etc.), all the way to the other side of the spectrum in the design and social sciences, and in between. Particularly, studies that do not limit themselves to one specific aspect of games are deemed to combine methods and techniques from various fields as well as from various schools of thought. This chapter, which is the last introductory chapter, first clearly defines the research agenda and the methods used throughout this thesis, and then describes the variations in the epistemological approach.

3.1. Research Scope

In this section, based on the potential directions for improvements explored in Section 2.2, a research agenda is proposed and the methods for answering the research questions are briefly explored. The research agenda is based on the four main areas of games, as those were first identified in Section 2.1. Figure 3.1 shows the different parts of the research and a brief demonstration of the different methodologies.

3.1.1. Design

Game design is characterised by a severe limitation in formal methods. Various frameworks and methods for formalising game design have been proposed with each one of

Parts of this chapter have been published in Roungas et al. (2018h).

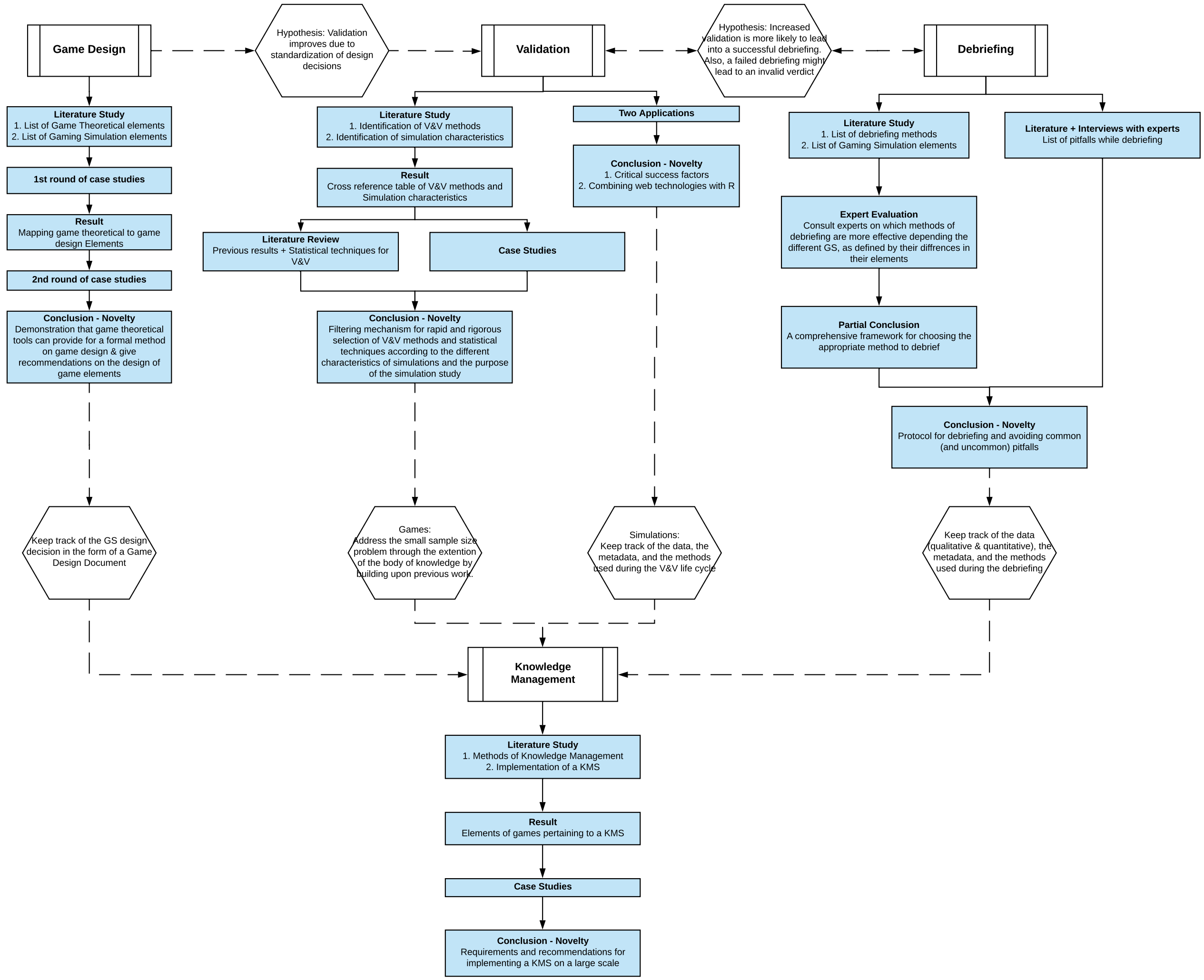


Figure 3.1: Illustration on the games and simulations explored in this thesis.

those having their own advantages and disadvantages. In this thesis, and particularly in Part I, the contributions and limitations of existing literature in game design are described and a novel framework based on game theory (GT) is proposed. While GT has been scarcely used in game design, it is a well-established tool for understanding and modelling the relationship of different actors in an analytical way. There have been supportive studies towards using GT in game design (Ritterfeld et al., 2009, Salen and Zimmerman, 2004). GT is a tool that can be coupled with existing methodologies and help bridge the gap between the design and analytical communities. Moreover, GT can systematically analyse real-world systems and pinpoint those areas within these systems that are problematic and could benefit from the use of games (Roungas et al., 2019b). Finally, GT can be directly used in the game design by formalising the game design choices and applying the optimal strategies wherever is required (Vorobeychik and Wellman, 2009).

The proposed framework consists of two phases: The CHARACTERISATION and the LINKS. The CHARACTERISATION is concerned with abstracting and modelling the real system using GT and it has been accomplished through an extensive literature review. The LINKS is concerned with connecting the elements of the model, i.e. the product of the CHARACTERISATION phase, with game design elements, thus creating a roadmap towards a formalised game design. This phase was construed first also through an extensive literature review and then by using two cases, i.e. games, from ProRail, in order to validate the connections. Finally, the complete framework has been validated through three case studies, two from ProRail and one from the Swedish healthcare system.

3.1.2. VALIDATION

Unlike pure simulations, games have a distinct characteristic, which is the human participation, or in other words games have what in Section 2.2.2 was defined to be the *Game Layer* on top of the *Simulation Layer*. Game validation, due to its nature of including humans, usually depends more on the subjective opinion of experts (van Lankveld et al., 2017), e.g. questionnaires, than formal methods. This limitation is related to the lack of design methods for games as well as to the usually low number of participants. The former is analysed in Part I. The latter, i.e. the sample size, plays a significant role on the applicability of game results. A small sample size is easy to obtain but has limited possibilities for analytical conclusions thus limited possibilities for generalising the observations from the game. A large sample size, while it solves the analytical problem and the generalisability of the results, it is usually expensive to obtain and difficult to coordinate (Grogan and Meijer, 2017).

Validation of the *Simulation Layer* has been vastly researched through the course of the last three decades (Balci, 1998, 2004, Sargent, 1996), where numerous formal methods and statistical techniques have been introduced. Moreover, methodologies for verifying that indeed a sample size is small (Lenth, 2001) have been proposed. However, this abundance of research has become the actual inhibitor due to the large number of methods and techniques. In Part II and particularly in Chapter 7 and Chapter 8, a framework for selecting the most appropriate simulation validation methods and statistical techniques, on the basis of the different characteristics of simulations, is proposed. In turn, in Chapter 9, a web tool that automates validation is explored. In the fourth chapter of Part II, Chapter 10, lessons learned from the application of the web tool in two cases from the railways are pinpointed.

Validation of the *Game Layer* due to its nature of including uncertainties pertaining to human activity, usually is not as straightforward as the *Simulation Layer*. On the one hand, the formalisation of game design can provide structure on game validation, as analysed in Part I. On the other hand, with regards to the sample size, the *Game Layer* would be benefited only through gradually extending the body of knowledge by building upon previous work. This aspect is directly linked with knowledge management, analysed in Part IV, in the sense that the greater the number of game sessions conducted, the more the evidence of a system's behaviour acquired and, as such, the cumulative sample size gradually becomes large enough to generalise the outcome of the game. These implications with regards to game validity are discussed in more detail in Chapter 6.

The games used in engineering organisations, like the primary case study in this thesis, tend to have a very dominant *Simulation Layer*, as is equally the case with all the games used in this thesis. As a result, Part II is heavily focused, though not solely, on the validation of seemingly pure simulations. Nonetheless, the implications of the validation of the simulation model in game validation become evident in Part II.

3.1.3. Game Session

A game session consists of three phases: briefing, gameplay, and debriefing, with the latter being considered the most important feature of games (Crookall, 2010). Nevertheless, their almost completely synthetic nature raises the question: are game sessions in general, and debriefing in particular, performed and analysed in a rigorous scientific way? In other words, are they consistently structured, given the different characteristics of games, and is it also clear what would constitute a successful game session and debriefing? The answer to all these questions is no. The reason for this negative outcome is that expertise regarding game sessions and debriefing resides almost entirely in the tacit knowledge spectrum. As a result, knowledge and best practices on how to conduct fruitful game sessions and debriefing are either disseminated without understanding the reason(s) for which a particular decision would be beneficial or not disseminated at all. Hence, the aim of Part III is to shed light on this tacit knowledge possessed by experts, to gain understanding on why certain practices are more prone to success than others and, eventually, to bring to light other practices that have remained largely hidden. In order to accomplish this goal, two rounds of interviews with facilitation experts and other game stakeholders were conducted.

3.1.4. Knowledge Management

Knowledge management (KM) and reuse of games is not, and should not be, of academic interest only. The effectiveness of a corporation depends heavily on how it manages and reuses knowledge (Markus, 2001), or in layman terms, how in the first place it obtains and thereafter maintains the so-called "Know-how" (Roungas et al., 2018d). As a corporation acquires and builds up on knowledge obtained through games, it improves its know-how, and thus sustains or even increases its competitive advantage (Dixon, 2000).

From a practical point of view, KM should be translated in a system, i.e. a knowledge management system (Alavi and Leidner, 2001). In Part IV, a methodology for defining the backbone structure of such a system is proposed. The proposed methodology is first constructed based on the state of the art and then tested using three cases from ProRail.

3.2. EPISTEMOLOGY

The broad scope of this study, as briefly illustrated in the previous section, requires a departure from the traditional monomethod research and dictates, instead, the adaptation of a mixed method approach (Johnson and Onwuegbuzie, 2004). This is also evident in the need for both communities of scientists involved in games, i.e. analytical and design, to bridge their differences.

From a purely epistemological point of view, this thesis adopts constructivism, although some areas in the study follow slightly different perspectives. More specifically, while research on game design follows a clear constructivism approach, the other three areas do not. Validation, being more mathematically oriented, adopts postpositivism; research in game sessions adopts interpretivism; and knowledge management should best be seen from a pragmatism point of view.

3.3. CONCLUSION OF THE INTRODUCTION

In this introduction spanning in three chapters, the building blocks of the thesis were set. In Chapter 1, the terms *systems* and *gaming simulations* that are extensively used throughout the thesis were defined and the primary case study, ProRail, was presented. In Chapter 2, research gaps and potential directions for improving them were identified, and the chapter concluded with the specification of the research questions. Finally, in Chapter 3, the research agenda of the thesis was set forth. The next four parts explore this research agenda within the areas of design, validation, game sessions, and knowledge management.

I

Design

We must free ourselves of the hope that the sea will ever rest.
We must learn to sail in high winds.

Aristotle Onassis

4

Design Framework

This chapter along with Chapter 5 constitute the first step towards formalising games in general and their design in particular. In brief, in this chapter, a framework for formalising, and consequently standardising, expediting, and simplifying the modelling of games is proposed. The proposed framework applies game concepts pertaining to game theory in the abstraction of the real system and the game design decisions.

4.1. Introduction

Game theory (GT) and games are two terms which, despite their lexical resemblance, are used to describe two seemingly unrelated fields. GT is the study of mathematical models of conflict and cooperation between intelligent rational actors, which results in the definition of Game Concepts (GC) (Bekius and Meijer, 2018, de Bruijn and ten Heuvelhof, 2008, Myerson, 1997, Rasmusen, 2007).

Games are imitations of real-world systems (RS) designed to solve a problem; their primary purpose is not merely to entertain but to educate, train, steer decision-making processes, etc. (Michael and Chen, 2005, Zyda, 2005). In this respect, building a game out of an RS necessitates the use of modelling to reduce the real-world complexity to a manageable level. The process of modelling and building a game out of an RS is characterised by the following challenges: i. it can be time consuming, which translates both into delay and cost, ii. it usually requires extensive experience on the part of the designers, as well as concrete knowledge of the system under study, and iii. depending on the actual size of the system, it dictates multiple decision-making, thus increasing the probability that mistakes will be made in the course of the modelling process, especially when the system includes hidden personal agendas and a notion of politics.

Yet since both GT and games aim to describe and interpret the behaviour of actors participating in complex systems (Holland, 1992), there does seem to be an area in which these two converge. The initial hypothesis in this study is that there are more correlations

Parts of this chapter have been published in Roungas et al. (2019b).

between these concepts than discrepancies, as may be inferred from both their definitions and tautological resemblance. This hypothesis is deemed to be verified by the fact that GT models are used (consciously or unconsciously) by game designers when designing games (Salen and Zimmerman, 2004).

The framework proposed in this chapter intends to improve the modelling of games in two ways. First, GC are used to analyse and abstract the RS, thus pinpointing the problematic areas within this system and its worst-case scenarios. Second, a link is made between GC elements (actors, strategies, issues etc.) and decisions regarding the game components (scenarios, goals etc.). By making such connections, many less relevant game decisions can be filtered out, thus accelerating the modelling and prototyping of games and making the design decisions more rigorous. As a result, the methodology presented can be applied by less experienced game designers. This study therefore aims to address the following research questions:

> Research Question 1: What aspects from game theoretical analysis can be translated to game design and in what way?

> Research Question 2: To what extent can the design of a meaningful game be determined from game theoretical analysis?

The term *Meaningful* refers to a game that addresses the problem at hand, and thus fulfils its purpose.

In Section 4.2, a literature review reveals the interconnections between GT and games. In Section 4.3, a framework is proposed for modelling games through GT. In Section 4.4, the development and validation process for that framework is described.

4.2. BACKGROUND WORK

This section presents literature on game design as a stand-alone practice and on its relationship with GT, then further reviews the application of GT to games.

4.2.1. GAME DESIGN

From the early days of games, there have been attempts to define and formalise game design. Duke (1974) proposed the use of conceptual maps combined with precise documentation of the design process. Such maps have the ability to ensure the games' correspondence with reality, ascertain that the appropriate level of abstraction is being adopted, and to confirm that the corresponding proposals can be implemented in the game design. The framework proposed uses GC in the same way. While a claim cannot be made on whether GC are more effective than conceptual maps, they are selected in this case because they are part of a larger framework.

Harteveld (2011) discusses balancing reality, meaning, and play in game design. For each of these three pillars, he proposes several ways to implement them successfully within a game.

> **Reality**. Incorporating reality into a game is of the utmost importance. Harteveld (2011) proposes achieving this not only through familiarisation with the RS under study, but by also enabling discussion with the client and subject matter experts.

This in turn will allow accurate identification of the actors and objects involved and enable the building of relationships between them.

Meaning. In the light of the RS it aims to imitate, the game has to have a specific purpose; it should transmit a particular message to the intended audience. The game designer should therefore define its purpose and develop a strategy on how to accomplish this, which in turn requires the implementation of certain game mechanics and feedback mechanisms, as well as reflection through debriefing.

Play. Finally, the development of an engaging, immersive, and aesthetically pleasing game can facilitate its positive reception by participants, and thus improve its outcome.

Harteveld (2011) implicitly utilises GT in several ways, but not fully. The goal of this study is to build upon his work, and more specifically to develop the first pillar of its triadic game design approach, i.e. reality, by explicitly linking game theoretical concepts with game design elements.

4.2.2. Game Theory in Games

With regard to GT approaches in game design, Meijer (2012a) comments on the differences between GT and games but goes on to conclude that these concepts are often intertwined. He defines game theory as “the mathematical approach of analysing calculated circumstances where a person’s success is based upon the choices of others”. In games, where the success of one player often depends on the choices made by others, GT hence provides a popular method for modelling artificial intelligence.

Bolton (2002) makes a case for the significance of GT in designing role-playing games, especially if these are to be successful at practical forecasting. Moreover, he observes that work to date on GT and role-playing games has dealt with highly simplified versions of the real world. In the same spirit, Ritterfeld et al. (2009) asserts that the systematic review process that GT provides could be a valuable heuristic for game designers.

Aligned with Bolton (2002), Salen and Zimmerman (2004) provide two reasons why GT can be useful for game designers. Firstly, GT analyses situations which resemble simple games in a detailed way. Secondly, it focuses on relationships between decisions and outcomes. In effect, the authors think of actions and outcomes as the building blocks of meaningful play. They therefore believe that applying GT concepts is useful when designing such games. Moreover, they take a step towards proposing concrete ways in which GT can be used to actually help game designers. By looking at games as a series of strategic decisions, they suggest the use of several GT elements, such as:

- decision trees, which allow the linking of different parts of a storyline, i.e., the order of the decisions made by the players;
- utilisation functions, which assist in quantifying players’ preferences;
- strategies, which can guide the players as they play; and,
- pay-off matrices, which show the relevant outcomes of a game, depending on the players’ decisions.

This approach by Salen and Zimmerman (2004) is a promising step towards formal use of GT as a game design tool, but it has a few restrictions.

1. The game has to be turn-based, or, in general, in discrete steps.

2. Players have to make a finite number of clear decisions with knowable outcomes.

3. The game has to be finite; it cannot go on forever.

These restrictions can be quite inhibiting in games with a decision-making purpose, where players do not take turns, the outcome of each decision is barely knowable, and the set of choices from which a decision can be chosen is infinite. In the proposed framework, decisions are not necessarily be linked to specific outcomes, thus freeing the designers to choose whichever they want to include.

Several case studies on the application of GT to games design have also been conducted. On of the most popular is the Beer Game (Sterman, 1989), which has formed the basis for further studies on optimisation (Meng et al., 2010, Thompson and Badizadegan, 2015) and the modelling of artificial intelligence (Kimbrough et al., 2002). Other less popular games, which have nevertheless incorporated GT in their game design are the approach by Mader et al. (2012) to developing a therapeutic game, where GT is used to examine the relationship between therapeutic activities and the players' motivation, and the attempt by Skardi et al. (2013) to apply cooperative GT to the control of total sediment yield in the watershed, vis-a'-vis landowners conflicting interests.

Further to the above, Guardiola and Natkin (2005) use GT as a tool to model and understand local properties of gameplay by building game matrices for a video game. In effect, though, they are using GT to understand a game rather than to design it. Finally, Fullerton (2004) proposes the utilisation of GT with various examples but only when the game involves dilemmas.

The literature reviewed in this second part illustrates that GT can contribute significantly to designing games. Nevertheless, research on this topic not only remains limited but also has severe limitations (restrictions, simplified games, etc.) or focuses only on specific games in the form of case studies. Throughout this study, it will be explicitly pinpointed wherever the proposed framework contributes in existing work and how it helps overcome limitations in previous research.

A point of criticism on the use of GT is that the method cannot cover the richness of empirical decision-making processes (Bennett, 1987, Binmore, 1987). It simplifies the situation to rational players who can only choose actions from a limited set of prescribed alternatives. When GT is applied directly to game design/science, the result is a game which scope is too narrow (Klabbers, 2018). Forcing the entire process into one game concept results in an oversimplification of the situation that is not useful for the decision-maker or game designer when applying it to real-world cases (Bekius et al., 2018). In order to mitigate the possible simplification multiple GC are used to characterise the process. The approach presented in this study is different from more general game theory applications since the concepts used are able to cover rich policy situations and give nuance to different incentives of different actors (Bekius et al., 2018).

As a conclusion to the literature review, one may identify the absence of a framework for formalising scientifically the application of GT on the whole spectrum of games. In the next section, the framework introduced aims at tackling this issue.

4.3. FRAMEWORK OF GAME DESIGN

This section proposes a framework for modelling RS through GT and for linking GC and games. The hypothesis tested in this respect is that the development of a game out of an RS - which undoubtedly translates into several design decisions and multiple individual game components - by default requires abstraction of the RS. As such, there is a need for a modelling framework able to guide the designer towards a game, which is an accurate representation of the system it simulates and is also feasible to build and maintain.

The proposed framework consists of: i) a methodology for abstracting the RS and describing it through one or more GC; and ii) a list of GC elements and, linked to it, the corresponding list of game design decisions. Establishment of the links is attempted through the use of the characteristics of the GC (actors, strategies, issues, etc.) and the different game design decisions (scenarios, goals, etc.).

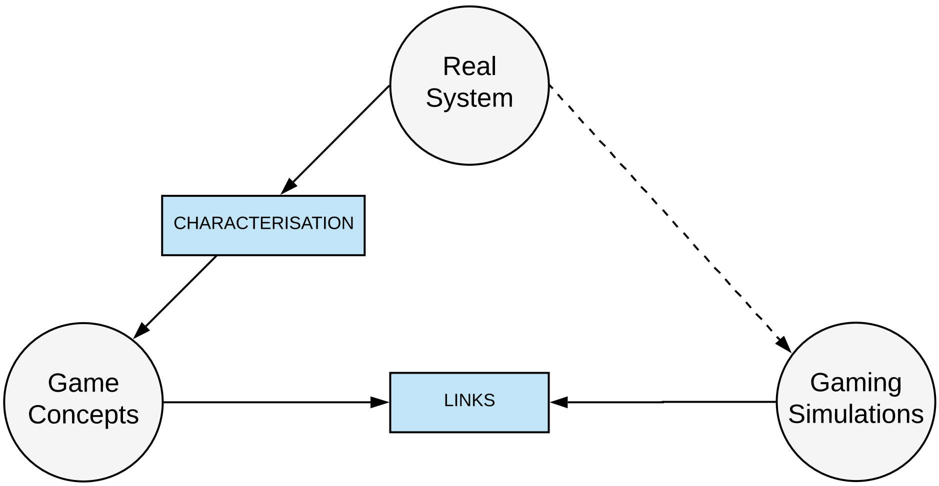


Figure 4.1: Framework for characterising the Real System and linking the Game Concepts to Game Design decisions.

The framework is depicted in Figure 4.1 and contains five blocks:

- The Real System (RS) represents the system under study. The RS contains actors operating in and on the system, as well as dynamics created by the interaction between the system and the actors. Depending on its complexity, the system can be characterised as either a complex adaptive system or a socio-technical system.

- The Game Concepts (GC) contains characteristics from the toolbox called Game Theory (Osborne and Rubinstein, 1994) representing the game elements of the RS under study. GC describe the interaction between and behaviour of actors who have to make a decision (Bekius and Meijer, 2018). Some game concepts are mathematically defined, such as the well-known Prisoners Dilemma (Rasmusen, 2007), while others have only been observed empirically, for example the Multi-Issue game (de Bruijn and ten Heuvelhof, 2008). Therefore, the characteristics of the GC vary between being empirically substantiated and mathematically proven.

- The Gaming Simulations (games) represents the game design decisions used in modelling the RS, after taking into account the complexity of the system the game is being designed for.

- The CHARACTERISATION of RS into GC is the first step in the methodological process. The resulting GC should enable identification of the problematic areas and worst-case scenarios within the system, thus answering the second research question.

- The LINKS between GC and games is the second step in the methodological process and subsequently answers the first research question. This is the part that is more directly connected with Harteveld (2011) and with his triadic game design, since it is the one that eventually leads to game design recommendations.

The dashed arrow represents the game design literature as of to date, thus making even more explicit the contributions of this study. The direct link from the RS to the game shows that game design is usually based on the experience of game designers and rarely based on formal methods.

The following section elaborates on how the proposed framework was developed and validated. Particular attention is given on the two capitalised blocks of the framework, the CHARACTERISATION and the LINKS.

4.4. METHODOLOGY

This section describes the methodology for developing and validating the proposed framework. That is in two parts, each corresponding with one of the rectangles in Figure 4.1:

- the CHARACTERISATION of RS into GC.

- the LINKS between GC and games.

Three organisations are involved in the case studies, which are more thoroughly discussed in Chapter 5; two are from the Netherlands and one is from Sweden. In the Netherlands, the organisations are ProRail and NS; ProRail is the government agency responsible for maintaining the national railway network infrastructure, allocating rail capacity and traffic control, whereas NS (Nederlandse Spoorwegen), also known as Dutch

Railways, is the principal passenger train operator. In Sweden, the organisation is the Stockholm County Council (Stockholms Läns Landsting, SLL), which is a regional government responsible for all healthcare provision in greater Stockholm.

In Section 4.4.1 and Section 4.4.2, an analysis of the two parts of the framework is provided. While in Figure 2, the complete methodology is depicted in a graph. The white background indicates artefacts observed either in the real world or in literature; the grey background indicates games or game-related projects used throughout the methodology; the cyan background indicates artefacts related to the framework; the involved organisations in each case are shown in parenthesis.

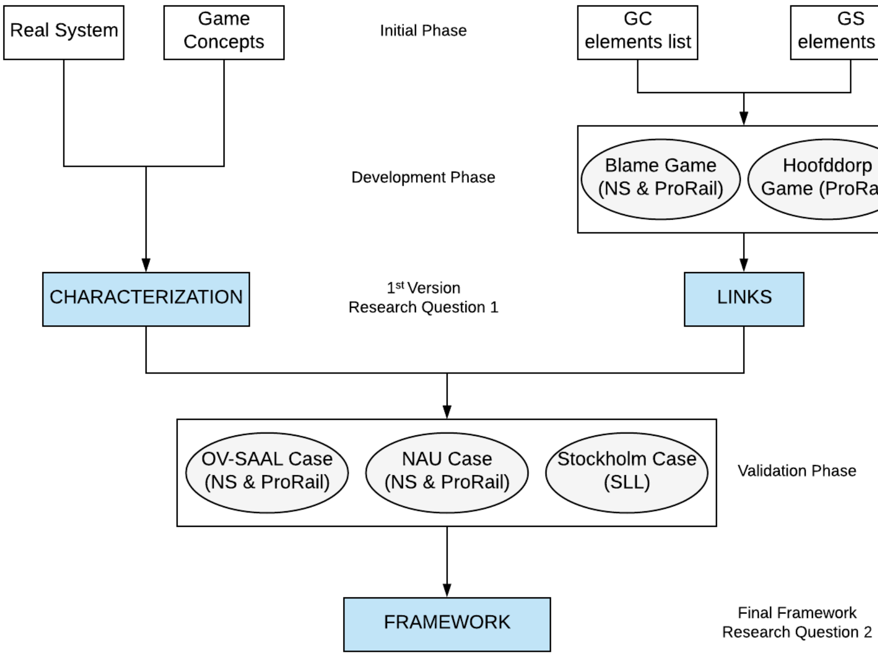


Figure 4.2: Methodology for the development and validation of the proposed framework.

4.4.1. THE CHARACTERISATION

GT models describe interactions between actors, who make decisions in order to reach a certain outcome. They can formalise the mechanisms and patterns actors perform in RS and thus be used to characterise these systems (Goeree and Holt, 1999, Helbing, 1994, Helbing and Balietti, 2011, Moss, 2001, Vollmer, 2013). Since several examples of such GT characterisations exist, choosing the right mechanism for the situation at hand is crucial, yet not always evident (Barreteau et al., 2007, Feld, 1997).

Bekius and Meijer (2018) present a taxonomy of GT concepts (GC). These originate from both formal GT and public administration, in order for these concepts to have a richer and more descriptive definition. The characteristics of GC therefore vary between being empirically substantiated and mathematically proven.

The criteria used to design the taxonomy, which originate from theory on complex real-world decision-making processes (de Bruijn and ten Heuvelhof, 2008, Koppenjan and Klijn, 2004, Teisman and Klijn, 2008), are important for selecting the right GC. Multiple actors are usually involved in these processes, forming a network of interdependencies., And hierarchical relations can exist within those networks, most frequently between two actors.

The aim of the process is to reach a collective decision. However, individual strategic behaviour plays an important role as well. Moreover, the decision-making process is dynamic. Therefore, the set of GC chosen from the taxonomy should, and does, cover a wide range of situations appearing in RS. A more detailed explanation of GC selection can be found in Bekius and Meijer (2018), while a comparison of this approach with other characterisation methods or decision-making tools can be found in Bekius et al. (2018).

With regard to games, the GT notions help us to analyse the situation and to predict worst-case scenarios. Since we obviously want to avoid such scenarios if at all possible, the ability to identify them in advance can be particularly helpful when making game design decisions.

4.4.2. THE LINKS

Two lists of GC characteristics and game design decisions are compiled in order to identify the elements linking GT and games. The compositions of these lists are based on literature. From a theoretical point of view, these two lists begin from a different start point, i.e. GT and game design, with the aim to be linked using two games. The two games are analysed in order to formulate an initial assumption regarding the links. For each of these games, the content of each element included in the corresponding GC and game design lists is identified. On this basis, elements from the two lists are then linked.

For the GC characteristics, a list of 16 GC elements (de Bruijn and ten Heuvelhof, 2008, Osborne and Rubinstein, 1994, Rasmusen, 2007) is used as a starting point. Overlaps between some of these elements necessitated the conduct of new research, which eventually introduced new elements. The resulting list of GC elements incorporated these, as well as merged versions of some of the original overlapping elements. The refined list is shown in the first column of Table 1.

For the game design decisions, additional literature is used in order to adapt and enhance the list of game elements for educational games, compiled by Roungas and Dalpiaz (2016), so as to fit in games with any purpose. The corresponding list of game design elements is shown in the second column of Table 1.

Two problems arise by creating these two lists.

1. The lists do not contain completely independent elements. 2. 1-to-1 correspondence between GC characteristics and game design decisions is not always applicable.

Problem 1 can be addressed by merging elements which appear within the same list. If one element is dependent or subordinate to another, it follows that the two can be merged. With regard to problem 2, 1-to-n, n-to-1 or no linking may be used as well.

Table 4.1: Links between GC characteristics and game design decisions by analysing the Hoofddorp Game and the Blame Game.

GC	Games	Comments
Actors	Characters	Both contain the same people, since clients and facilitator(s) are included to the Characters.
Actions	Challenges, Rules, Tasks	Challenges, Rules, and Tasks are not completely independent and they all have overlaps with the action set from game theory. For the Blame Game, Rules are not part of Actions because the game is quite open with very few rules. Also, the few rules that exist do not correspond to the actions taken by the participants.
Strategies	Challenges, Motivation	Strategies are about the how (Challenges) but also about the why (Motivation).
Payoffs	Motivation, Rewards	Pay-offs are the sum of explicit (e.g. money) and implicit (e.g. satisfaction) rewards.
Information set	Feedback	Information set is influenced by many game elements but we only connect elements from the 2 lists when they match content-wise. Feedback is particularly included.
Context	Scenario, Fidelity, Type of game	Definitions and content between context and scenario fit almost completely. Context can define the level of fidelity and the type of game. For the Blame Game, Fidelity level is not linked because it is not a design decision.
Issues	Challenges, Pitfalls	Only for those issues that correspond to the content of the game and not issues related with the design. Define a new element "Pitfalls" for issues related with the game design.
Outcome	Goals, Debriefing, Purpose	With goals due to similar definition. With debriefing due to the fact that debriefing aims at maximizing the outcome of the game. Define a new element "Purpose" showing the purpose the game is designed for.
Iterative game	Repetition	Due to similar definition

References: Alessi (1988), Apperley (2006), Barreteau et al. (2007), Bekius et al. (2016), de Bruijn and ten Heuvelhof (2008), Rasmusen (2007), Roungas and Dalpiaz (2016)

THE GAMES

To verify the proposed methodology, two games were used. Both are related to the railway sector. They have been called the Hoofddorp Game and the Blame Game. Interviews were conducted with the designer of each, asking specific questions in order to gain an insight into their design decisions and to retrieve the requisite information needed to identify the GC characteristics. Subsequent to each interview and establishment of its results, the substance of each element of the GC and game lists was ascertained independently. In other words, one researcher identified the GC characteristics and another the game design decisions. In this way, the probability of bias in creating the links was minimised. As described above, moreover, elements were merged when one was dependent on or subordinate to another. Such merges were effected only for elements appearing on the same list.

The Hoofddorp Game is a board game with a low fidelity level, which tests changes affecting the railway infrastructure in and around Hoofddorp station. Hoofddorp itself is a small town between Amsterdam and Leiden, but in the Dutch national railway network it is strategically situated close to the country's largest airport, Schiphol, on the main line linking it to some of the Netherlands' biggest cities, like The Hague and Rotterdam. Any changes affecting the infrastructure at Hoofddorp can thus have a severe impact on the connection between these cities and the airport.

The game has two different scenarios:

1. What happens if a fire breaks out in the railway tunnel under Schiphol Airport?

2. What happens in the event of disruption on the line to Leiden?

The output of the game has been used as input to help decide whether changes to the infrastructure at Hoofddorp are necessary or, alternatively, a whole new plan should be compiled.

The Blame Game is a role-playing game with a high fidelity level, which simulates a situation where two groups are "blaming" each other for incorrect planning. Subsequently, each player writes a report in which they nominate one or more members of the opposing group for dismissal.

The game has two scenarios:

1. In 2016 a decision was made to modify part of the railway infrastructure. But now, in 2018, the resulting performance has proven disappointing and passengers face frequent delays.

2. In 2016 it was decided that it would not be beneficial to make any changes to the infrastructure. But now, in 2018, performance is disappointing and passengers face frequent delays.

This game has been used to raise awareness amongst strategic decision-makers of the interdependencies within the system and of the importance of team play.

The Final Links

Actors (GC) have a 1-1 link with Characters (Games), because the latter are a subset of the former in the sense that they include the participants in the game, the client and the facilitator(s), all of whom are included as Actors (GC) along with the game designer(s).

The Action Set (GC) has a 1-n link with Rules, Challenges and Tasks (Games). The Action Set is defined by what participants need to accomplish in the game (Tasks), which in turn heavily influences the way they pursue their objectives (Challenges) based on the applicable restrictions (Rules). The only exception to the above are games like the Blame Game, which can be described as open games with minimal or no predefined rules. In these cases, the Action Set has a 1-n link with Challenges and Tasks (Games) only.

Strategies (GC) have a 1-n link with Challenges and Motivation (Games), because Strategies (GC) are about how a specific action from the Action Set (GC) is chosen (Challenges) and why (Motivation).

Pay-Offs (GC) have a 1-n link with Motivation and Rewards (Games) because, being the utility an actor receives, Pay-Offs (GC) can also generally be described as the sum of implicit (e.g. satisfaction) and explicit (e.g. money) rewards.

The Information Set (GC) has a 1-1 link with Feedback (Games) because the latter produces information about past (reaction to an action) and future (knowledge that can be used in the future) actions.

Context (GC) has a 1-n link with Scenario, Fidelity Level and Type of Game (Games). The definitions and content of Context (GC) and Scenario (Games) are almost identical, since both refer to the general situation surrounding the game. In addition, Context (GC) can determine Fidelity Level (Games) - low, medium or high - as well as the Type of Game (Games).

Issues (GC) have a 1-1 link with Challenges (Games), but only in the case of those issues which relate to the content of the game - not for the issues which relate to its design. Although Challenges (Games) seem to be the best match to Issues (GC), given the current list of game design elements it would be more appropriate to introduce a new game element including information pertaining to issues related to the game design. We have therefore introduced the term Pitfalls (Games), which is defined as any problems or mistakes occurring during the process of designing a game.

Outcome (GC) has a 1-n link with Goals and Debriefing (Games), due to its similar definition to Goals (Games) and the fact that Debriefing (Games) aims at optimising the outcome of the game. Although Goals and Debriefing (Games) appear to be almost a full match with Outcomes (GC), there seems to be a gap regarding the purpose this outcome is used for; in other words, a game design element describing the purpose of the game is missing. Therefore, Purpose (Games) has been introduced as a new game design element defined as the function the game is designed for, e.g. training, decision-making, etc.

Iterative Game (GC) has a 1-1 link with Repetition (Games), because of their very similar definitions. Both terms show whether any, and if so how many, repetitions are needed in order to optimise the outcome of the game.

4.5. Conclusion on the Methodology

The methodology proposed is different from that of Salen and Zimmerman (2004), in the sense that it does not directly imply how the order of actions should be or which actions should belong to which outcomes. Instead, it specifies on a more high level the sets of actions or sets of outcomes that belong to certain design choices. Finally, when someone knows which game is "played" and is able to gain insight on the different GT elements, the table can be used as a structure to make your game design choices.

In the next chapter, the proposed framework and the subsequent methodology are validated through two case studies from the Dutch railway sector. A third case study, i.e. the Stockholm case, is used as a way to show (inexperienced) game designers how to use the proposed framework for future game design.

5

Application of Framework

In this chapter, the application of the framework, proposed in Chapter 4, in three case studies reveals several advantages of incorporating game theory into game design, such as formally defining the game design elements and identifying the worst-case scenarios in the real-systems, to name but two.

5.1. OV-SAAL Case

The so-called OV-SAAL corridor (Schiphol-Amsterdam- Almere-Lelystad public transport) is part of the High-Frequency Rail Programme (Programma Hoogfrequent Spoor), which aims to increase the number of trains operating per hour in the Randstad conurbation. In order to raise capacity in this corridor, several options have been proposed by the various actors involved, two of which are ProRail and NS. One of these options is doubling the tracks around Weesp station.

5.1.1. The Game

The game was paper-based, using real timetables. The participants, playing the role of traffic controllers, were people from ProRail with at least with some experience in the role. The purpose of the game was to test the robustness of five pre-designed infrastructure enhancements in the face of medium-scale disruptions. The game resulted in an expanded set of solutions, which inhibited the participants from reaching a consensus, hence no final decision was made. As a result, the game was negatively received and criticised since it had failed to fulfil its purpose.

5.1.2. Game Concepts

In this case, four GC based on a game-concept-selection tool (Bekius and Meijer, 2018) were identified (this same tool was used in all the case studies). Those GC were the Multi-

Parts of this chapter have been published in Roungas et al. (2019b).

Issue Game, the Cascade Game, the Volunteer's Dilemma and the Battle of the Sexes. The predominant one was the Multi-Issue Game, so that is further analysed below.

MULTI-ISSUE GAME

A Multi-Issue Game is when multiple actors with different incentives form a network of interdependencies and finally reach consensus in a decision-making process that was initially deadlocked. A large number of actors results in multiple issues coming to the table, which intensify and increase as the moment when a decision needs to be made approaches. Actors have broad agendas, which usually create room for consensus, negotiations, cooperation and participation in the process.

However, too many issues and actors with different ideas about them can result in over-complex situations. And, because of all these issues and actors, the process can be delayed. That was the case with OV-SAAL. During the design and the gameplay phase, the number of issues involved - and hence the number of scenarios (and runs) to be played out - increased. As a result, the complexity of the game did not reflect that of the actual situation. Moreover, major issues which were not supposed to be solved at that particular moment were also introduced, further increasing the complexity and the frustration among participants.

5.1.3. GAME DESIGN RECOMMENDATIONS

In the OV-SAAL case, the game could have been significantly improved using the proposed methodology. First, characterisation of the problem as Multi-Issue could have been completed within a relatively short time, which in turn would have provided insights into the game design. Then, given the links shown in Table 4.1 in Chapter 4, several game design elements could have been better defined.

The game should have included not only participants from the operational layer of the organisation but also from management, thus engaging the actual decision-makers with the process. Alternatively, had that not been possible, the challenges and tasks within the game should have been simpler. That is, they should have involved fewer decisions in order to avoid over-complex situations with multiple issues per actor. This could easily have been achieved if Actions and Strategies for the Multi-Issue Game had been explicitly defined. Finally, defining a set of Outcomes for the Multi-Issue Game would have maintained the focus of the game on its initial purpose, thus providing an additional safeguard that the game would deliver valid and meaningful results and so be considered successful by the stakeholders.

5.1.4. CONCLUSION OF THE CASE

From the above analysis, it is clear how GC can help develop more robust and meaningful games. There are several areas of game design which GC can improve, but the most noticeable is the never-ending struggle of every game designer to create a realistic game while maintaining complexity at a reasonable level. The OV-SAAL case therefore provides a positive step towards validation of the proposed framework.

5.2. NAU CASE

Utrecht Centraal is the most centrally located and busiest railway station in the Netherlands. It is within an hour by train from Amsterdam, The Hague, Rotterdam, Nijmegen, and Eindhoven. Consequently, disruptions there can affect almost every other major station in the country. The primary purpose of the NAU case was to address such disruptions and to make Utrecht more resilient (Van den Hoogen and Meijer, 2012). Its secondary purpose was to alleviate the workload of the rail traffic controllers.

The complexity of the situation, which necessitated the use of games, lay in the fact that the operational layer of organisations like ProRail tends to resist implementing decisions made by the managerial layer (strategic decisions), thus increasing the uncertainty of their effectiveness. In this particular case, an additional reason for characterising the situation as over-complex was the conflicting incentives of the actors involved. ProRail was focused on improving system performance, whereas NS was most keen to reduce the workload of the controllers.

5.2.1. THE GAME

The game used a paper-based model of the infrastructure, with low-tech interfaces but real timetables. The participants, playing the role of traffic controllers, were from different entities, including ProRail and NS (Van den Hoogen and Meijer, 2012). The purpose of the game was fourfold.

1. To test a pre-designed separation of traffic-control tasks into de-clustered zones of control (Van den Hoogen and Meijer, 2012).
2. To test a different traffic-control concept intended to mitigate second-order delays (Lo et al., 2013).
3. To limit abnormalities during major disruptions.
4. To adjust the division of labour at the traffic control centre (Meijer, 2012a).

5.2.2. GAME CONCEPTS

In this case, two GC were identified: the Hub-Spoke and the Battle of the Sexes. Of these, the Hub-Spoke was predominant and so is further analysed below.

HUB-SPOKE

Multiple actors (the spokes) with different incentives are steered by one actor (the hub) using a command-and-control style. The game creates an incentive to make inflated claims, as the spokes can make agreements amongst themselves and create strategic issues for the hub.

In the NAU case, the strategic level at ProRail is the hub and the different operational departments (including NS and other actors) are the spokes. The former wants to see its decisions implemented, while the latter need to be convinced of the usefulness and necessity of those decisions, which influence their way of working - a highly culturally sensitive factor. If the spokes are unwilling to implement the decisions and able to cooperate with each other, they make life difficult for the hub.

5.2.3. Game Design Recommendations

In the NAU case, the recommendations defined using the proposed methodology are mostly in line with those resulting from the game itself. This managed to actively involve both the strategic and the operational levels and to make the latter aware of its necessity and usefulness in creating a more resilient system. The two most important recommendations are as follows.

1. Reduce the number of decisions to be made and thus limit the design space of the game. This recommendation is similar to that resulting from the OV-SAAL case, but differs in the fact that, for NAU, the game designers explicitly limited the number of decisions (Van den Hoogen and Meijer, 2012).

2. Acknowledge the potential conflicts in the incentives driving the different actors. This increases the external validity of the game. Such incentive-based conflicts can occur both between organisations (in this case, between ProRail and NS) within them (in this case, between controllers and managers). Involving all relevant stakeholders in a game raises awareness of those conflicts, which in turn provides a more realistic overview of the situation.

5.2.4. Conclusion of the Case

The NAU case resulted in a proof-of-concept which was later considered largely successful. In retrospect, then, the fact that the recommendations provided by this study were mostly in line with the actual implementation of the game is yet another positive step towards validation of the proposed framework.

5.3. Stockholm Case

In Sweden, Stockholm County Council (SLL) is a regional government responsible for all healthcare provision in greater Stockholm. However, home-care services (non-medical decision-making) are provided by local authorities, not SLL. Psychological and social care provision is split between local government and SLL. This makes the institutional environment rigid and not so easy to change.

The demand for healthcare in Stockholm is enormous and rising. The current system faces difficulties in meeting this demand, which makes it vulnerable. A large proportion of the demand comes from older people, who have multiple health issues. One possible solution is the use of sensors to control their well-being at home, thus potentially reducing unnecessary visits to healthcare facilities.

SLL wants to introduce digital innovations in healthcare by conducting tests at local teaching hospitals. Specifically, it wishes to start with three testbeds of 100 elderly people each. They will be supplied with sensors, which will be monitored.

5.3.1. Game Concepts

In this case, three GC were identified: the Volunteer's Dilemma, the Principal-Agent Game and Hub-Spoke. The predominant one was the Principal-Agent Game, so that is further analysed below.

PRINCIPAL-AGENT GAME

A Principal-Agent Game describes a hierarchical relationship between a principal and an agent, in which the former is dependent upon the latter because of their expertise in a certain subject. This GC reveals the power position of the subordinate in such a relationship.

In this case, SLL is the principal and the teaching hospitals are the agents. Similarly, a local authority could be seen as the principal and home-care centres as the agents. This immediately reveals the complexity of the decision-making processes, since multiple Principal-Agent Games can take place simultaneously. This analysis focuses on the interaction between SLL and the hospitals.

The knowledge and expertise concerning the digitisation of healthcare in general and the introduction of the testbeds, in particular is possessed by the hospitals. SLL is therefore dependent upon them unless it acquires more power.

During the game, the agent makes a decision regarding the test and the principal either accepts or rejects it. The agent's decision is modelled using the variable $y \in 0, 1$. This is defined as follows:

- y=0 means that the agent is fully objective and not at all influenced by the hierarchical power and expectations of the principal.
- y=1 means that the agent is fully subjective and makes the decision expected by the principal.

Given these two extremes, eight possible outcomes exist. Based on their probability of occurrence, the worst-case scenario can be identified. An overview of Principal-Agent Game is shown in Figure 5.1.

From SLL's perspective, the worst-case scenario is when it wants the test performed and would thus prefer a "Yes" decision but the hospitals are fully objective and decide "No" (scenario 3 in Figure 5.1). SLL needs the hospitals to co-operate with it, otherwise it cannot solve the region's healthcare problems. When the worst-case scenario occurs, that damages the relationship between principal and agent, which is not beneficial for either of them.

5.3.2. GAME DESIGN RECOMMENDATIONS

Unlike the OV-SAAL and NAU cases, this project is still ongoing and a game has not yet been designed. Therefore, any recommendation provided would not be for research and validation purposes only but could also serve as an actual input for the forthcoming game.

Based on the analysis from the Principal-Agent Game, the worst-case scenario for SLL is when doctors decide not to go ahead with the tests. Of course, if such a decision is based purely on their medical or scientific assessment, then SLL should probably accept it. But if it is based on a lack of knowledge of new technologies and how they work, this is something a game can prevent. Hence, one design recommendation would be to develop a game for doctors focusing especially on the worst-case scenario. In other words, design a game which raises the doctors' awareness of modern sensors, how they work and how they can simplify their everyday job-related activities.

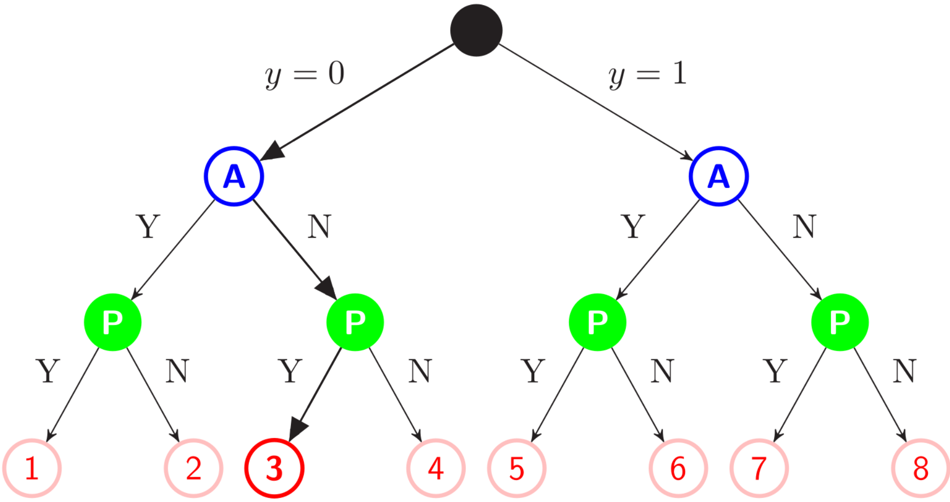


Figure 5.1: The Principal-Agent game in the Stockholm case.

5.3.3. CONCLUSION OF THE CASE

The application of GC in this case shows yet another way in which GT can benefit the development of games. GC pinpointed the worst-case scenario in a quick and formal way, whereafter a game can be used to further explore and perhaps prevent it.

5.4. CONCLUSION

This study, spanning in Chapter 4 and Chapter 5, proposes a framework which contemplates a more efficient and effective modelling of games by formalising their design decisions using GT concepts. Based on the reviewed literature and to the best of our knowledge, such a framework has never been proposed before. At present, however, it relies heavily on case studies in order to be fine-tuned and validated.

5.4.1. ANSWER TO RESEARCH QUESTIONS

With regard to Research Question 1 (What aspects from game theoretical analysis can be translated to game design, and in what way?), the answer lies within a continuum. At one end of this there is the direct translation of GC elements to game design choices (e.g., Actors → Characters, Pay-offs → Rewards, Outcome → Purpose), at the other the purely qualitative information (e.g., Actions, Strategies) which should be entrusted to the game designers, since their interpretation depends heavily not only on the purpose of the game but also on the particular requests made by the client (i.e., the person or company which owns and assigns development of the game).

With regard to Research Question 2 (To what extent can the design of a meaningful game be determined from game theoretical analysis?), the answer lies in the advantages of the proposed framework.

- The links between GC elements and game design choices, as defined in Research

Question 1.

- Identification of the purpose of the game (the WHAT) by including the context of the decision to be made in the analysis of the situation.

- Identification of worst-case scenarios and problematic areas, as particularly shown in Section 5.3.

- Prediction of the possible outcomes of the game. Even when the game does not explicitly steer participants towards a certain outcome and designers want to keep this broad, during the debriefing this could be a way to structure the discussion (e.g., What if you had chosen A instead of B?).

- Prediction of how a situation characterised by a specific GC can evolve in the future into another GC.

Given these advantages in general, and the game design recommendations in particular, it is safe to conclude that, for the cases presented in this chapter, a meaningful game can be designed based on the proposed framework.

In addition to answering the two research questions, whether and the extent to which the proposed framework resolves the challenges associated with modelling games, as those were identified in Section 4.1 in Chapter 4, should also be addressed. While the introduction of the intermediate step of GT between the RS and games is not trivial, it does enhance the information for designers that can subsequently be part of the game design. For example:

In the OV-SAAL case, as it was shown, it was difficult even for experienced designers, let alone for inexperienced, to acknowledge how the multiple needs and wants of each actor would significantly increase the complexity of the game and as a result inhibit the final decision making process. In the Stockholm case, using a simple tree-like graph (Figure 5.1), the worst-case scenario was pinpointed relatively quickly. Most probably an experienced game designer would have found the same result but it would have been difficult to do it equally rapidly. Moreover, in this particular case, inexperienced designers would have had a hard time understanding the complexity of the Swedish healthcare system, abstract it and then identify the worst-case scenario. The LINKS part of the framework provides a roadmap for translating parts of the RS, through GT, to game design choices. While it can also be helpful for experienced designers as a reference, it is particularly useful for inexperienced designers because it gives them a “dictionary” on how a real world problem can be abstracted and translated into a game.

Finally, it should be noted that the proposed approach overcomes the restrictions imposed by the game theoretical approach of Salen and Zimmerman (2004), i.e. discrete steps, knowable outcomes and finite gameplay.

5.4.2. LIMITATIONS

Naturally, the novelty of the proposed framework entails some risks. Moreover, GT as a discipline has also its own limitations. Therefore, acknowledging and either eliminating or mitigating these inhibitors is of paramount importance.

The two lists were constructed based on a literature review in the fields of GT and game design. The use of literature almost entirely eliminates the risk of incorporating incorrect elements in either list, but only mitigates the risk of neglecting to include further relevant elements. With regard to the game design decisions list, the risk of not including important elements is further mitigated by the fact that this list is based mostly upon interviews with game design experts, who were called to comment as to whether an element was missing from the list.

The games used in this analysis have different fidelity levels, serve different purposes and, most importantly, address different professionals. Nonetheless, the first two games (OV-SAAL and NAU), which are the ones used to validate the framework, share one common characteristic: both relate to the railway sector. This represents a risk in respect of the validity of the framework.

While GT offers a vast toolbox for exploring social systems, it also comes with certain restrictions. The most important of these is that GT assumes that actors behave rationally, whereas more often than not social systems tend to behave in a seemingly irrational way. In order to mitigate that risk, the proposed framework does not force designers to choose a rational path for their game design; that is left open. Another significant limitation of GT is that it has only a restricted ability to reveal the reasoning behind certain choices made by actors. It is only during the actual gameplay of the game that their rationale may be revealed.

5.4.3. Future Work

The risk of inadvertently omitting certain GT elements can be mitigated by interviewing GT experts who have experience with complex real-world decision-making and are thus able to pinpoint whether any element has not been incorporated in the framework. Furthermore, additional case studies in fields other than the railways, as well as with games that have different characteristics (in terms of fidelity, purpose, intended audience and perceived success or failure) from the ones used in this study, will add further value to the proposed framework. Finally, game designers should test the framework in a real-world design situation in order for its validity to be further strengthened.

II

Validation

Since we cannot change reality,
let us change the eyes which see reality.

Nikos Kazantzakis

6

VALIDATION IN GAMES

THIS first chapter of part II aims to debate on the nature of validation in games and particularly on the extent to which validation methods from simulations can be successfully applied in games and on the relationship of game validity to credibility and usability. Before diving into the specifics, it is essential to first define what is validation in games. Adapting Schlesinger et al.'s (1979) and Balci's (2003) definitions from simulations, validation in games can be defined as the degree to which the game imitates the underline system in a satisfactorily level, or in layman terms game validation addresses the question of whether the game is the "right" one. Yet, Peters et al. (1998) argues that the scope of this definition is restrictive and it does not account for more abstract, perhaps even metaphorical, games. They instead adopt a more broad definition, initial proposed by Raser (1969), who identified four criteria for the validation of games: psychological reality, structural validity, process validity, and predictive validity. In this thesis, while the importance of a strict definition, which would subsequently clearly provide an acceptability threshold, is acknowledged and supported, the need for a broader understanding of what the validation of games entails is also considered. The aim of this chapter is therefore not to propose one particular methodology for game validation but rather to pinpoint that in most cases game validation is not as straightforward as the validation of simulations.

With regards to validity in more broad terms, games can also be seen themselves as a mechanism for validation. Regardless the technology used or the area of application, games have been used as a means to validate certain hypotheses or future scenarios (Meijer, 2015). The concepts of validation and games are therefore intertwined, forming a more complex relationship than initially anticipated. In this thesis and particularly in this part, while several of the games examined have been used for validation purposes, the research is primarily concerned with the validation of games as opposed to how games can be used to validate artefacts or hypotheses.

6.1. FROM SIMULATIONS TO GAMES

Unlike pure simulations, games have a distinct characteristic, which is the human participation, or in other words games have a *Game Layer* on top of the *Simulation Layer*. Game validation, due to its nature of including humans, usually depends more on the subjective opinion of experts (van Lankveld et al., 2017), e.g. questionnaires, than formal methods. This limitation is related to the lack of design methods for games as well as to the usually low number of participants. The former was analysed in Part I. The latter, i.e. the sample size, plays a significant role on the applicability of game results. A small sample size is easy to obtain but has limited possibilities for analytical conclusions thus limited possibilities for generalising the observations from the game. A large sample size, while it solves the analytical problem and the generalisability of the results, it is usually expensive to obtain and difficult to coordinate.

As it will be shown in Chapter 7, validation of the *Simulation Layer* has been vastly researched through the course of the last three decades (Balci, 1998, 2004, Sargent, 1996), where numerous formal methods and statistical techniques have been introduced. Moreover, methodologies for first verifying that indeed the sample size is small (Lenth, 2001), then selecting the most appropriate validation methods and statistical techniques among the numerous existing ones, shown in Chapter 8, and finally automating validation, shown in Chapter 9, have been proposed. Furthermore, for games in which participants do not need to physically be present, technology can be used to reach a greater audience, thus increasing the sample size (Katsaliaki and Mustafee, 2012). Though, the real contribution in games that the knowledge acquired from the validation of simulations has to offer is the ability of the validation methods from the simulation field to tackle and potentially address three of the criteria proposed by Raser (1969), i.e. structural validity, process validity, and predictive validity.

Validation of the *Game Layer* due to its nature of including uncertainties pertaining to human activity, usually is not so straightforward. On the one hand, the formalisation of game design can provide more structure on game validation, as analysed in Part I. On the other hand, with regards to the sample size, the *Game Layer* would be benefited only through gradually extending the body of knowledge by building upon previous work. This aspect is directly linked with knowledge management, analysed in Part IV, in the sense that the more game sessions are conducted the more evidences of a system's behaviour are discovered and the cumulative sample size gradually becomes large enough to generalise the outcome of the game. Furthermore, the *Game Layer* dictates the need for validating a game also with regards to its psychological reality, which is the degree to which a game provides an environment that seems realistic to the players.

6.2. VALIDITY VS. CREDIBILITY AND USABILITY

Apart from validity, two more terms are often associated with a game's successful implementation and application, credibility and usability. Credibility is defined as whether, and the degree to which, key stakeholders in a project consider the game, and subsequently its results, to be "correct", always vis à vis the particular objectives of the study (Law, 2008). While credibility does not conceptually has a 100% correspondence with prediction, this thesis posits that within the scope of games they are strongly correlated, in the sense that if a game has high predictive validity would provide credible results, and vice versa, a game that provides credible results has high predictive validity.

The second term, usability, is defined as the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use (ISO/IEC-9241-11, 2018). Given the nature of the games this thesis is covering, usability is strongly correlated with psychological reality, in the sense that games that aim to imitate a particular system, even in a metaphorical level, tend to be considered realistic when they provide a familiar to the users interface to interact with, which in turn this familiarity is perceived as a more usable artefact.

Given those two terms along with the different kinds of validity, proposed by Raser (1969), validation of games should be seen as an extended version of simulation validation, incorporating psychological and user experience factors, as opposed to be seen as mutually exclusive to usability (van den Hoogen, 2019). In that sense, the claim that a game is valid cannot be stated unless it has been established that the game is usable and its results credible, always for its intended purpose of use. Indeed, credibility has been recognised by several researchers to be a validity criterion (Eisenhart and Howe, 1992, Leininger, 1994, Lincoln and Guba, 1985). On the other hand, while to the best of our knowledge usability has not been explicitly identified to be a criterion for validity, there have been studies that acknowledge a strong connection between them in general (Greenberg and Buxton, 2008) and in games as well (Fernandez et al., 2012, Kortmann and Sehic, 2010).

6.3. Overview of Part II

While game validation can be strongly benefited from analytical methods, it cannot solely rely on them since it heavily depends on contextual and behavioural factors. These factors are not in conflict or even separate from validation, as the latter is defined in simulations, but rather complimentary, all of which in this thesis are put under the umbrella of validation in games. Moreover, game validation does not only depend on the game itself but also on how the game is executed. In other words, it depends on the briefing, game session, and debriefing, with particularly the latter being of paramount importance. Validation and game sessions have a reciprocal relationship. Increased validation is more likely to lead to a fruitful and more successful game session, and a successful game session boosts the game outcome and thus further increases its validity. But then the question that rises is: How is a successful game session defined particularly in games for decision making? A question that is further explored in Part III.

Due to the nature of the games explored in this thesis, i.e. games for decision making in engineering systems, which incorporate a very dominant simulation layer, the emphasis on validation is given on the simulation part of games. The subsequent chapters in this thesis are therefore primarily focused on simulations. Particularly, Chapter 7 and Chapter 8 are the combined work on a framework for selecting the most appropriate validation and verification methods and statistical techniques given the simulation and the real-system at hand. In Chapter 9, a methodology for automating the validation of simulations is proposed. Finally, in Chapter 10, the lessons learned from validating two simulation models in the Dutch railways are presented.

7

Simulation Validation

This chapter along with Chapter 8 are the combined work on a framework for selecting the most appropriate validation and verification methods and statistical techniques given the simulation and the real-system at hand. In this first chapter, an extensive literature review is conducted as to build comprehensive lists of validation and verification methods, and statistical techniques. Then, in Chapter 8, the actual framework is proposed. The influence of this work in the gaming field becomes apparent in Chapter 8, where one of the criteria for selecting validation and verification methods, and statistical techniques is whether they would be applicable in games.

7.1. Introduction

Back in 1972, based on Forrester's work (Forrester, 1971), Meadows et al. (1972, 1974) introduced World 3, a simulation of the world for the years 1900-2100. The purpose of the simulation model was to project the dynamic behaviour of population, capital, food, non-renewable resources, and pollution. The model's forecast was that during the contemplated two centuries the world will experience a major industrial collapse, which will be followed by a significant decrease in human population. The model became very popular especially because of the increasing interest in environmental degradation encountered because of human activities (Janssen and De Vries, 1999). Even though the model gained support for being "of some use to decision makers" (Meadows et al., 1974) and generated the spark for many later global models, it had several shortcomings, for which it received a lot of criticism (Nordhaus, 1973). In turn, this criticism raised the question of whether, and to what extent, such simulation models are validated and verified. This is just one example of the notion that validation and verification (V&V) is a fundamental part of a simulation study (Balci, 1994).

The term V&V is used to characterise two relatively different approaches that almost always go hand by hand, namely validation and verification. Validation is this phase of

Parts of this chapter have been published in Roungas et al. (2017) and Roungas et al. (2018f).

a study that ensures that the simulation imitates the underline system, to a greater or lesser extent, and in any case satisfactorily (Schlesinger et al., 1979), or in layman terms validation addresses the question of whether the built model is the also the "right" one (Balci, 2003). On the other hand, verification is the phase of the study that ensures that the model and its implementation are correct (Sargent, 2005), or in layman terms verification addresses the question of whether the model was built in the "right" way (Balci, 2003). V&V has become a well-researched field with a significant amount of produced literature and commercial case studies. The large number of V&V methods and statistical techniques created or adopted by this wide range of research, is the greatest impediment to designing a V&V study.

The predetermined budget of a simulation study usually limits the amount of time and resources that can be spent on V&V. Additionally, the nature and the diverse characteristics of simulations limit the number of V&V methods and statistical techniques that are applicable to each simulation. In other words, not all V&V methods (hereinafter referred to as methods) and V&V-applicable statistical techniques (hereinafter referred to as techniques) are suitable for every simulation. To the best of our knowledge, a taxonomy for characterising methods and techniques and, subsequently, matching them with different simulations does not exist. Therefore, the research question that this study will address is:

> How can the selection of V&V methods and V&V-applicable statistical techniques given the simulation and the real-world system at hand be optimised as to be more time efficient and rigorous?

This chapter aims at identifying the majority of the available methods and techniques in order, in Chapter 8, to classify them on the basis of their different characteristics and on whether they can be used to validate or verify a simulation, and eventually match them with characteristics of simulation models.

7.2. LITERATURE ANALYSIS

In this section, a 3-step literature analysis is presented. The initial hypothesis of this study is that simulations exhibit certain properties that influence the effectiveness and applicability of methods and techniques. Therefore, the 3 steps of the literature analysis are the following:

Step 1: Identification of methods and techniques.

Step 2: Identification of simulations' properties potentially influencing the selection of methods and identification of simulations' and systems' characteristics potentially influencing the selection of techniques.

Step 3: Identification of the phases of a simulation study.

7.2.1. STEP 1: V&V METHODS AND STATISTICAL TECHNIQUES

Methods are different in many aspects; some methods are strictly mathematical whereas others accommodate the more qualitative aspects of simulations, etc. Balci (1998) identified more than 70 methods, which in turn categorised into four categories: informal, static, dynamic, and formal. Balci's (1998) list is the most accurate representation of the body of work on methods and, even to date, is considered as the most extensive one. This study adopts the list in reference - but not the categorisation - and goes as far as to propose a new classification of methods.

On the other hand, numerous techniques have been proposed throughout the years, a subset of which are applicable in V&V studies. Moreover, techniques can be characterised in various ways, e.g., depending on the input they require (numerical, categorical etc.), or the purpose they are used for (goodness-of-fit, time series etc.).

Below, an introduction on some of the V&V methods and statistical techniques the author used is provided, while the complete lists of methods and techniques, along with the definition of each one of them, can be found in Appendix A and Appendix B respectively.

V&V METHODS

One of the most popular methods for validation is *Face Validation*, in which people knowledgeable about the system under study subjectively compare the model's and system's behaviour and judge whether the model and its results are reasonable (Hermann, 1967). *Face Validation* is often used along with *Graphical Comparison*, in which case graphs produced from the model are compared to graphs produced by the system under study, in order to detect similarities and differences between the two (Miller et al., 1993). A variation of these two tests, which in some extent copes with bias, is *Turing Test*, in which experts are presented with the model's and system's output data and without knowing which one is which, they are asked to differentiate the two (Schruben, 1980). While all these methods are informal, they have been quite effective in simulation, as well as game, validation, hence their popularity.

Another very popular method for the operational validation of a model is *Predictive Validation*, in which the model executes with past input data and the results are then compared with data from the real system (Emshoff and Sisson, 1970). This method works well under the assumption that exogenous, and thus unpredictable, factors will have equal, or even less, effect in the model in the future than they did in the past.

With regards to the conceptual model validation, one of the most common validation methods is *Inspections*, in which a team of four to six members conduct a study on all the development phases of the model such as the system and objectives definition, the conceptual model design, and the communicative model design. The end result of the study is not only the validation but also suggestions for improvements and a follow-up (Schach, 2011). Similar to *Inspections* are *Reviews* and *Walkthroughs*, but in the former the review team also includes managers, while in the latter only the model developer from the development team should be part of the team conducting the study.

STATISTICAL TECHNIQUES

When it comes to using statistics in simulation validation, *t-Test* is by far the most popular and perhaps the easiest to use technique for comparing two different datasets. But a positive result in a *t-Test* does not automatically imply that these two datasets - in the case of validation these datasets would be the model's and the real system's - behave in the same way; for exploring similarities in the behaviour of two datasets, the *Goodness*

7

of Fit techniques are used, with one of the most commonly used being the *Kolmogorov-Smirnov Test (K-S Test)*. *K-S Test* can be one-sample, i.e., test whether a sample is distributed according to a known theoretical distribution (e.g., normal, binomial etc.), or two-sample, i.e., test whether two different samples are drawn from the same empirical distribution (Chakravarty et al., 1967). In simulation model V&V, the two-sample K-S test is the most common, i.e., comparing whether the data from the model and from reality are derived from the same distribution.

7.2.2. STEP 2: SIMULATIONS' AND SYSTEMS' PROPERTIES AND CHARACTERISTICS

This step aims at identifying the properties and characteristics of simulations and the real-world system (hereinafter referred to as system) under study that can potentially influence the selection of methods and techniques.

SIMULATIONS' PROPERTIES

Since simulations differ from one another in various ways, distinctions are made on whether they represent an existing system, or whether they simulate a system at a microscopic or macroscopic level, or whether they are intended for learning or decision making, and so forth. This is an indication that simulations can be characterised by various properties. Based on literature, this study has identified 10 properties of simulations. The rationale behind selecting those properties was to describe simulations with as much detail as possible. Hence, the properties span on multiple levels. Not all identified properties necessarily influence the selection of V&V methods, therefore, this step is not only about identifying the properties themselves but also determining which are the ones that really influence the effectiveness of a method; in other words, this step serves as the rationale for choosing those properties of simulations that are applicable to specific V&V methods, and provides for the reasons behind this selection.

The 10 identified properties of simulations are the following:

1. Access to the source code of the simulation. Accessibility, or lack of it, influences the selection of a V&V method (van Gunsteren and Mark, 1998), since several methods require some sort of a check on the code level. Hence, this property is included in the analysis.

2. The simulation represents an existing real-system for which real data exist (Kleijnen, 1995). The existence of, or more importantly the lack of, real data heavily influences the selection process since several methods require real data and thus cannot be used when no real data is available. Hence, this property is also included in the analysis.

3. The formalism the simulation is based on, like Discrete Event System Specification (DEVS), Differential Equation Specified System (DESS), System Dynamics, etc. (Vangheluwe et al., 2002). Several frameworks and methods have been proposed on how to verify and validate DEVS (Byun et al., 2009, Saadawi and Wainer, 2009), DESS (Di Benedetto et al., 2007, Jo et al., 2012), or system dynamics models (Barlas, 1994, Forrester and Senge, 1980), but they are either application specific or the

same method can be used in more than one formalisms, making it independent of the actual formalism. Therefore, while formalisms are an important aspect of simulation modelling, their influence on the V&V method selection is minimal, ergo excluded from the analysis.

4. The simulation's worldviews: i) Process Interaction/Locality of Object, ii) Event Scheduling/Locality of Time, iii) Activity Scanning/Locality of State (Overstreet and Nance, 2004). While worldviews allow for more concise model descriptions by allowing a model specifier to take advantage of contextual information, there is not any evidence from a literature point of view that they have an influence on the V&V method selection, hence, they are excluded from the analysis.

5. The fidelity level of the simulation (Low, Medium, High) (Liu et al., 2008). While from a literature point of view there is no evidence to support the influence of the level of fidelity on the V&V method selection, common sense dictates that there must be some. Indeed, in order to characterise a simulation as of high fidelity, it must imitate an existing system and real-world data must exist, thus making the comparison and the final characterisation possible. Therefore, as discussed in the second property and shown in Table 8.1, the existence of data of the real system influences the V&V method selection, as does the level of fidelity. Yet, since the correlation between real data and high fidelity is almost 1-to-1, the fidelity level is excluded from the analysis for reasons of simplification.

6. The type of the simulation (Constructive, Virtual, Live) (Morrison and Meliza, 1999). This classification, which is adopted by the U.S. Department of Defense (DoD-5000.59-M, 1997), should be seen more as a continuum rather than as a discrete characterisation. Once a simulation moves towards the Virtual or the Live side of the continuum, it can also be referred to as 'game'. A game has the distinct characteristic that the game session is succeeded by debriefing, whereby the participants reflect upon the game session to link the content presented during the session with reality (Fanning and Gaba, 2007). It has been demonstrated that debriefing can in general facilitate validation (Lo et al., 2013, van den Hoogen et al., 2014). Moreover, while all methods identified in this study are suitable for pure simulations (constructive), not all of them are appropriate for games. It would be interesting to examine which of the methods can also be used for validating games. Hence, this property is included in the analysis.

7. The purpose the simulation was built for (learning, decision making, etc.). Several case studies on V&V of simulations for different purposes have been reported; in training (Morgan et al., 2004, Zevin et al., 2012), in decision making (Gass, 1983), in concept testing (Nemani and Running, 1989), etc., but there are no reports of specific V&V methods being more effective for a certain purpose. Hence, this property is excluded from the analysis.

8. The simulation imitates a strictly technical, a socio technical system (STS), or a complex adaptive system (CAS) with multiple agents. There are several studies on modelling and validating simulations for STS (Mavin and Maiden, 2003) and CAS

with multiple agents (Louie and Carley, 2008, Nilsson and Darley, 2006) but there are no indications that certain V&V methods are more effective for an STS or a CAS. Therefore, this property is excluded from the analysis.

9. The application domain of the simulation (logistics, business, physics, etc.). Although the application domain of the simulation plays a significant role in the modelling process, since different approaches are required (Newtonian physics for object movement, Navier-Stokes equations for fluid behavior, etc.) for modelling different systems (Landriscina, 2013), literature, or more precisely the lack of it, suggests that the V&V process and thus the V&V method selection is not affected by the application domain. Hence, this property is excluded from the analysis.

10. The functional (hard goals) and non-functional (soft goals) requirements of the simulation (Mylopoulos et al., 1999). Validating the simulation's requirements is indeed an important part of the V&V process (Balci, 2004), since validation is always relative to the intended use (Pace, 2004), in other words the use defined in the requirements. Hence, making a distinction between the hard and soft goals is paramount and as such this property is included in the analysis.

Simulations' and systems' characteristics

Simulations and the systems they imitate can produce a variety of data, which can be characterised in various ways. Moreover, depending on the type of data and on the purpose of the V&V study, different statistical tests are usually necessary, which in turn depend on the produced output. Based on the literature review on the techniques presented in Section 7.2.1, the characteristics of simulations and systems that influence the selection of techniques are the following:

1. Number of datasets. The most usual case in simulation model validation is to have two datasets (model and reality). Nevertheless, there are cases where the number of datasets can be either one, e.g., when testing whether the model derives from a known distribution like the normal or gamma distributions, or more than two, e.g., when testing the results of more than one models against the operational data.

2. Number of variables. The most usual case in simulation model validation is to test one variable, e.g. in railway simulations, this variable is usually the amount of delay. Nevertheless, there are cases where the number of testing variables is more than one, e.g. simultaneously testing longitude and latitude values between model and reality.

3. Purpose of the statistical technique. A statistical technique can test for equality of means, the extent to which the data from the model and reality are similarly distributed, the extent to which two time series are equivalent, or it can be used to reduce the model's complexity.

4. Known parameters. Statistical techniques are divided in two major categories: parametric and non-parametric. Parametric techniques are the ones that require the mean and variance (μ, σ^2) to be known, whereas non-parametric techniques can deal with cases where these parameters are not known.

5. Type of data. The type of data simulations and systems produce range from strictly quantitative to purely qualitative. Usually, statistical techniques suitable for a V&V study should be able to deal with data that are either numerical or categorical (binary).

6. Size of samples. Simulation and system data are almost impossible to be normally distributed. Nevertheless, due to the Central Limit Theorem (Feller, 2015), when the size of a sample exceeds 30 (or 40 depending how close to be normally distributed the data are), it is assumed that it follows the normal distribution thus the techniques that work for the normal distribution are applicable.

7.2.3. STEP 3: PHASES OF A SIMULATION STUDY

According to Sargent (2000), there are 4 distinct phases of V&V: *Data Validation*, *Conceptual Model Validation*, *Model Verification*, and *Operational Validation*. *Data Validation* is concerned with the accuracy of the raw data, as well as the accuracy of any transformation performed on this data. *Conceptual Model Validation* determines whether the theories and assumptions underlying the conceptual model are correct, and whether the model's structure, logic, and mathematical and causal relationships are "reasonable" for the intended purpose of the model. *Model Verification* ensures that the implementation of the conceptual model is correct. Finally, *Operational Validation* is concerned with determining that the model behaves accurately based on its intended purpose. This study adopts Sargent's (2000) characterisation and aims at using it to classify the methods, in addition to the simulations' properties.

7.2.4. CONCLUSION OF THE LITERATURE REVIEW

It is evident that selecting one method or technique over another for a V&V study depends on several characteristics of the simulation, the system, the methods, and the techniques, as well as the phase of the simulation study. In Chapter 8, a methodology that combines all three steps aiming at the development of a framework for method and technique selection is proposed.

8

Framework for Validation Method Selection

In this chapter, a methodology for selecting the most appropriate V&V methods (Section 8.1) and statistical techniques (Section 8.2) for a V&V study is proposed and a case study illustrates how the framework, through the use of Table 8.1 and Table 8.2, can be used.

8.1. V&V method selection methodology

As discussed in Section 7.2.2, dimensions 3, 4, 5, 7, 8, and 9 are perceived to have little influence on the method selection, hence, there are excluded from the analysis. On the other hand, the purpose of the method selection, discussed in Section 7.2.3, seems to be crucial; in other words, it is important to differentiate on whether the selected method will be used for data validation, conceptual model validation, model verification, or operational validation. Therefore, the list of the dimensions is refined, and is expressed in questions, as follows:

1. Does the V&V method require access to the simulation model's source code? *Possible answers: Yes or No.* A positive answer to this question means that this method can only be used when the person or persons performing the V&V have access to the simulation's source code, whereas a negative answer means that it can be used in any occasion regardless of the accessibility to the simulation model's source code. It should be noted that the current study - and consequently this dimension - is not concerned with the specific programming language the simulation is built on (Assembly, C++, NetLogo, etc.), but solely with whether the application of a V&V method depends upon having access to the source code.

Parts of this chapter have been published in Roungas et al. (2017) and Roungas et al. (2018f).

Table 8.1. List of V&V methods & properties of simulations.

Method	Source Code	Real Data	Game	Requirements	Purpose
1. Acceptance Testing	No	No	Yes	Both	O. Val.
2. Alpha Testing	No	No	Yes	Both	O. Val.
3. Assertion Checking	Yes	No	No	Hard	M. Ver.
4. Audit	Yes	No	Yes	Soft	M. Ver.
5. Beta Testing	No	No	Yes	Both	O. Val.
6. Bottom-Up Testing	Yes	No	No	Both	M. Ver.
7. Cause-Effect Graphing	Yes	No	No	Hard	M. Ver.
8. Comparison Testing	No	No	No	Both	C.M. Val.
9. Compliance Testing: a) Authorization Testing	No	No	No	Soft	M. Ver.
b) Performance Testing	No	No	No	Soft	M. Ver.
c) Security Testing	No	No	No	Soft	M. Ver.
d) Standards Testing	No	No	No	Soft	M. Ver.
10. Control Analysis: a) Calling Structure Analysis	Yes	No	No	Hard	C.M. Val.
b) Concurrent Process Analysis	Yes	No	No	Hard	M. Ver.
c) Control Flow Analysis	Yes	No	No	Hard	C.M. Val.
d) State Transition Analysis	Yes	No	No	Hard	D. Val. & M. Ver.
Data Analysis: a) Data Dependency Analysis	Yes	No	No	Hard	D. Val. & M. Ver.
b) Data Flow Analysis	Yes	No	No	Hard	D. Val. & M. Ver.
11. Debugging	Yes	No	No	Both	M. Ver.
12. Desk Checking	Yes	No	Yes	Both	M. Ver.
13. Documentation Checking	Yes	No	Yes	Both	C.M. Val.
14. Execution Testing: a) Execution Monitoring	No	No	No	Hard	C.M. Val.
b) Execution Profiling	No	No	No	Hard	C.M. Val.
c) Execution Tracing	Yes	No	No	Hard	C.M. Val.
15. Face Validation	No	Yes	Yes	Both	O. Val.
16. Fault/Failure Analysis	No	No	No	Hard	C.M. Val.
17. Fault/Failure Insertion Testing	No	No	No	Hard	C.M. Val.
18. Field Testing	No	Yes	No	Both	O. Val.
19. Functional (Black-Box) Testing	No	Yes	No	Hard	C.M. Val.
20. Graphical Comparisons	No	Yes	Yes	Both	O. Val.
21. Induction	No	No	No	Both	C.M. Val.
22. Inference	No	No	No	Both	C.M. Val.
23. Inspections	No	No	No	Both	C.M. Val.
24. Interface Analysis: a) Model Interface Analysis	No	No	No	Soft	C.M. Val.
b) User Interface Analysis	No	No	Yes	Soft	O. Val.
25. Interface Testing: a) Data Interface Testing	No	No	No	Soft	D. Val.
b) Model Interface Testing	No	No	No	Soft	C.M. Val.
c) User Interface Testing	No	No	Yes	Soft	O. Val.
26. Lambda Calculus	Yes	No	No	Hard	M. Ver.
27. Logical Deduction	No	No	No	Both	All
28. Object-Flow Testing	No	No	No	Hard	O. Val.
29. Partition Testing	Yes	No	No	Hard	C.M. Val.
30. Predicate Calculus	Yes	No	No	Hard	M. Ver.
31. Predicate Transformations	No	Yes	No	Hard	M. Ver.
32. Predictive Validation	No	Yes	No	Hard	O. Val.
33. Product Testing	No	No	Yes	Both	O. Val.
34. Proof of Correctness	Yes	No	No	Hard	C.M. Val. & M. Ver.
35. Regression Testing	Yes	No	No	Hard	M. Ver.
36. Reviews	No	No	Yes	Both	C.M. Val.
37. Semantic Analysis	Yes	No	No	Both	M. Ver.
38. Sensitivity Analysis	No	No	No	Hard	O. Val.
39. Special input testing: a) Boundary Value Testing	Yes	No	No	Both	M. Ver.
b) Equivalence Partitioning Testing	No	No	No	Hard	O. Val.
c) Extreme Input Testing	No	No	No	Hard	O. Val.
d) Invalid Input Testing	No	No	No	Hard	O. Val.
e) Real-Time Input Testing	No	Yes	No	Hard	O. Val.
f) Self-Driven Input Testing	No	No	No	Hard	O. Val.
g) Stress Testing	No	No	No	Hard	O. Val.
h) Trace-Driven Input Testing	Yes	Yes	No	Both	D. Val. & C.M. Val.
40. Structural (White-box) Testing: a) Branch Testing	Yes	No	No	Both	C.M. Val. & M. Ver.
b) Condition Testing	Yes	No	No	Both	C.M. Val. & M. Ver.
c) Data Flow Testing	Yes	No	No	Both	C.M. Val. & M. Ver.
d) Loop Testing	Yes	No	No	Both	C.M. Val. & M. Ver.
e) Path Testing	Yes	No	No	Both	C.M. Val. & M. Ver.
f) Statement Testing	Yes	No	No	Both	C.M. Val. & M. Ver.
41. Structural Analysis	No	No	No	Hard	C.M. Val.
42. Submodel/Module Testing	No	No	No	Both	C.M. Val.
43. Symbolic Debugging	Yes	No	No	Hard	M. Ver.
44. Symbolic Evaluation	Yes	No	No	Hard	C.M. Val.
45. Syntax Analysis	Yes	No	No	Hard	M. Ver.
46. Top-Down Testing	Yes	No	No	Both	C.M. Val.
47. Traceability Assessment	Yes	Yes	No	Both	C.M. Val.
48. Turing Test	No	Yes	No	Both	O. Val.
49. Visualization/Animation	No	Yes	Yes	Both	O. Val.
50. Walkthroughs	No	No	Yes	Both	C.M. Val.

8

2. Does the V&V method require data from the real system?
   Possible answers: Yes or No. A positive answer to this question means that this method can only be used when data from the real system are available, whereas a negative answer means that it can be used in any occasion regardless of the availability of data from the real system. It should be noted that the current study - and consequently this dimension - is not concerned with the nature of the data in general (qualitative or quantitative), but solely with their existence and availability.

3. Is the V&V method suitable for a game V&V study?
   Possible answers: Yes or No. While all methods are suitable for pure simulations, some of them will be also suitable for games in particular. Although games often have a simulation model running on the background, in which case all methods would be applicable, in this study the term *game* is used to describe the layer that is on top of the simulation model and refers to the players' interaction.

4. For what type of requirements is the V&V method more suitable?
   Possible answers: Hard (Functional), or Soft (Non-Functional), or Both. A method might be focused on either the functional part or the non-functional part of the model or on both.

5. For which type of study is the V&V method more suitable?
   Possible answers: Data Validation (D. Val.), Conceptual Model Validation (C.M. Val.), Model Verification (M. Ver.), or Operational Validation (O. Val.). A method might be suitable for one or more of the available categories.

8

Table 8.1 summarises the results of the analysis. The intended use of Table 8.1 is to act as a filtering mechanism. Whenever an individual or a team wants to verify and/or validate a simulation model, they can utilise this table to narrow down the applicable methods according to the different properties of the simulation at hand. The selection process is shown in Figure 8.1.

With regards to the first property, i.e., the accessibility to the source code, and in contrary to the second property, access to the source code does not imply that the methods categorised under "Yes" are stronger. Usually, access to the source code is associated with verification and in some cases conceptual model validation.

With regards to the second property, i.e., the availability of data from the real system, the methods categorized under "No" can be used irrespective of whether real data exist. Nevertheless, the methods categorized under "Yes" are more powerful in the sense that, if used appropriately, they provide evidence or a data trace of how the simulation should work. Hence, whenever real data are available, the methods categorized under "Yes" should be preferred, unless an alternative method is definitely more suitable.

With regards to the third property, i.e., the suitability of certain methods for the V&V of games, informal methods (Balci, 1998) seem to be the ones suitable for games. This is a preliminary conclusion that is expected to an extent. In games representing Complex Adaptive Systems (CAS), experts' opinion plays an important, and perhaps the most important, role (Meijer, 2012a), regardless of the game's level of fidelity (van Lankveld et al., 2017) or use of technology (Meijer, 2012a). It should be noted that although the term

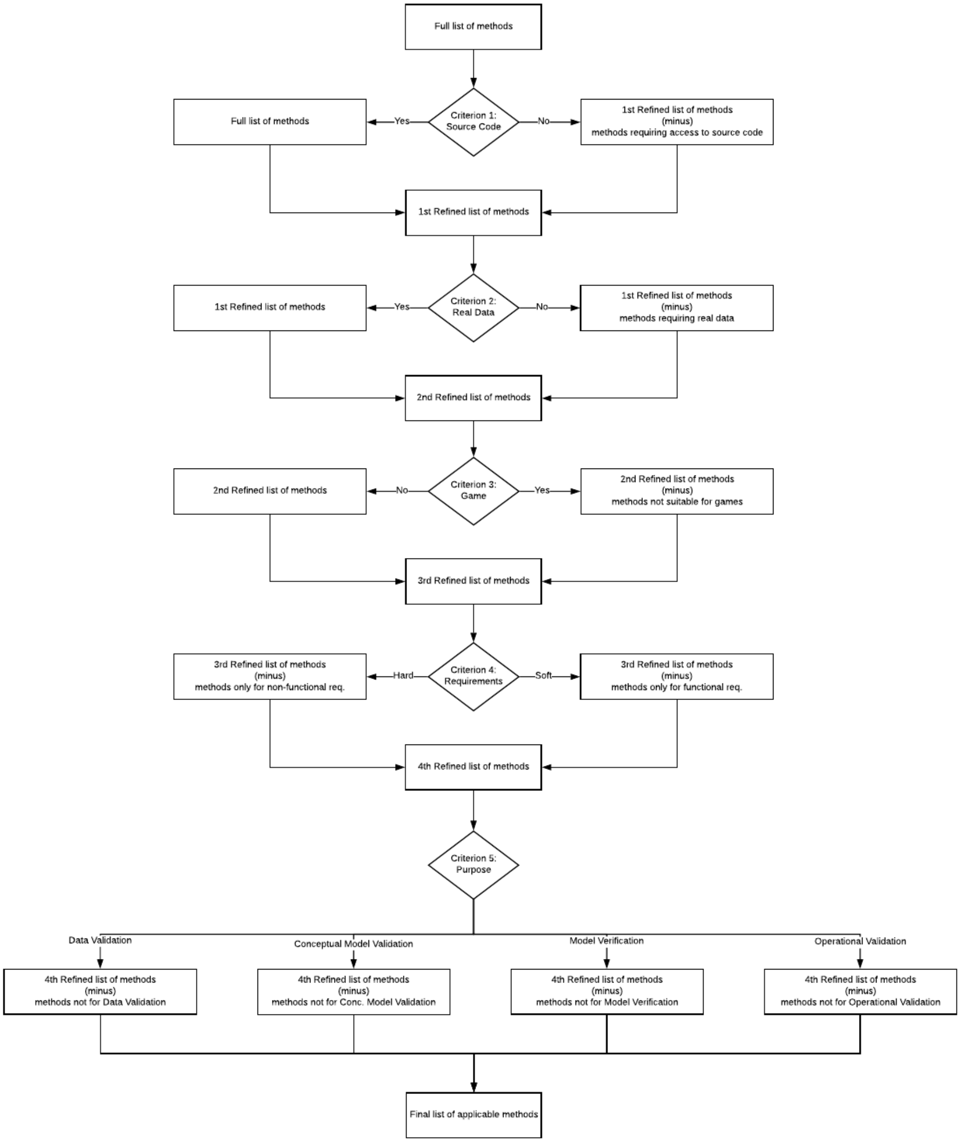


Figure 8.1. The flow diagram of the selection process of methods.

Games assumes both high-tech (computer-based) and low-tech (e.g., tabletop) games, the selection in Table 8.1 was made with a bias towards the high-tech games.

With regards to the fourth and fifth property, i.e., the type of requirements being tested and the purpose of the V&V study respectively, the answers are more or less self-explanatory. Some methods are more suitable for testing one type of requirement. As an example, regression testing is more appropriate for functional requirements (hard goals). Other V&V methods are better suited for one purpose, such as Structural (White-box) testing, which is more appropriate for conceptual model validation, while others are more suitable for testing both types of requirements (e.g., Graphical comparisons), or for more than one purpose (e.g., Trace-Driven Input Testing).

The novelty of the proposed framework does not lie in the content of Table 8.1 per se, but on the idea that the list of methods can be narrowed down to a manageable level, thus making the V&V of a simulation better grounded, faster, more accurate, and more cost effective.

There is a threat towards the validity of the content on Table 8.1. The line between whether data from the real system are needed, or whether access to the source code is needed, or whether a specific requirement is definitely functional or non-functional, or whether the purpose is to validate the data, the conceptual model, the operational ability of the model, or to just verify the model, is not always clear and well defined. In Section 8.4, future steps are proposed aiming at addressing and mitigating the above mentioned threat.

8.2. STATISTICAL TECHNIQUE SELECTION METHODOLOGY

The list of simulations' and systems' characteristics that influence the selection of techniques, which are explained in more detail in Section 7.2.2, are expressed in questions, as follows:

1. How many different datasets are going to be examined?
   Possible answers: 1, 2, and/or >2.

2. How many different variables are going to be examined?
   Possible answers: 1 or >1.

3. What is the purpose of the statistical test?
   Possible answers: Mean equality, Complexity reduction, Goodness-of-fit, Heteroscedasticity, or Time Series analysis.

4. Are the sample parameters (μ, σ^2) known?
   Possible answers: Yes or No.

5. What kind of data are going to be examined?
   Possible answers: Numerical or Categorical.

6. What is the sample size?
   Possible answers: Large, Small, or Any.

Table 8.2 summarises the results of the analysis. Similarly to Table 8.1, the intended use of Table 8.2 is to act as a filtering mechanism. Whenever an individual or a team wants to verify and/or validate a simulation model, they can utilise this table to narrow down the applicable techniques according to the different characteristics of the simulation at hand and the system. It should be noted that another significant factor in selecting a technique is the statistical power of the technique, i.e., the probability that the null hypothesis (H_0) is correctly rejected for the alternative hypothesis (H_1). The statistical

TABLE 8.2. LIST OF STATISTICAL TECHNIQUES.

Statistical Techniques	# of datasets	# of variables	Purpose	known parameters	Type of data	Sample size
t-Test	1 or 2	1	Mean equality	Yes	Numerical	Any
Hotteling's T^2 Test	1 or 2	>1	Mean equality	Yes	Numerical	Any
Analysis of Variance	>2	1	Mean equality	Yes	Numerical	Any
Multivariate Analysis of Variance	>2	>1	Mean equality	Yes	Numerical	Any
Simultaneous Confidence Intervals	1 or 2	>1	Mean equality	Yes	Numerical	Any
Factor Analysis	1	>1	Complexity reduction	Yes	Numerical	Any
Principal Component Analysis	1	>1	Complexity reduction	Yes	Numerical	Any
Kolmogorov-Smirnov Test	1 or 2	1	Goodness-of-fit	No	Numerical	Any
Chi-squared Test	1 or 2	1	Goodness-of-fit	No	Numerical & Categorical	Any
Anderson-Darling Test	1 or 2	1	Goodness-of-fit	No	Numerical	Any
Cramér-von Mises Criterion	1 or 2	1	Goodness-of-fit	No	Numerical	Any
Kuiper's Test	1 or 2	1	Goodness-of-fit	No	Numerical	Any
Coefficient of Determination	2	1	Goodness-of-fit	Yes	Numerical	Any
Mann-Whitney-Wilcoxon Test	2	1	Mean equality	No	Numerical & Categorical	Small
White Test	2	1	Heteroscedasticity	Yes	Numerical	Any
Glejser Test	2	1	Heteroscedasticity	Yes	Numerical	Any
Spectral Analysis	2	1	Time Series analysis	Yes	Numerical	Any
Durbin-Watson Statistic	2	1	Time Series analysis	Yes	Numerical	Any

power of a technique is not predetermined, which is the reason it is not included in this analysis. Nevertheless, the Neyman-Pearson lemma (Neyman and Pearson, 1933) is a test that determines which technique is the one with the greatest statistical power given several attributes, like the sample size and the statistical significance.

8.3. A CASE STUDY

In this section, a case study illustrates how the framework, through the use of Table 8.1 and Table 8.2, can be used. The case study is a computer simulation of a particular instantiation of the Dutch railways. The authors were assigned to validate the simulation model with regards to punctuality, in other words the precision of the delays of trains in the model.

In more detail, the simulation model was built on the FRISO simulation package (Middelkoop and Loeve, 2006), which is ProRail's in-house simulation environment. Being a microscopic simulation environment, FRISO has the potential to - and depending on the model it usually does so - simulate the railway network in a detailed manner; it has the ability to depict the network down to a switch level, which is the case with this model. The model was built in 2014 and it simulates the train operations in one of the most heavily utilised sections (Amsterdam Central station - Utrecht Central station) of one of the largest corridors in the Netherlands (A2), during the whole month of June 2013. The intended use of the model was to examine the punctuality of the timetable with the particular focus being the Amsterdam and Utrecht central stations. A more in depth description of the model, including its input, output, and the final results can be found in (Roungas et al., 2018e).

With regards to the methods, the initial list, as it is shown in Table 8.1, consists of 75 methods. Then with every step, the list is narrowed down. For this particular study, the selection process for each property, as shown in Figure 8.2, was as follows:

1. Access to the source code was not available; *Answer: No.* Using this criterion the available methods were reduced to 42.

2. There were available data from the real system; *Answer: Yes.* Using this criterion

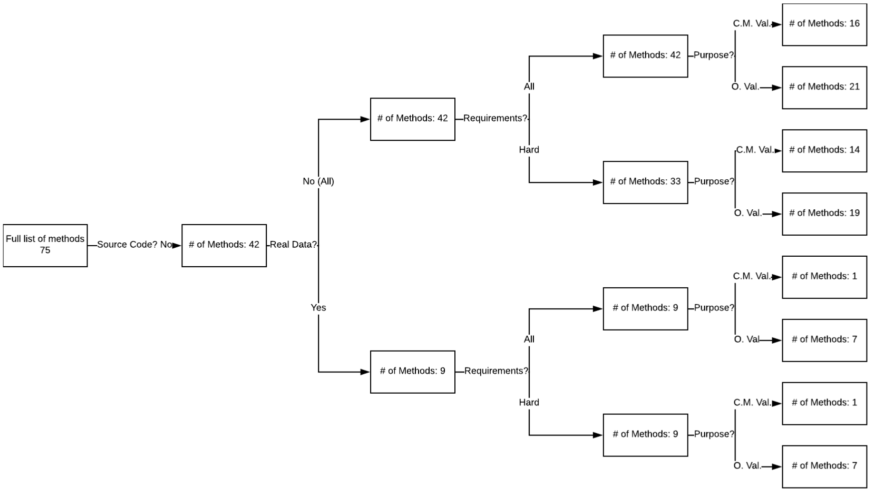


Figure 8.2. A tree graph of the method selection process.

eliminated 33 more methods returning a total of 9 available methods. Nonetheless, all 42 methods could have be used in this particular case.

3. The main focus was on the punctuality, ergo functional (hard) requirements, but comments were also expected on the non-functional (soft) requirements; *Answer: Both (but main focus on hard)*. If on the previous criterion *Yes* was chosen as an option, choosing either *Both* or *Hard* on this criterion would leave the list intact (Total 9 methods).

4. The study was mainly concerned with the operational validity of the simulation, but to a degree also with the conceptual model validity; *Answer: C.M. Val & O. Val.*. Using this criterion and based on the selections on the previous criteria, the final number of available methods was reduced to 1 for the conceptual model validation and 7 for the operational validation.

For the operational validation, which was the primary interest of the study, the final list of the seven methods is shown in Table 8.3. Out of this list, in total four methods were used, namely: the *Face Validation*, *Graphical Comparisons*, *Predictive Validation*, and *Turing Test*. *Predictive Validation* was first used to handle the initial datasets (simulation dataset & operational dataset) and to produce results for the different statistical tests. Then, a combination of the remaining three methods was used to ascertain the validity of the simulation.

With regards to the techniques, the initial list, as shown in Table 8.2, consists of 18 techniques. Then with every step, the list is narrowed down. For this particular study, the selection process for each characteristic was as follows:

1. The model's and reality's output were examined;
Answer: 1 or 2. Using this criteria reduced the available techniques to 16.

2. The amount of delays was the focus;
   Answer: 1. Using this criterion eliminated 4 more techniques totalling in 12 available techniques.

3. The study was concerned with whether the delays between the model and reality were similarly distributed and whether the averages were significantly different;
   Answer: Mean equality and Goodness-of-fit. Using this criterion resulted in 2 suitable techniques for mean equality and 6 for goodness-of-fit.

4. The sample parameters (μ, σ^2) were known;
   Answer: Yes. This is a criterion that only influences the results if the answer is *No*, since the non-parametric techniques can still be used when the mean and variance are known. Therefore the number of techniques remained the same.

5. The delays were in seconds, hence numerical;
   Answer: Numerical. Using this criterion eliminated 1 techniques for the mean equality, resulting in 1 technique for mean equality and 6 for goodness-of-fit.

6. Each sample was larger than 100;
   Answer: Large. This last criterion did not further reduced the number of techniques, since the only techniques suitable for small datasets (*Mann-Whitney-Wilcoxon Test*) had been eliminated in a previous step.

For testing the equality of means, the only suitable techniques, i.e., t-test, was used. Whereas for testing the goodness-of-fit, from the 6 suitable techniques, the Kolmogorov-Smirnov and chi-squared test were used.

Table 8.3. Refined list of V&V methods of the Case Study.

Method	**1**	**2**	**3**	**4**
Face Validation	No	Yes	Both	O. Val.
Field Testing	No	Yes	Both	O. Val.
Graphical Comparisons	No	Yes	Both	O. Val.
Predictive Validation	No	Yes	Hard	O. Val.
Real-Time Input Testing	No	Yes	Hard	O. Val.
Turing Test	No	Yes	Both	O. Val.
Visualization/Animation	No	Yes	Both	O. Val.

In this section, the use of the proposed framework demonstrates clearly its effectiveness. As shown in Table 8.3, the initial list of 75 methods was narrowed down in a matter of minutes to the manageable level of seven; and similar reduction occurred in the techniques. By all means, the effectiveness of the framework is not only evident due to its time-saving nature but also due to the fact that it ensures that the chosen methods and techniques are appropriate for the simulation and the system at hand as well as for the purpose of the V&V study.

8.4. Final Remarks on the Methodology

In this chapter, based on the literature examined in Chapter 7, a framework for simulation validation and verification method and statistical technique selection was proposed. Various properties and characteristics of simulations and systems were taken into account and it was shown that indeed some of these influence the method and technique selection and thus, the final results of the simulation study.

Moreover, the framework was applied on a case study, as a first step towards verifying its effectiveness. The case study showed that the framework is an effective time-saving tool, which also provides a safety net for choosing the methods and techniques that best serve the intended purpose of the simulation and the V&V study.

With regards to future work, additional simulation properties should be identified that may potentially influence the method selection, or some of the discarded properties, identified in Section 7.2.2, might prove to be more influential than initially acknowledged. Moreover, there is a need to further verify the connection of each method to the simulation model's properties and the purpose for which they are more suitable; in other words, it should be verified that the answers on columns 2-6 in Table 8.1 are correct. With regards to the techniques, a more extensive list analysed in the same way as in Section 8.2 would provide for an improved guide towards selecting the most effective techniques given the problem at hand. Finally, more case studies, from the authors and more importantly from researchers unrelated to the authors, both in pure simulations and in games, would further strengthen the validity and applicability of the framework.

Nevertheless, this study paves the way for future research in the topic, and as discussed earlier, the main contribution of the framework does not lie in the results presented in Table 8.1 and Table 8.2, but is related to the identification of the relationships between the methods, the techniques, the simulation's and system's properties, and the purpose of the V&V study. Therefore, it is of utmost importance that any future research be focused on these relationships.

9

Automation in Simulation Validation

In this chapter, a methodology for automating the validation of simulations is proposed. The case studies analysed in this chapter are two simulation models that have been used in games, hence the relevance with, and the potential impact to, the gaming field.

9.1. Introduction

During the last few decades, systems characterised as Complex Adaptive Systems (CAS) (Axelrod and Cohen, 1999) have become increasingly popular. This complexity, as insinuated by their name, does not arise just from the increase of their size but also from the high level of involvement of humans in many of those systems' internal processes. In turn, since human behavior cannot be characterised as 100% rational, these systems often tend to behave in a seemingly irrational way. Such an example are the decision making processes in areas like healthcare (Rouse, 2008) and transportation (Rinaldi et al., 2001).

The societal and financial impact of decisions made in CAS necessitates for an affordable, ethical, and risk-free way to test potential changes or threats to the system at hand. As such, simulations, which are one of the most popular - if not the most popular - ways of accomplishing that, are deemed to be the most appropriate tool.

Despite the significant help simulations can provide, they should not be, and usually are not, trusted blindly. Simulations should be thoroughly validated, ensuring - at least to some extent - that their results are credible and can be used for the intended purpose. There are multiple methods for validating simulations (Balci, 1998), but these methods are not the only aspect validation success relies upon; the type of validation (Conceptual Model & Operational) (Sargent, 2000), as well as the intended purpose and audience play

Parts of this chapter have been published in Roungas et al. (2018g).

an important role on how to approach the validation results. This chapter is concerned with the operational validation of simulation models (Sargent, 2000), and more specifically with how Web 3.0 technologies and the R statistical language can help mitigate several restrictions that occur during the standard validation life-cycle (Balci, 1994).

More specifically, simulations are characterised by the fact that they regularly need updates either in the form of verification (e.g. debugging), or in the form of conceptual model validation (e.g. improvements in the core algorithms of the simulation model). In turn, following each update, the model should run again in order to assert its operational validation. As a result, a team of validation experts should often be available to perform both the conceptual model and the operational validation (Bergmann and Strassburger, 2010); this substantially increases i. the cost of the validation study, ii. the time between the simulation is finalised by the modellers and the moment it can be used for a formal study, iii. the probability for a human error to occur due to the numerous calculations involved in the procedure (Balci, 1997), and iv. the dependence on specific validation experts, who are knowledgeable of the particularities of the simulation at hand (Balci, 1998).

The conceptual model validation, being a technical area, is quite difficult to automate, albeit not impossible. On the other hand, the operational validation is more straightforward in terms of modelling and automation, since in many cases it utilises statistical techniques. As such, the research question that will be covered in this chapter is:

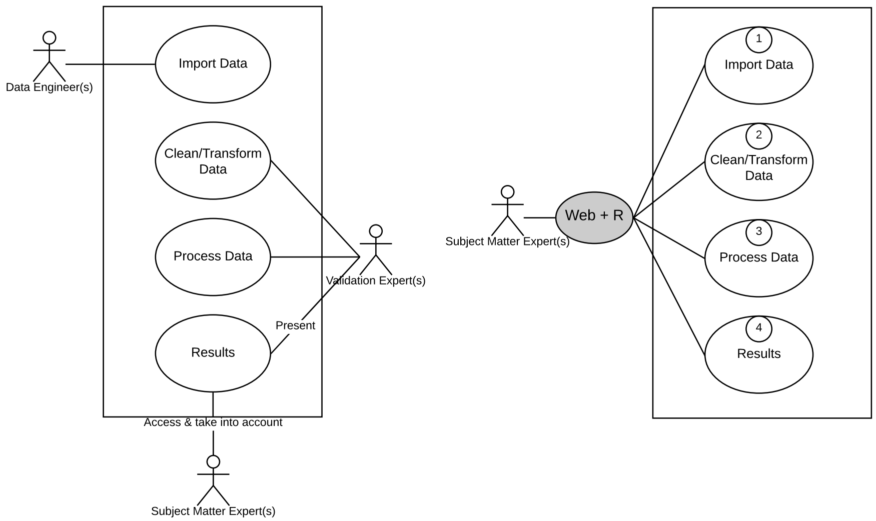
> *Research Question:* How can the operational validation of a simulation model be automated or semi-automated, in order to reduce the time, cost, and human error associated with it?

This chapter starts with two assumptions: i. automation, or at the very least semi-automation, can be achieved on an operational validation (hereinafter referred to as validation) study, and ii. Web 3.0 technologies and the R statistical language, given their numerous advantages described in Section 9.3.4, can be used to accomplish that. The chapter aims at answering the research question by first proposing a framework for automated simulation model validation and then demonstrating how this framework was used, in order to automate the operational validation study of two simulation models in the railway sector.

In Section 9.2, the state of the art on automated simulation model validation is identified. In Section 9.3, a web-based framework for simulation model validation is proposed. In Section 9.4, a proof of concept of the framework based on two different case studies is presented. Finally, in Section 9.5, the future steps are illustrated and final remarks are made.

9.2. BACKGROUND WORK

Automated simulation model validation is an issue that has increasingly gaining awareness within the simulation community (Balci, 1998). In various areas, automated validation has been proposed and operationalised as a way to mitigate one or more of the risks associated with validation. Studies on automated validation have been performed



(a) Manual validation (b) Automated validation

Figure 9.1. Use case diagram of the validation life-cycle.

in the automotive industry (Albers et al., 2012, Kum et al., 2006), in pedestrian simulations (Porzycki et al., 2015), in biological models (Cooper, 2008), even in human-device interfaces (Bolton, 2013). Nevertheless, the amount of research and subsequently the amount of practical applications is still rather limited, and most of the approaches are either domain specific or lack many of the traits that can help mitigate the risks associated with validation, as the latter were identified in Section 9.1.

On the other hand, despite the huge influence and usage of the World Wide Web (Chandrasekaran et al., 2002), and consequently of web technologies, to the best of our knowledge, there are no frameworks or tools proposing or supporting the utilisation of web technologies for building simulation model validation solutions. While, web technologies have been used widely to build simulation models in various fields (Byrne et al., 2010), their application usually stops after the modelling and before the validation phase. The closest attempt towards a web-based simulation model validation environment has been the Evaluation Environment (EE) (Balci et al., 2002), which is a web-based client/server software system that enables geographically dispersed people to conduct complex evaluation projects in a collaborative manner. Nevertheless, EE is not a validation tool but rather a tool more suitable for complex evaluations, such as modelling and simulation credibility assessment, which requires rigorous collaboration among technical people, subject matter experts, engineers, project managers, and program managers.

The final decision regarding the validity of a simulation model is made by subject matter experts (SMEs). Hence, any tool aimed at helping those SMEs can be considered to be a decision support system (DSS) (Landry et al., 1983). Unlike in simulation model validation, in decision support, web technologies have been used in several different occasions (Bhargava et al., 2007). In particular, the type of DSS that bears a significant resemblance to a simulation model validation process is the model-driven DSS (Power, 2004). A model-driven DSS uses formal representations of decision models and provide analytical support using the tools of decision analysis, optimization, stochastic modeling, simulation, statistics, and logic modeling (Bhargava et al., 2007).

Web-based DSSs, including model-driven DSSs, have experienced a significant increase research-wise in the last decade (Blomqvist, 2014), which is a result of their numerous advantages. The field of simulation model validation could be similarly benefited by web technologies, despite the so far infinitesimal amount of research. The forthcoming sections of this chapter demonstrate a framework and two applications of this framework in which the advantages of web technologies for simulation model validation are illustrated and the research question stated in Section 9.1 is addressed.

9.3. THE AUTOMATED VALIDATION FRAMEWORK

In this section, a web-based framework for simulation model validation is proposed. The framework has three main components: i. the steps taken throughout the standard validation life-cycle, ii. the actors involved in the whole process, and iii. the architecture of the validation tool. In Section 9.3.1, Section 9.3.2, and Section 9.3.3 the three main components of the framework are presented respectively.

9.3.1. VALIDATION STEPS

Every validation study is different, even if it is about the same simulation software and model. This is due to the multiple elements that define simulations (Roungas et al., 2017). Depending on the nature of the study, these elements might be the formalism the simulation is based on (Vangheluwe et al., 2002), the fidelity level of the simulation (Liu et al., 2008), the type of the simulation (Constructive, Virtual, Live) (Morrison and Meliza, 1999), to name a few. Despite the uniqueness of every validation study, the steps that need to be taken throughout the life-cycle of the validation are common. Namely, these steps are:

1. **Import data.** Data from both the simulation and the real system under study need to be imported for further analysis. Data import can be performed in multiple ways. The simplest one is through a user interface (UI), while the most efficient one is by directly running a script in the database. The former is easy but usually extremely slow especially in today's big data era. The latter is fast but requires technical expertise that many SMEs do not possess. Hence, the automation of data import requires a hybrid approach. In the two case studies presented in Section 9.4, two different approaches are tested.

2. **Clean & transform data.** Cleaning data from outliers, thus ensuring accuracy, and transforming data to the same units, thus making data between the model and reality comparable, is usually required. The cleaning and transformation of data can be fully automated, semi-automated, or a combination of both.

   Full automation is implemented for known cleaning and transformation issues, such as converting kilometers to meters or geodesic coordinates to longitude and latitude. Fully automated scripts can run either immediately upon importing the data into the database or after a user's request.

Semi-automation is implemented when the cleaning or transformation criteria are not predetermined, but instead should be defined by the user. Such an example is the acceptable deviation of GPS data, in which case depending on the study, a user might choose a more strict or more loose threshold.

Regardless of the level of automation, data cleaning and transformation is a tedious task aiming at ensuring a high quality of data, which depends on multiple criteria (Huang, 2013).

3. **Process data.** The 3^{rd} step of the validation, the processing of data, is the core of the implementation. This is the step in which the actual validation of the model takes place. It includes the statistical analysis and the design of all the visualizations. In the proposed framework, in this phase, R is used to perform the statistical calculations and create the necessary graphs.

4. **Present results.** The last step of the validation study, the presentation of the results, is concerned with how the processed data and the visualizations produced in step 3 are presented to the SMEs. This is the step in which the question: *Is the model valid?* is answered. This question can only be answered in a given context, and presenting the results of the validation study in light of its context to the responsible stakeholders accomplishes that. All three previous steps are preparatory for this phase, in which SMEs can finally assess the simulation results and adjudicate on their validity for the intended purpose of use. While steps 1 and 2 are relatively independent with each other and with the remaining two steps, steps 3 and 4 are closely connected to each other. Design decision made in step 3 pertaining to the statistical tests and the visualizations, are directly affecting the way results are presented to the SMEs and thus, influence not only in terms of content but also in terms of context the final verdict with regards to the validation of the model.

9.3.2. Actors

In the current manual state of the validation life-cycle, each of the steps identified in Section 9.3.1 requires one or more actors to perform it and usually more than one type of actors are needed to perform the complete life-cycle (Landry et al., 1983). Depending on the nature and scale of the model, the number of different actors required for the validation of the model can vary. With regards to simulations for decision making, an example of which for the current manual state is shown in Figure 9.1(a), usually requires three type of actors. Namely, these actors are:

1. Data engineers, who are the ones responsible for fetching the data from the output of the simulation model and provide an environment in which the validation experts can interact with the data.

2. Validation experts, who are the ones responsible for cleaning, transforming, and processing the data providing results for the subject matter experts.Validation ex-

perts should have knowledge of the problem domain and expertise in the modeling and validation methodology (Balci, 1998).

3. Subject matter experts (SMEs), who are the ones responsible for evaluating the results provided from the validation experts and deciding about the validity of the simulation model.

While, the SMEs are necessary for the final decision on the validation of the simulation model, the other actors can be potentially omitted from the validation life-cycle. Figure 9.1(b) shows what the ideal situation of an automated validation life-cycle would look like, which is also the final goal of this chapter. In Figure 9.1(b), the numbers indicate the different steps. These numbers are used throughout this chapter to indicate the parts of the architecture and the implementation that correspond to each step. The idea behind Figure 9.1(b) is for the SMEs to interact as less as possible with the tool during the first three steps and focus mainly on evaluating the results presented to them during step 4. The different levels of automation and the ways to accomplish that are described in detail in Section 9.4.

9.3.3. ARCHITECTURE

The architecture of the proposed web-based framework consists of three main components:

- The web browser, which incorporates the user interface (UI) and is how the SMEs interact with the data and the results.
- The web server, which includes all the web files (e.g. HTML, CSS, etc.), the R scripts, and the necessary interfaces for their communication with the web browser and the database, and
- The database server, which houses the database.

These three main components are the minimum requirements of an implementation based on the proposed framework. Each component can have different *flavors* depending on the nature and the size of the final implementation.

9.3.4. ADVANTAGES & DISADVANTAGES

The utilization of web technologies and R improves multiple Non-Functional Requirements (NFR) (De Weck et al., 2011), also known as *-ilities*. Namely, some of these NFR are:

- **Usability.** Usability is a quality attribute that assesses how easy a UI is to use (Nielsen, 2003). There is an abundance of research and case studies on how modern applications can be user-friendly. Particularly, web content has drawn most of the attention. As a result, building web interfaces has become increasingly more straightforward due to the numerous guidelines, which are also targeted for different user profiles. Hence, a web based system, like the one proposed in this chapter, can provide an intuitive UI to SMEs for easily previewing and processing data.

- **Affordability.** Affordability is a collection of attributes, alternatives, and decision-making processes that allow someone to determine if a product is affordable (Bever and Collofello, 2002). A web based system, like the one proposed in this chapter, is quite affordable since web tools, like PHP and JavaScript, and R are open source programming languages that bear no cost to license them. Additionally, those tools' popularity makes it affordable to hire or train people, who can then help build and maintain such a system. Finally, such an implementation in which all the complex calculations are performed in the server minimizes the cost of investment in end devices.

- **Portability.** Portability is a measure of the ease with which a program can be transfered from one environment to another (Tanenbaum et al., 1978). Modern devices (laptops, tablets, smartphones) coupled with the responsive design of web content converts these devices into portable working stations. Moreover, due to the fact that all complex calculations can be set up to be performed in the server, even the simplest hardware can be adequate to satisfy a user's needs. Hence, a web based validation system can provide instant access to SMEs regardless their hardware, software environment, or even geographic location.

- **Interoperability.** Interoperability is the ability of different systems and software applications to communicate, to exchange data accurately, effectively, and consistently, and to use the information that has been exchanged (Heubusch, 2006). The specific implementation presented in this chapter is an example on the interoperability of web technologies. Web technologies (like HTML, CSS, JavaScript, PHP, SQL) are known for communicating well with each other. What makes these tools remarkable in terms of interoperability is their ability to also seamlessly bidirectionally communicate with external scripting or even programming languages, like it is demonstrated in this chapter with R.

- **Accessibility.** Accessibility refers to the design of products, devices, services, or environments for people who experience disabilities (Henry et al., 2014). Internet browsers are increasingly becoming both directly (without assistance) and indirectly (with assistance like screen readers, braille writing devices etc.) accessible to users who experience disabilities. Hence, a web based validation system can enable disabled SMEs to be more engaged in the validation process and offer their expertise.

Along with the NFR, there are also other advantages of using web technologies and R throughout a validation study. These technologies offer high levels of customizability both in the backend and frontend, thus fitting different needs, depending on the simulation at hand. Moreover, the use of Ajax or in the case of R a package like $shiny$, enables to build interactive web applications, where different parts of the data can be used in real-time. Finally, web technologies' and R's versatility promotes both quick prototyping (proof of concept) and full scale commercial implementations with animations, interactive content etc.

Nevertheless, web technologies do have some disadvantages. Web applications, built upon a framework like the one proposed in this chapter, do not depend on a local implementation. The browser serves only as the UI for the end user and a server to perform all the calculations is necessary, which means that Internet access is required at all times. A workaround to this limitation is for the end users to have an exact copy of the server implementation in their *localhost* but this requires some level of expertise on managing databases and perhaps web applications in general, as well as it eliminates the portability and affordability advantages. Another drawback of web technologies is that they are still an evolving field. While, there have been major steps forward the past decade, there are still performance issues especially with animated and interactive content.

9.4. IMPLEMENTATION

In this section, a proof of concept based on the proposed framework for the validation of two distinct simulation models in the railway sector is presented. The models were built to run on two separate simulation environments, FRISO (Middelkoop and Loeve, 2006) and OpenTrack (Nash and Huerlimann, 2004), which are both railway microscopic simulation environments. Thus, they have the potential to, and depending on the model usually do, simulate the railway network in a detailed manner; both have the ability to depict the network down to a switch level. Despite the pointed similarities, these two simulation packages are different, thus suitable for different usages, as revealed by being used in different studies, namely in a punctuality and a conflict detection study respectively. Therefore, the nature of the simulation studies combined with the significantly dissimilar data asked for a diverse approach with regards to their validation. As a result, while the first 2 steps of the validation life-cycle (Data import & Data cleaning and transformation) required similar methods to implement, for steps 3 and 4 two different tools were built. In spite of the use of Web technologies and R on both of them, there was a significant distinction on their design approach.

Web 3.0 technologies include a vast selection of tools and programming languages spanning from frontend to backend. This study and the resulted proof of concept were developed using a MySQL database schema, PHP on the backend, and basic JavaScript with HTML and CSS on the frontend. Therefore, the syntax and commands used throughout this chapter are that of PHP, SQL, JavaScript, shell script, and R. The deployment diagram of the implementation is depicted in Figure 9.2. The circled numbers correspond to the steps shown in Figure 9.1(b) and indicate where in the implementation each step takes place.

9.4.1. DATA IMPORT

In this study, two different approaches of data import are tested. The most common file extensions used to import data into a database are the .sql and .csv (Comma Separated Values), thus the subsequent analysis is concerned only with these two extensions.

The first approach is using exclusively the back-end programming language of choice. In PHP, for a CSV file, this is accomplished by using the function $fgetcsv()$. In Java, the equivalent is function $Scanner()$, while in Python the module csv. Similarly, for an SQL file, this is accomplished by using the function $file\_get\_contents()$. In Java, the equivalent is class $SqlReader$, while in Python the function $read$.

The second approach is using an execution function of the programming language

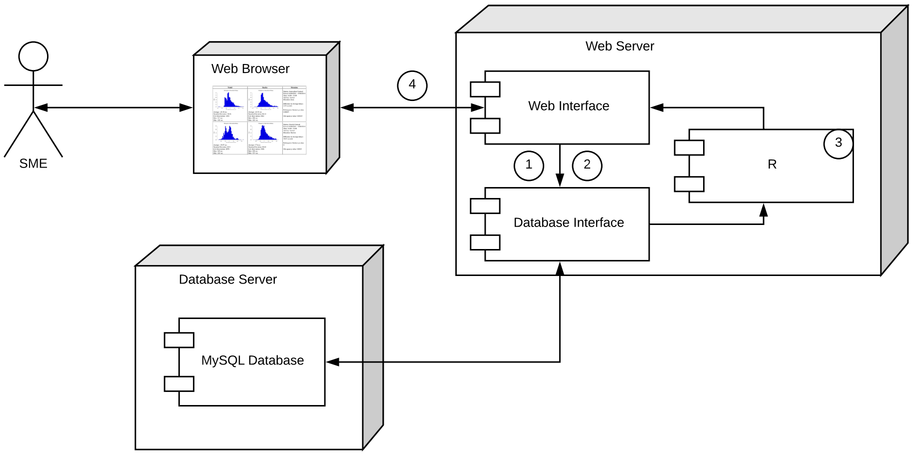


Figure 9.2. Deployment diagram of the implementation.

of choice. In PHP, this is accomplished by using the function *exec*() and executing a shell script. The procedure is almost identical for both SQL and CSV files. In Java, the equivalent for PHP's *exec*() function is also a function called *exec*(), while in Python the equivalent is subprocess *call*.

In a fully commercial implementation, there are several issues someone should consider, like user permissions, name consistency in the database tables, data sanitization etc., but they are out of scope for the purpose of this study; hence they are not analysed further.

Both approaches can be combined with an easy to use UI and they have their advantages and disadvantages. On the one hand, the first approach has the major advantage that it can be sanitised and used safely in the public domain, but at the same time is relatively slow, especially with CSV or SQL files with millions, or even billions, records. On the other hand, the second approach is much faster and should be preferred for large files, nevertheless it is prone to SQL injections making it unsafe for the public domain but suitable for internal use within a company.

9.4.2. DATA CLEANING AND TRANSFORMATION

The proof of concept implemented for this project has a simple UI for cleaning and transforming data and depending on the task at hand, it is a combination of fully and semi-automated cleaning and transformation SQL scripts. In the two case studies examined, five data quality issues were identified. In Table 9.1, these issues are listed along with an example from the data and the level of automation used to address them.

9.4.3. DATA PROCESSING

Every model, depending on its intended purpose, usually requires different statistical techniques and visualizations to enable its validation. The two examples in this chapter

TABLE 9.1. EXAMPLES OF DATA QUALITY PROBLEMS.

Quality Criterion	Definition	Example	Automation
Syntactic Consistency	The uniformity in the syntactic representation of data values that have the same or similar semantics (Pipino et al., 2002).	Inconsistent: friso.train_series='120NB' operational.train_series=120	Fully Automated
Semantic Accuracy	The conformity of a data value v to its real-world value v' that is considered correct (Fox et al., 1994).	Inaccurate: opentrack.distance=31971.90 operational.distance=33879.40 for acceptable difference <1000	Semi-Automated
Mapping Consistency	The uniformity in the key values of data representing the same external instance (Price and Shanks, 2005).	Inconsistent: opentrack.position=31971.9 operational.longitude=5.293365 operational.latitude=51.69003	Fully Automated
Semantic Completeness	The degree to which existing values are included in data relevant to the purpose for which the data is stored (Bovee et al., 2003).	Incomplete: friso.arrival='18:42:21' operational.arrival='00:00:00'	Fully Automated
Presentation Suitability	The degree to which the data format, unit, precision, and type-sufficiency are appropriate for the purpose of data use (Price and Shanks, 2005).	Kilometers VS Meters: opentrack.position=0.935898 operational.position=935.9	Fully Automated

are not the exception. Nevertheless, for validating both models, the same tools (HTML, CSS, JavaScript, PHP, SQL, & R) were used and in some cases in the same way. Particularly for this step, the tools used were JavaScript, PHP, SQL, and R. Below, the common usage of these tools on both models is described.

PHP: Used to dynamically load information, like stations, train series etc., from the database and to trigger the R script with the appropriate arguments using the $exec()$ function. In this example, the first argument of the $exec()$ function is a file that allows R to run as a scripting language, the second argument is the R script to be executed, and the rest of the arguments are the ones passed on the script and utilized within it.

JavaScript: Initially used to dynamically load train series and stations in dropdown menus given previous choices. Then, in one of the interfaces, the library $Leaflet$ was used to create an interactive map of conflicts.

R: Used the feature $args \leftarrow commandsArgs(TRUE)$ to fetch the arguments passed from the $exec()$ function in PHP, and the $RMySQL$ package to allow to run SQL queries from within the R script (R can also directly interact with NoSQL databases, like MongoDB, using the package $rmongodb$, or Cassandra, using the package $RCassandra-package$).

In Section 9.4.3 and Section 9.4.3 the tools developed to validate FRISO and OpenTrack are presented respectively.

FRISO

The model in FRISO was used to test the punctuality of timetables. As a result, the tool for validating the model should offer a statistical comparison of the delays between the model and reality. Moreover, it should offer a way to perform graphical comparison (Balci, 1998) of the delays. The latter is needed by SMEs in order to observe how the delays of specific train series or at specific stations are distributed and perhaps cause indirect delays to other trains.

Therefore, the tool incorporates three main components, all in one interface:

1. histograms depicting the delays of both the simulation model and reality in the middle,

2. descriptive statistics (average delay, standard deviation, min and max delay etc.) below the histograms, and

3. metadata (date, location etc.) and tests to verify the equality of the distributions between model and reality (Kolmogorov-Smirnov test and Pearson Chi-square test) on the right.

In R, the function $hist()$ was used to build the histograms and the functions $ks.test()$ and $chisq.test()$ to calculate the Kolmogorov-Smirnov test and the Pearson Chi-square test respectively (In Python, all these function can be found in the package $SciPy$).

OPENTRACK

The model in OpenTrack was used to test the conflicts in a timetable. As a result, the tool for validating the model should offer a statistical comparison on the frequency of conflicts between the model and reality. Moreover, it should offer a microscopic visualisation of the train driving behavior, which can allow SMEs to pinpoint problematic regions in the infrastructure or the rolling stock.

Therefore, the tool incorporates two main components in different interfaces:

1. a simple UI in which SMEs can identify in a geographical map whether certain conflicts that exist in reality also exist in the model and vice versa, as well as the frequency and the root cause of each conflict.

2. a detailed graph with the driving behavior of the model and reality, including the totality of the realisation data and several percentile lines, the number of which varies depending on the observations.

In R, the library $ggplot$ was used to build the visualisation of the driving behavior (In Python, advanced visualisations can be found in the package $matplotlib$). Additionally, several other libraries were used to fine-tune the graph, like $scales$ that allows to automatically determine breaks and labels, and $directlabels$ that allows to put a label on a line-graph outside the legend and directly next to the line.

9.4.4. PRESENTATION OF RESULTS

For FRISO, a proof of concept of the tool developed is shown in Figure 9.3, in which case PHP was used to fetch the results of the statistical tests, and HTML & CSS to fetch the histograms and present all the outcomes of the analysis in the most appropriate way. For OpenTrack, which had two different interfaces as mentioned above, an example of the first interface using the Javascript library $Leaflet$ for identifying the conflicts in a map, is shown in Figure 9.4, whereas an example of the resulting graph of the second interface is shown in Figure 9.5, in which case all the work for the resulting graph was performed

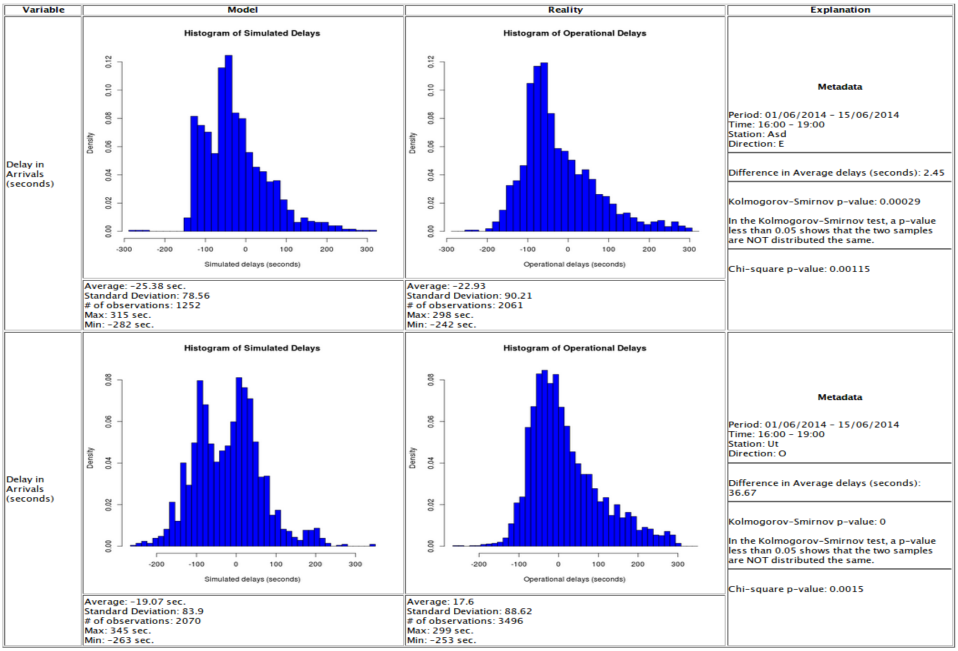


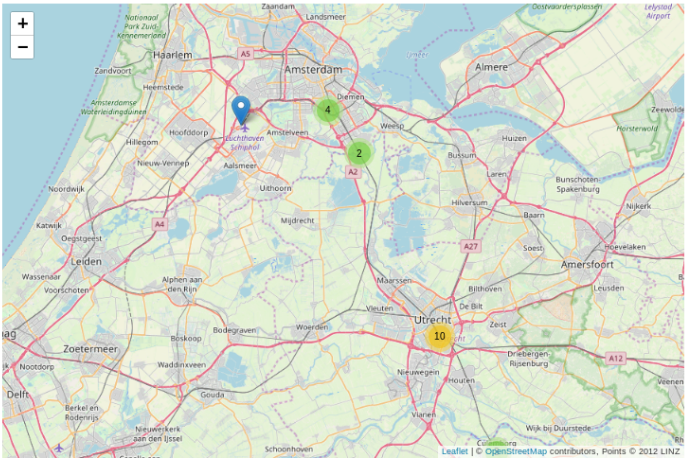
FIGURE 9.3. SIMULATED AND REAL DELAYS IN AMSTERDAM AND UTRECHT CENTRAL STATIONS.

9

in R. For the first interface, PHP, JavaScript, HTML & CSS were used to build the UI, which has simple dropdown menus, and compares the conflicts between the train series from the simulation and operational data.

9.5. FINAL REMARKS ON THE AUTOMATED VALIDATION FRAMEWORK

In this chapter, a framework that combines web technologies and the R statistical language was explored as a mean to mitigate problems pertaining to the validation of simulation models. Web technologies are commonly used in numerous occasions, but there was no indication of them being used in simulation model validation, despite the overwhelming evidence that these technologies can help towards mitigating the risks associated with validation. Indeed, the application of the proposed framework to two case studies showed that web technologies offer a vast toolbox that can help towards developing a more automated validation than the current almost completely manual state. Moreover, the interoperability of those tools further widens the toolbox by enabling more accurate and well-established technologies to be directly implemented within the same environment. The latter is demonstrated in the implementation presented in this chapter, in which case R, a well-established statistical language, was used to perform all the necessary calculations and create elaborate graphs, without the end user even be aware of. Finally, the fact that the end result of such an implementation can be presented in



Delayed Train	Train Causing Delay	Direction	Location	Exists	Frequency
3500	4300	E	AsdS	Realization & OpenTrack	R: 18 | OT: 8
3500	3000	E	Ut	Realization & OpenTrack	R: 62 | OT: 4
3500	3000	E	UtN	Realization & OpenTrack	R: 79 | OT: 8
3500	6900	E	Cl	Realization & OpenTrack	R: 38 | OT: 4
3500	3000	E	Ac	Realization & OpenTrack	R: 12 | OT: 4
3500	9600	E	Ht	Realization	R: 117
3500	1100	E	Ehv	Realization	R: 64
3500	6900	E	Gdm	Realization	R: 32
3500	9600	E	Btl	Realization	R: 18
3500	6900	E	Htn	Realization	R: 15
3500	800	E	Ut	Realization	R: 13
3500	9600	E	Ehv	Realization	R: 13

FIGURE 9.4. CONFLICTS AROUND DIFFERENT STATIONS IN THE NETHERLANDS DEPICTED IN A MAP.

a web browser translates not only to easy accessibility but also to an affordable solution regarding the hardware and software of the end user.

A potential threat to the applicability of the framework is the limited focus of the case studies, which were both from the railway sector, and the limited focus on the simulation packages, which were only two, i.e. FRISO and OpenTrack. Nevertheless, the data used on both case studies did not have domain-specific or package-specific particularities, which translates that the framework is applicable to different domains and simulation packages. By all means, this assessment should not be taken for granted and future work should include the application of the framework to different domains and simulation packages.

Another potential threat arises from the fact that the combination of the particular web tools used in the case studies (PHP, plain JavaScript etc.) and R is not the only way that this system could have been designed. A PHP library for statistics, like *statistics*, coupled with a JavaScript framework for visualisations, like *d3.js*, or an end-to-end solution using Python are also viable and worthy exploring solutions. Nevertheless, this

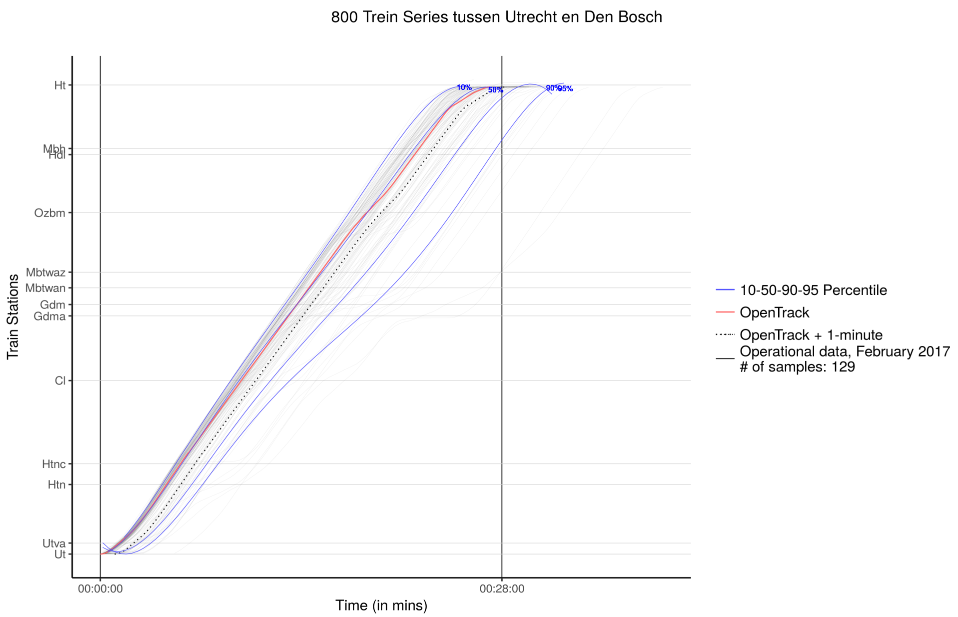


FIGURE 9.5. DRIVING AND BREAKING BEHAVIOR IN OPENTRACK.

choice was made based on what the authors considered to be the *best tool for the job*, which again should not be considered as a universally best solution. Implementations with different tools, like Java or Python, can lead into a comparative analysis on which tools are preferable for a validation study.

Moreover, in the future, an implementation based on a NoSQL database will provide more functionality with semi-structured and unstructured data, which can enhance even further the applicability of the validation tool. Finally, commercial use of the tool would be possible through a full scale implementation, which will take advantages of a modern JavaScript framework (like Angular.js, React.js, Backbone.js etc.) for a fully customisable user interface, Ajax for asynchronous communication with the database, and more optimised SQL (or NoSQL) and shell scripts for importing, cleaning, and transforming data. Using the aforementioned technologies would benefit modellers, validation experts, and SMEs from developments in other domains, e.g. JavaScript, query optimisation etc., hence preventing them from reinventing the wheel and focusing on what is important to them.

10

LESSONS LEARNED FROM THE CASE STUDIES

10.1. INTRODUCTION

A nation-wide railway network is not just the sum of its individual components, but should, instead, be seen as a Complex Adaptive System (CAS) (Rinaldi et al., 2001). Upon analysing such a system, it becomes evident that it consists of multiple socio-technical systems (Trist and Bamforth, 1951) (trains with drivers, infrastructure with maintenance engineers and train traffic controllers etc.) and several independent actors (passengers, politicians etc.), which adapt their behaviour based on internal and external stimuli and form complex and emergent relationships with each other. At the same time, a railway network can also be seen as a System of Systems (SoS) despite the lack of a definitive definition of SoS. It has been shown that railway networks demonstrate the behaviour of what experts describe as SoS. According to De Laurentis' (2005) work, they seem to adequately satisfy the distinguishing traits of SoS: operational & managerial independence, geographic distribution, evolutionary behaviour, emergent behaviour, to name a few.

On the other hand, by definition, simulations are the imitation of the operations of a real-world process or system over time (Banks et al., 1984), and as such they are an abstraction, or simplification, of the respective process or system. Despite their abstractive nature, simulations are perhaps the best way that systems characterised as SoS can be understood and tested in an affordable, risk-free, and ethical way (Zeigler and Sarjoughian, 2012). These three terms, i.e. affordable, risk-free, and ethical, are the *holy grail* of most railway companies. Even a small change in a railway infrastructure can cost several millions. Moreover, it bears significant risks both in terms of construction (e.g. wrong materials, mistakenly positioned switch etc.) and operation (e.g. not alleviating the load on the network, interfere with the normal operations etc.), whereas their miti-

Parts of this chapter have been published in Roungas et al. (2018e).

gation further increases the cost. Finally, since such a system is used by hundred of thousands or even millions of people on a daily basis, the extent to which a railway company exhausts all possible solutions, in order to provide the best possible service, becomes an ethical issue. This is a small example of the complexity of just one decision. But not only changes in the physical infrastructure need extensive testing. Changes in the timetable also need testing, in order to ascertain that the railway resources (infrastructure, rolling stock etc.) successfully accommodate these changes, that any unexpected situation are dealt with in the best possible way, and that the probability the service will become unavailable is minimised.

In effect, the use of simulation in the planning and operations of railways has become increasingly popular. The said popularity has not passed unnoticed by ProRail, which several years ago started building simulations both internally and through the use of third-party packages, and has todate developed a wide range of simulations, extending from microscopic (Yuan and Hansen, 2007) and macroscopic (Middelkoop and Bouwman, 2001) to gaming simulations (Meijer, 2012b, 2015). Microscopic railway simulations simulate every aspect of the system in a detailed manner; the train's motion at any given moment is determined according to dynamic equations and every aspect of the infrastructure is taken into account (Asuka and Komaya, 1996). On the other hand, macroscopic railway simulations simulate only the arrival and departure times of trains, and the general characteristics of the infrastructure (Asuka and Komaya, 1996). Finally, gaming simulations can be either microscopic or macroscopic and what differentiates them is the human input. Each of these simulations has a different purpose. Whilst macroscopic simulations are time efficient, microscopic simulations are more precise and thus preferred for networks with high speed trains and high density of train traffic Asuka and Komaya (1996). Gaming simulations are used when human input is necessary.

Several of these simulation packages are quite similar both in terms of input and output data, and in terms of their intended purpose. An example of such similarity is FRISO and OpenTrack, which the authors were assigned to validate, and thus form the two case studies examined in the present chapter. FRISO is ProRail's in-house simulation environment (Middelkoop and Loeve, 2006) whereas OpenTrack is a well-established program developed at the Swiss Federal Institute of Technology's Institute for Transportation Planning and Systems (ETH IVT) (Nash and Huerlimann, 2004). Since they are both microscopic simulation environments, FRISO and OpenTrack have the potential to, and depending on the model usually do, simulate the railway network in a detailed manner; both have the ability to depict the network down to a switch level. In view of the pointed similarities, a comparison of one simulation package over the other seems rather inevitable. In this respect, the authors' initial hypothesis is that let alone their similarities, FRISO and OpenTrack are different, thus suitable for different usages. This is a hypothesis that will be accepted or rejected, once the comparison of the two software packages has been concluded.

In this study, two different instantiations of models on both packages, which led in the development of customised tools for their validation, are presented. Through the analysis and the comparison of the two models, and the development of the respective tools for the ensuing validation, this study aims at identifying critical factors that influence the success of simulation models. Therefore, the intention of the comparison is not to decide which simulation package is more valid but to demonstrate the steps that were followed during their validation. Particularly, the comparison of the two packages aims at pinpointing the common practices during the validation of a simulation model. Subsequently, the analysis of these common practices would be of great interest, since they could be considered good candidates for factors that can critically influence the validation study of a simulation model and, as a result, the model itself.

An important distinction should be made between the validation methodologies applied in each one of FRISO's and OpenTrack's models, and the methodology used to accomplish the aim of this study. The former are two methodologies for validating punctuality and the train driving behaviour, respectively. The latter is built on the steps of this chapter, which by definition also includes the two aforementioned methodologies. The purpose of this study is to identify the similarities and differences between the two packages and their subsequent models, pinpoint the impediments during the validation study, and in turn, present the most striking results and the identified critical success factors.

In Section 10.2, FRISO's and OpenTrack's architecture is described. In Section 10.3 and Section 10.4 the detailed models for FRISO and OpenTrack are presented along with their respective validation studies. In Section 10.5, the results of both models are compared to each other and a conclusion is drawn in regard to the initial aim of this chapter. Finally, in Section 10.6, future steps that can improve the research on this field are outlined and final remarks are made.

10.2. SIMULATIONS' ARCHITECTURE

In this section, the architecture of both simulation packages is described. The authors did not have access to any of the packages' source code nor to the internal structure of the models; they were only assigned to validate the models. Therefore, the architecture analysis in both cases is limited to the input the simulations require and the output they produce.

The overall scheme of both packages seems to have a remarkable resemblance. Arguably, this is a product of both packages being microscopic simulations, thus requiring more or less the same input and producing the same output. As such, both FRISO and OpenTrack require three basic components as input:

- Timetable, which for an experimental study can be an hourly pattern timetable whereas for a more operationally oriented study it should be as close as possible to the actual timetable.
- Rolling Stock, which includes all the different trains along with their technical characteristics.
- Infrastructure, which for a microscopic simulation means not just the railway tracks but also switches, signals, and any other detail that influences train operations.

The visualisation features of both packages are quite similar in that they both offer animation and interactive capabilities. Finally, the output is equally similar for both, with OpenTrack having a few more features when it comes to graphs and tables. Particularly, OpenTrack offers two additional options: a. rail occupation statistics followed by occupation diagrams, and b. a train power and energy consumption output. A depiction of OpenTrack's main elements is shown in Figure 10.1. What differentiates FRISO from OpenTrack, and it is the main reason ProRail built it in the first place, is its ability to be incorporated in a High Level Architecture (HLA) scheme (Kuhl et al., 1999), which allows the interaction with other computer simulations regardless of the computing platforms, and provides for a wide application in gaming simulations (Middelkoop et al., 2012).

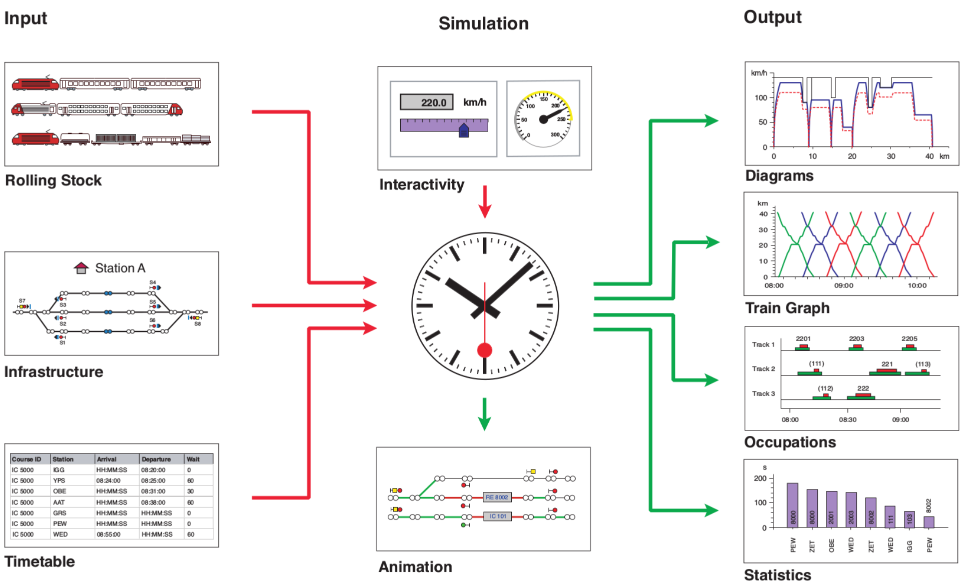


FIGURE 10.1. MAIN ELEMENTS OF OPENTRACK (NASH AND HUERLIMANN, 2004).

In the next two sections, the input, output, and the validation study of two models instantiated in FRISO and OpenTrack are described.

10.3. THE FRISO MODEL

In this section, the model used in FRISO (hereinafter references to FRISO and OpenTrack will be references to their respective models, unless otherwise stated) and its subsequent validation is described in detail. The model was built in 2014 and it simulates the train operations in one of the most heavily utilized sections (Amsterdam Centraal - Utrecht Centraal) of one of the largest corridors in the Netherlands (A2), during the whole month of June 2013. The intended use of the model was to examine the punctuality of the timetable with the particular focus being the Amsterdam and Utrecht central stations. The input elements of the model were the:

- Timetable, which was the theoretical timetable for the month of June 2013. The term theoretical is used in order to denote that the actual timetable slightly changes every day, so as to accommodate urgent and unplanned events, e.g. an unplanned freight train, heavy weather conditions etc.

- Rolling Stock, which was the exact rolling stock operating between Amsterdam and Utrecht central stations in June 2013. This included the major train series 120 (Operator: DB, Type: ICE), 3000 (Operator: NS, Type: Intercity), 4000 (Operator: NS, Type: Sprinter), 7400 (Operator: NS, Type: Sprinter), 800 (Operator: NS, Type: Intercity), and the minor train series 47700, 48700, 77400.

- Infrastructure, which was the exact infrastructure (tracks, traffic signals, switches etc.) between Amsterdam and Utrecht central stations in June 2013.

Moreover, several variations of the Gamma, Normal, and Negative Exponential distributions were used for the delays in arrivals.

The analysis of the output of the model required extensive data cleaning for both the model output data and the operational data. In more detail, the data cleaning comprised of queries that:

- Deleted several columns from both datasets that were not relevant for the study (like columns with specific codes used in planning),

- Deleted rows containing regions and train series that were not common on both datasets (ensuring that both datasets were of the exact same region and containing the same train series), and

- Deleted from the realization data all rows with unrecorded arrival or departure time.

The data cleaning resulted in a reduction of the number of variables (i.e. columns) for both datasets and of their overall size down to half, which significantly reduced the execution times of the queries. Finally, renaming certain columns became necessary, in order to allow the customized validation tool to automatically calculate the various statistical tests and produce the necessary graphs.

As a whole, the model shows remarkable precision by being only two seconds off in estimating the average delay in arrivals. Even at a station level, for all stations, the difference in the average arrival delay between the model and reality is less than 30 seconds (Table 10.1). Despite this rigour in a macroscopic and in a station level, there are three striking observations:

Train	All	120	3000	4000	47700	48700	7400	77400	800
Delay	-2.04	19.82	24.53	-20.66	-99.89	-101.98	-9.73	-65.31	27.4

Station	All	Ut	Mas	Dvd	Asa	Asdm	Asdma	Asd
Delay	-2.04	17.05	20.97	-19.99	9.94	1.9	-1.3	-3.34

TABLE 10.1. DIFFERENCE IN DELAYS IN ARRIVALS AT A TRAIN AND AT A STATION LEVEL BETWEEN FRISO AND REALITY.
ABBREVIATIONS: UT: UTRECHT CENTRAAL, MAS: MAARSSEN, DVD: DUIVENDRECHT, ASA: AMSTERDAM AMSTEL, ASDM: AMSTERDAM MUIDERPOORT, ASDMA: AMSTERDAM MUIDERPOORT AANSLUITING (PASSING POINT), ASD: AMSTERDAM CENTRAAL

10

1. At a train level, the behaviour of the model does not appear to be consistent. The major train series (120, 3000, 4000, 7400, and 800) exhibit relatively good behaviour (difference between reality and model is less then 30 seconds), whereas the minor train series (47700 and 48700) seem to experience big delays throughout the whole route from Amsterdam to Utrecht and vice versa. This is due to the minor series accounting for less than 5% of the total traffic in this particular route, resulting in modellers focusing mainly on the major train series. The train series 77400 is not an actual independent series but the 7400 series with an addition of a 7 in front, in order to indicate that the train performs shunting movements, i.e. parking or sorting the rolling stock. Therefore, given the significant influence of the major train series in the model, and the rather insignificant impact of the minor train series, any negative effect resulting from the latter is diluted.

2. Due to the nature of FRISO's output data, i.e., the availability of the arrival and departure times alone, the only visualization that could show the richness of the data is a histogram of delays, either at a station (Table 10.2) or at a train level. The histograms were built by creating 20-second intervals (bins) of the delays occurred in the model and in reality. The large number of observations gave histograms a fine granularity and allowed for concrete conclusions. Hence, the second striking observation is how delays in arrivals are distributed. Both in Amsterdam and in Utrecht central stations, the operational delays seem to follow a right skewed distribution, while in the model, the delays seem to follow a bimodal distribution, although slightly less sharp for Amsterdam central station. However, since the positive delays (i.e., delays greater than zero), for both cities - and especially in Amsterdam - appear to be distributed similarly in the model and in reality, experts were confident that this discrepancy was not enough to invalidate the model.

3. The third striking observation is that the more a train drives away from Amsterdam central station (Asd), which is the epicentre of the model, the more the precision of the model decreases; and as the train reaches Utrecht central station (Ut), which is the second most important station in the model, the precision of the model slightly rises again. This observation can only be valid if the negative delays (early arrivals) are disregarded. Unless an early arrival causes an indirect delay to another train, which was not observed in this particular case, then it does not cause any negative effect to the system, thus it can be disregarded. Table 10.3 serves as a heatmap, in which the difference in delays between reality and the model are marked as follows:

   - Under 15 seconds with green.
   - Between 15 and 30 seconds with orange.
   - More than 30 seconds with red.

   The stations' sequence in the table is the same a train follows from Amsterdam to Utrecht and vice versa. The colour distinction used in Table 10.3 is meant to help observers identify patterns in the delays. Indeed, upon carefully examining Table 10.3, it becomes easier to notice that the model experiences a ripple effect.

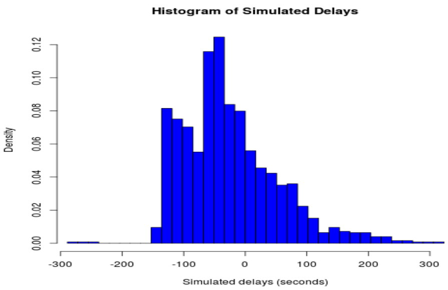
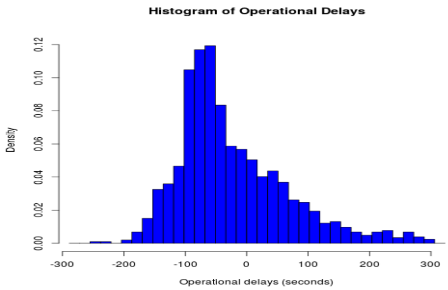
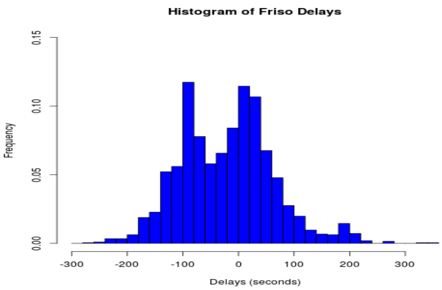
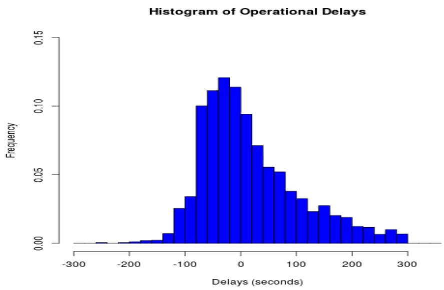
Model	Reality
Amsterdam	
	
Average: -25.38 sec. Standard Deviation: 78.56 # of observations: 1252 Max: 315 sec. Min: -282 sec.	Average: -22.93 sec. Standard Deviation: 90.21 # of observations: 2061 Max: 298 sec. Min: -242 sec.
Utrecht	
	
Average: -19.07 sec. Standard Deviation: 83.9 # of observations: 2070 Max: 345 sec. Min: -263 sec.	Average: 17.6 sec. Standard Deviation: 88.62 # of observations: 3496 Max: 299 sec. Min: -253 sec.

TABLE 10.2. SIMULATED AND REAL DELAYS IN AMSTERDAM AND UTRECHT CENTRAL STATION.

The more a train diverges from the main area of focus in the model, i.e., primarily Amsterdam central station and secondarily Utrecht central station, the less the factors that influence the predictability of the model are taken into account and calculated with precision. Nevertheless, similarly to the minor train series, experts concluded that, due to the fact that those cities are outside the main area of focus, their overall impact in the model is insignificant. Hence, this observation was also not enough to invalidate the model.

In this section, FRISO, and its subsequent validation, were described and the most

		Stations						
		Ut	**Mas**	**Dvd**	**Asa**	**Asdm**	**Asdma**	**Asd**
Trains	**120**	10.89	24.66	40.86	22.83	15.16	4.31	4.24
	3000	15.12	47.6	38.9	9.24	0	-1.27	-1.21
	4000	-	-	-32.32	-2.43	-0.79	-11.16	-14.73
	47700	29.36	-106.31	-121.72	-52.82	-42.23	-52.15	15.64
	48700	3.16	-209.71	-66.05	-55.83	-41.99	-58.06	-0.95
	7400	8.14	-3.9	-24.22	-0.8	-38.1	-11.26	-14.94
	77400	-	-	16.6	36.2	17.21	-8.86	-22.78
	800	34.57	44.31	40.43	18.17	11.25	9.19	10.64

TABLE 10.3. HEATMAP WITH DIFFERENCE IN DELAYS IN ARRIVALS BETWEEN FRISO AND REALITY. **ABBREVIATIONS:** SAME AS IN TABLE 10.1

striking observations were presented. All these striking observations are deemed by experts to constitute insufficient evidence for invalidating the model because:

1. The train series accounting for more than 95% of the traffic were modelled with high precision.

2. Disregarding the early arrivals in Friso, which did not appear to cause direct or indirect delays to other trains, resulted in transforming the bimodal distribution of delays in a right skewed distribution very similar to the one of the operational data.

3. Any stations other than Amsterdam and Utrecht central stations were not the focus in this model, hence any inconsistency between the Friso and operational data in these stations was not taken strongly into account.

As a result, experts considered the model to be valid for its intended purpose, which is to test the punctuality in major train stations like the ones in Amsterdam and Utrecht. In addition to the three aforementioned reasons, the model exhibits striking resilience especially in the two stations of focus (Amsterdam and Utrecht). This is evident from the fact that despite the relatively large differences in the off-focus stations, the model adapts and covers a significant amount of these differences, either by accelerating or decelerating. A more in depth analysis and a comparison of FRISO with OpenTrack, which is described in Section 10.4, takes place in Section 10.5.

10.4. THE OPENTRACK MODEL

In this section, OpenTrack and its subsequent validation is described in detail. The model was built in 2016 and it simulates the train operations in a heavily utilized section (Eindhoven Centraal - Utrecht Centraal) of one of the largest corridors in the Netherlands (A2), according to the newly designed timetable, which was to start in January 2017. The input elements of the model were the:

- Timetable, which was the newly developed timetable intended to be put in use in January 2017.
- Rolling Stock, which was part of the rolling stock operating between Eindhoven and Utrecht central stations according to the new timetable. This included four train series: 800 (Operator: NS, Type: Intercity), 3500 (Operator: NS, Type: Intercity), 6000 (Operator: NS, Type: Sprinter), and 9600 (Operator: NS, Type: Sprinter).
- Infrastructure, which was the planned infrastructure (tracks, traffic signals, switches etc.) between Eindhoven and Utrecht central stations for January 2017.

Similarly to FRISO, the analysis of OpenTrack's output required data cleaning for both the model and the operational data. But unlike FRISO, with OpenTrack the data cleaning was more about transformation of the data in comparable units (model) and correction or deletion of GPS data that were either distorted or off certain limits (reality).

The initial dataset included four train series running through one of the major corridors of the Netherlands (A2), namely from Utrecht to Eindhoven. Unlike FRISO, in which the intention of the study was to test punctuality, with OpenTrack the purpose was to test the conflicts occurring throughout the timetable by examining the train driving behaviour. Despite this intention, due to the nature of its data, OpenTrack could also provide for a punctuality test by extracting the arrival time of trains. Hence, OpenTrack's validation study was divided in two independent studies: validation of the driving behaviour and validation of the punctuality. Since the models in Friso and OpenTrack depict different instantiations of the railway system, it could be argued that they cannot be compared. Ideally, there should have been a model for each simulation package simulating the exact same scenario, thus allowing the two models to be directly and indisputably compared. Nevertheless, in a commercial setting this is hardly ever possible due to time and budget restrictions. Companies cannot usually afford such additional costs and, further, do not have the luxury of time to build and run experiments of the same scenario on multiple platforms for testing purposes. This results in situations like the one described in this chapter, where a comparison is performed between the different models by incorporating experts' knowledge about the system. In other words, a comparison in such a situation can become fruitful when along with the results of each model, experts provide insights based on their experience that provide context, which in turn mitigates the risk arising from the different instantiations.

10

10.4.1. DRIVING BEHAVIOUR

The driving behaviour modelled in OpenTrack depends on five parameters: the acceleration rate, the minimum speed, the maximum speed, the braking (deceleration) rate, and a performance coefficient, and it should be separated into two different categories, namely the actual driving behaviour and the braking behaviour. The driving behaviour is determined by the acceleration rate, and the minimum and the maximum speeds after all of these have been multiplied for adjustment by the performance coefficient. The braking behaviour is only determined by the braking rate. The acceleration rate, the minimum speed, the maximum speed, and the braking rate are predetermined by the Dutch railway operator ProRail and they are fixed. Therefore, any variation on the driving and braking behaviour depends on the performance coefficient. The performance coefficient is determined by the modellers through trial and error based on observation from past operational data, in order for the train behaviour to be as realistic as possible. For this particular model, the performance coefficient fluctuates from 97.5% to 100.5% depending on the train type, i.e., fast train (intercity) or slow train (sprinter).

A visualization of the driving and braking behaviour is shown in Figure 10.5. The graphs were construed in order to, initially, have all the available information visualized, allowing for the deduction of the necessary information afterwards. The x-axis shows time in minutes and the y-axis shows the starting and stopping station, as well as all intermediate stations trains pass through without stopping. The graphs include four different kinds of line graphs:

- All operational data from February 2017 depicted in a low opacity black colour line, which create a grey shadow that can reveal patterns.

- The three or four percentile lines (10^{th}, 50^{th} (median), 90^{th}, 95^{th} percentile) depending on the number of observations, depicted in a blue colour line. The 95^{th} percentile line is shown only in cases where the sample size of the operational data is more than 100. The percentiles were calculated by creating 10-second intervals (0 to 10 sec., 10 to 20 sec. etc.). Then, all available data points from the GPS data that belonged at each interval were gathered and the position of all trains within each interval was linearly extrapolated to the floor (e.g. getting the position of a train 11 seconds after starting would mean that its position would be linearly extrapolated at 10 seconds, getting the position of a train 23 seconds after starting would mean that its position would be linearly extrapolated at 20 seconds and so forth). Finally, the list of data points was sorted in an ascending order, and the 10^{th}, 50^{th}, 90^{th}, and 95^{th} percentiles were chosen.

- OpenTrack's data obtained in March 2016 depicted in a red colour line.

- OpenTrack's data obtained in March 2016 increased by one minute is depicted in a dotted black colour line. The use of this line aimed at planning and visualizing potential conflicts with trains based on the minimum running time (minimum difference between two consecutive trains that are on the same track) allowed by ProRail.

The braking behaviour of the model, as shown in all three subfigures in Figure 10.5, is modelled with extreme precision when compared to the median (50^{th} percentile). On the other hand, regarding the actual driving behaviour, in some cases (Figure 10.2) the train drives almost on the exact path of the median (50^{th} percentile), which is the desired driving behaviour, but on some other occasions it either drives along the 90^{th} percentile (Figure 10.3), which is the slowest 10%, or exhibits some sort of irrational behaviour (Figure 10.4), which upon further examination is caused due to a conflict with another train.

As a result, this fluctuation in the driving behaviour brings up the following question: To what extent can a single driving profile adjusted only by a coefficient simulate a realistic driving behaviour? In order to answer this question, the purpose of the simulation study, and consequently the extent to which this model is valid, should be taken into account. This particular simulation study was intended to test the newly developed timetable against conflicts between trains rather than the driving behaviour per se. Modelling the driving behaviour was a means to an end.

The conflict identified in Figure 10.4 was observed in 11.1% of the cases in the model and approximately in 10-15% of the cases in reality, depending on the train series. By all means, not every conflict resulted in exactly the same behaviour, neither in reality nor in the model. Nevertheless, the model managed not only to anticipate the possibility of conflicts but also to approximate the probability of the occurrence of these conflicts.

10.4.2. PUNCTUALITY

The richness of OpenTrack's output also provided for a punctuality test, similar to the one performed on FRISO, which allows for a direct comparison of the two models. Unlike FRISO, in which the focus of the punctuality test revolved around the central stations of Amsterdam and Utrecht, the focus in OpenTrack's punctuality test was in the central stations of Eindhoven and Utrecht. OpenTrack provided an output file with all the delays, and the operational delays were calculated by extracting the last value of the GPS data from each sample.

The histograms of delays from the model and reality are shown in Table 10.4 for Eindhoven and Utrecht central stations. While the average difference in the delays between the model and reality for both cities is approximately 10 seconds, which indicates a good estimation of punctuality, the delays are distributed completely differently between the model and reality.

10.4.3. CONCLUSION ON OPENTRACK

This section focused primarily on OpenTrack and, subsequently, on its validation parameters. The validity of the model was tested with regards to the train driving behaviour and the punctuality of the contemplated timetable. Experts considered the model to be valid for the purpose of simulating the driving behaviour and for identifying, in turn, conflicts between trains. On the contrary, the difference in the distributions of delays has been so significant that the model cannot be considered valid for a punctuality study at a microscopic level.

10.5. DISCUSSION

The analysis in Section 10.3 and Section 10.4 showed the advantages and disadvantages of FRISO and OpenTrack, which were both deemed by experts to be valid for their intended purpose. As mentioned in Section 10.1, a comparison between similar simulation packages and their subsequent models is inevitable, but in this particular case, doing so would be a mistake. Comparing FRISO with OpenTrack would not be fruitful due to multiple reasons. First and foremost, despite both being microscopic simulation packages, the intended purpose of each model is different (FRISO: punctuality, OpenTrack: conflict detection).

Moreover, the datasets used in the models and the parameters tweaked within the simulation software have significant differences. The dataset used in FRISO has only the arrival and departure times of trains, whereas in OpenTrack, the exact location of

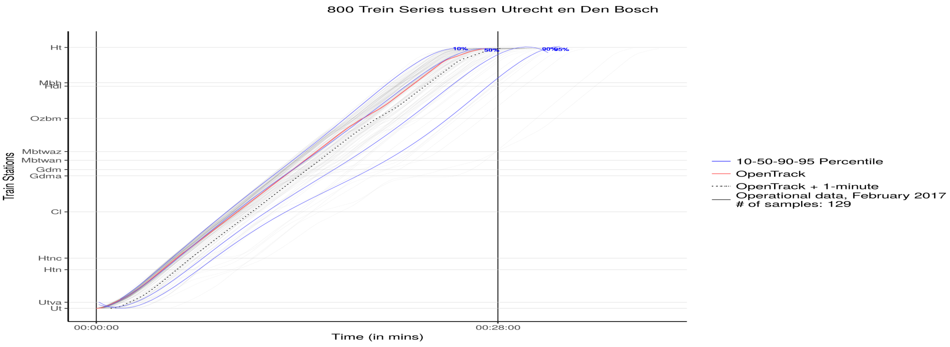


FIGURE 10.2. TRAIN SERIES 800 BETWEEN UTRECHT CENTRAL STATION AND DEN BOSCH

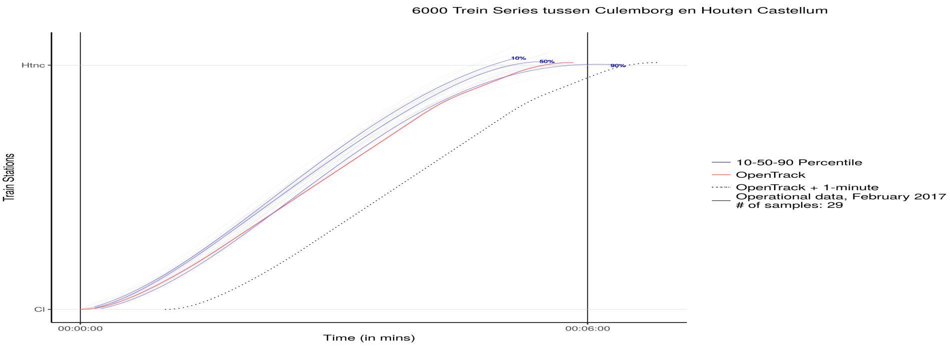


FIGURE 10.3. TRAIN SERIES 6000 BETWEEN CULEMBORG AND HOUTEN CASTELLUM

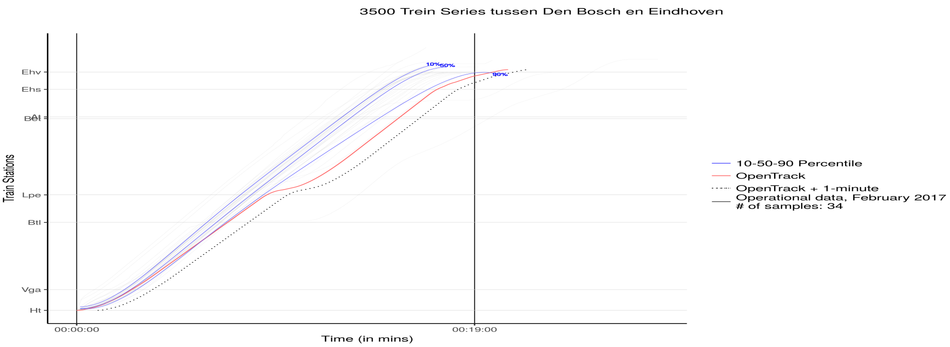
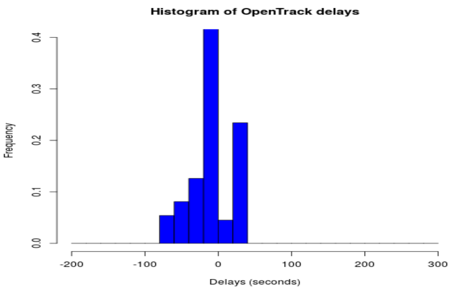
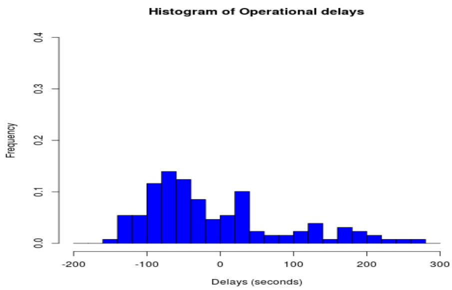
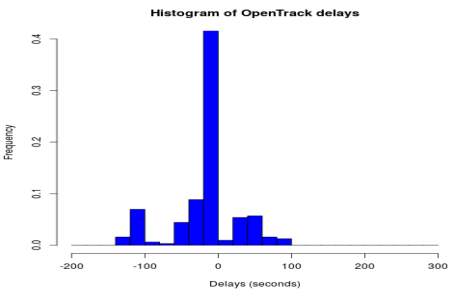
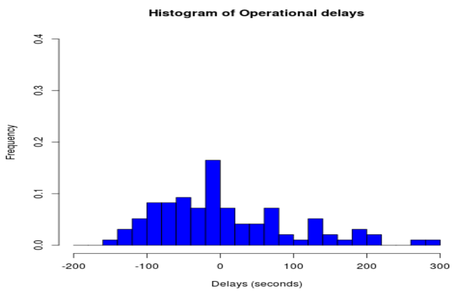


FIGURE 10.4. TRAIN SERIES 3500 BETWEEN DEN BOSCH AND EINDHOVEN

FIGURE 10.5. DRIVING AND BRAKING BEHAVIOUR OF OPENTRACK.

TABLE 10.4. SIMULATED AND REAL DELAYS IN EINDHOVEN AND UTRECHT CENTRAL STATION.

Model	Reality
Eindhoven	
 Average: -3.73 sec. Standard Deviation: 27.39 # of observations: 111 Max: 36 sec. Min: -68 sec.	 Average: -12.85 sec. Standard Deviation: 94.71 # of observations: 129 Max: 262 sec. Min: -159 sec.
Utrecht	
 Average: -8.74 sec. Standard Deviation: 41.58 # of observations: 316 Max: 96 sec. Min: -136 sec.	 Average: 2.85 sec. Standard Deviation: 93.99 # of observations: 97 Max: 282 sec. Min: -149 sec.

each train is available every two seconds. Additionally, FRISO was built based on the existing infrastructure, rolling stock, and timetable, for which delays were known. On the contrary, OpenTrack was built based on the existing infrastructure and rolling stock, but based on a future timetable for which delays were not known and could only be assumed according to the modellers' knowledge. Finally, the two datasets used in the models were almost four years apart, focused on different cities (with a small overlap in

Utrecht), and were based completely different timetables.

Therefore, there are several critical factors, identified in this study, for a simulation model to be successful, both in a conceptual or design level as well as in a practical or analytic level. With regards to the conceptual or design level, these critical factors are the following:

- Whereas it is very difficult and perhaps of no use for all stakeholders to know in detail how the model works, they should have an understanding of the intended purpose of the model, in order to build it, validate it, and eventually use it effectively. Public transportation in general, including the railway sector in particular, is a multi-disciplinary field, and as such it should be assumed that not all involved stakeholders have similar background, either professional or educational. Hence, information should be carefully and appropriately disseminated among the different stakeholders (Balci, 1990).

- Whereas it is common knowledge (not only among validation experts) that there is no such thing as *absolute validity* (Martis, 2006), it is often overlooked, even by experts. In other words, imperfect models can still be valid. An example of such case was demonstrated in Section 10.3, where FRISO showed a few discrepancies between the model and reality but it was nevertheless considered valid by experts for its intended purpose.

- Similar to the previous point, the degree to which the model deviates from reality but still remains within acceptable limits is neither certain nor the same for all models (Balci, 2004). In other words, a discrepancy observed in one model might not invalidate it, whereas, if it is observed in another model it might do. An example of such case was demonstrated in both Section 10.3 and Section 10.4, where in the former the differences on how delays were distributed did not invalidate FRISO, while in the latter, OpenTrack was invalidated.

- One model can be valid for one purpose but invalid for another (Balci, 1990). An example of such case was demonstrated in Section 10.4, where OpenTrack was deemed valid for conflict detection but invalid for testing the punctuality of the timetable. If OpenTrack had been used only to test the punctuality of the timetable, it would have been invalidated and thrown away, resulting in some sort of a Type II error (false negative).

- Similar to the previous point, the methodological approach on validation changes in accordance with the purpose of the model (Balci, 2003). An example of such case was demonstrated in Section 10.4, where one methodology was used to validate the driving behaviour and the conflict detection of the model, as opposed to a different methodology used to validate the model's ability to accurately assess the punctuality of the timetable.

With regards to the practical or analytic level of simulation models, and specifically the validation of simulation models, these critical factors are the following:

- In public transportation, the validation of a simulation model can focus on a geographical level (for the railway sector that means *stations*), an example of which is shown in Table 10.2, on a vehicle level (for the railway sector that means *trains*), an example of which is shown in Figure 10.5, or on a mixture of those two, an example of which is shown in Table 10.3. Therefore, the appropriate tools should be used to validate a model depending on its focus.

- Given the problem at hand, e.g. testing a new timetable or assessing the punctuality of the current timetable etc., companies and researchers should carefully craft the methodology that would result in a successful validation. This methodology includes the selection of the most appropriate validation methods (Roungas et al., 2017) and the acceptability criteria, as well as the requisite tools for the implementation of the methods in reference and the presentation of the validation results to the experts, who are, in turn, the ones to assess the validity of a model or models.

- During the data analysis stage, one should prefer to initially plot all available information - e.g. Figure 10.5, including all the operational samples in a graph; depending on the subsequent needs the information may be reduced and more simplistic graphs or tables will be created incorporating only the necessary data under examination. The other way around bears a significant risk of neglecting pieces of information in the beginning, which would later on be proven crucial.

10.6. FINAL REMARKS ON THE LESSONS LEARNED

Simulations are sine qua non in the planning and operations of the railway sector and their success, or failure, depends on several critical factors. In this chapter, two simulation models, which were instantiated in two different simulation packages, and their respective validation studies, were presented. While the models per se might not be of great interest to the transportation community as a whole, the lessons learned from their validation certainly are. Some of these lessons are old, yet still applicable, and some of them are new. The lack of absolute validity - even if old - is always applicable and yet neglected many times, resulting in an endless pursuit of the *perfect* model. On the other hand, the visualisation of multivariate datasets is an emerging field in which studies often have some new insight to offer. Moreover, what is also new are the methodologies needed to analyse systems with multi-disciplinary nature. These methodologies, along with the system at hand, evolve and adapt, or should at least do so. In turn, the role of simulation models is to give life to those methodologies in an affordable, risk-free, and ethical way.

As a conclusion, the critical success factors identified in this study can be summarised as follows:

- Information and knowledge dissemination should be conducted by taking into account the diversity of the stakeholders, who are involved in such complex systems. As a result, companies should either develop or adjust, as the case may be, their knowledge management protocols, so as to accommodate the difference in the educational and professional background of all the involved actors.

- The validation of a simulation model is heavily determined by its intended purpose. Hence, the intended purpose of the model dictates the methodology that has to be followed, in order for the validation study to be fruitful and avoid Type I & II errors.

- The analysis and the visualisation of data should initially include all the available information, which analysts can then reduce, abstract, or simplify in view of fulfilling their respective goals.

III

Game Sessions

Don't talk to me about rules, dear.
Wherever I stay I make the goddamn rules.

Maria Callas

11

DEBRIEFING

THIS part of the thesis dives into the intricacies of game sessions and debriefing. Particularly this chapter establishes what debriefing is, why it is important in games, and what debriefing methods exist.

11.1. INTRODUCTION

To reflect is to look back over what has been done so as to extract the net meanings which are the capital stock for intelligent dealing with further experiences Dewey (1986), and debriefing does exactly that. It is a facilitated reflection in the cycle of experiential learning Fanning and Gaba (2007).

Historically, debriefing originated in the military (Pearson and Smith, 1985) and it is generally described as a process to elicit information pertaining to an experienced event, in order to gain a better understanding of it (Seymour, 2012). Debriefing has been identified to be the most important feature of games (Bond et al., 2007, Cantrell, 2008, Crookall, 2010, Issenberg et al., 2005, Shinnick et al., 2011) and it is a competency that simulation educators need to master (McCausland et al., 2004, Nehring and Lashley, 2004). Debriefing aims at linking the content presented during the game session with reality and thus, it is where learning occurs (Barreteau et al., 2007). Despite this fact, there are a limited amount of articles on how to debrief, how to learn to debrief, what methods of debriefing exist, and how effective these methods are at achieving learning objectives and goals (Dufrene and Young, 2014). As a result of this lack of research, several factors that inhibit debriefing remain either unknown or disregarded, or are poorly addressed. Section 11.2 serves as an introduction on debriefing.

11.2. DEBRIEFING METHODS

In this section, a brief introduction on debriefing is given. Then, several models and types of debriefing as well as factors and aspects that influence debriefing are presented.

Parts of this chapter have been published in Roungas et al. (2018a).

The purpose of this chapter is to provide a good understanding of the multiple dimensions of debriefing. The reason for that is for the reader to be able to comprehend how a factor inhibits debriefing and how a solution can potentially solve this problem.

Pearson and Smith (1985) distinguish two main forms of debriefing; Formal and Informal. *Formal* debriefing is that which occurs in the presence of a group leader and is deliberately structured in some manners. The group leader should possess skills in structuring, organising, group processes, communication, conflict resolution and counselling. *Informal* debriefing is that which occurs individually or with others after formal debriefing. This part is primarily concerned with formal debriefing.

Debriefing is an important aspect of games, which should be taken into account during the requirement elicitation phase Easterbrook (1991). The first decision regarding debriefing is whether a game session actually requires debriefing. Thiagarajan (1992) tackled this issue by posing two questions:

1. Do participants lack a sense of closure? and

2. Can we derive useful insights through a discussion of the experience?

Answering both questions positively means that debriefing is required and therefore increases the probability that it will have a positive effect.

The next step is to identify the structural elements involved in the debriefing, which can enable the characterisation of the debriefing process. (Raemer et al., 2011) posed five questions, aiming at characterising the debriefing process. Namely, these questions are:

1. Who. Who is debriefing? Raemer et al. (2011) mean who is the debriefer, but this question may also include the participants to debrief (Lederman, 1992) as well as any other stakeholder, who has a direct or indirect interest on the outcome of the debriefing, like the facilitator(s) of the G&S or decision-makers.

2. What. What is the content of debriefing? This includes the experience or the simulation scenario; which debriefing is about, the impact of the experience, the recollection upon the experience, and potentially the report that might be necessary to communicate the results of the debriefing (Lederman, 1992).

3. When. What is the timing of the debriefing? In other words, when does the debriefing take place, especially with regards to the time the G&S session took place. In Chapter 12 and Chapter 13, it is shown how the choice of when to debrief can influence the outcome of the debriefing.

4. Where. What is the environment of the debriefing? This question is referred to both the geographical location and the overall physical setup of the room(s), where the debriefing takes place. As it is shown later in Chapter 12 and Chapter 13, the environment plays a significant role on the outcome of the debriefing.

5. Why. Which is the theoretical framework supporting the debriefing? In other words, is there a debriefing protocol? The usage of a protocol while debriefing has several advantages and disadvantages; which was analysed in Chapter 12.

According to both Raemer et al. (2011) and Lederman (1992), the debriefer is a key element of the debriefing, but the degree to which he or she is involved differs. Fanning and Gaba (2007) stated that, during the debriefing process, the degree to which the debriefer is involved can depend on a variety of generic factors, such us the objective of the experiential exercise, the complexity of the scenarios, the experience level of the participants as individuals or a team, the familiarity of the participants with the simulation environment, the time available for the session, the role of simulations in the overall curriculum, and the individual personalities and relationships between the participants.

Dismukes and Smith (2000) identified three levels of involvement on behalf of the debriefer; High, where the participants largely debrief themselves, Intermediate, where an increased level of instructor involvement may be useful, and Low, where an intensive level of instructor involvement may be necessary.

Several researchers have tried to come up with a debriefing model. Below, five models with high impact are analysed.

THATCHER AND ROBINSON (1985)

According to Thatcher and Robinson (1985), debriefing should enable participants to:

1. Identify the impact of the experience.
2. Identify and consider the processes; which were developed.
3. Clarify the facts, concepts, and principles.
4. Identify the ways, in which emotion was involved.
5. Identify the different views, each of the participants formed.

Thatcher and Robinson (1985)'s model is considered to be one of the first approaches towards formalising debriefing, and as such the model's focus is more on the outcome of the debriefing, with continuous consideration of the importance of experience-based learning.

LEDERMAN (1991)

According to Lederman (1991), the debriefing process should consist of three phases; 1. The introduction to the systematic reflection and analysis, 2. The intensification and personalisation of the analysis of the experience, and 3. The generalization and application of the experience. Lederman (1991) considers communication between stakeholders to be fundamental, in order for the debriefing to have a positive outcome. As such, the model is more focused on the exploration of the emotional state of the participants.

PETRANEK (1994)

According to Petranek (1994), the debriefing process should include Events, Emotions, Empathy, Explanations and analysis, Everyday applicability, Employment of information, and Evaluation. His approach is more component-based. In his model, he identifies the main components of debriefing and additionally he introduces written debriefing. Written debriefing is a concept, which is elaborated more in Chapter 12.

RUDOLPH ET AL. (2008)

Rudolph et al. (2008) defined their debriefing model based on a summary of multiple debriefing models, with the most important ones being the three models by Thatcher and Robinson (1985), Lederman (1991), and Petranek (1994), which are presented above. Their model consists of three stages:

1. Reaction, where the participants, along with the debriefer, clear the air, review the facts and set the stage for addressing the learning objectives.

2. Understanding, where the participants explore what happened, unpack frames through advocacy-inquiry, apply good judgement (Rudolph et al., 2006) and teach; which moves participants to new skills and generalise lessons learned to real situations.

3. Summarise, where the participants review and discuss the takeaways and the lessons learned that will be applied in future events.

ZIGMONT ET AL. (2011)

Zigmont et al. (2011) developed the 3D model of debriefing, which is based on Kolb's (1984) Experiential Learning Cycle, the Adult Learning Theory (Sheckley et al., 2008), and the Learning Outcomes Model (Zigmont, 2010). It incorporates well-established debriefing phases and addresses the adult learner, the key experiences, and the learning environment(s). The latter, the learning environment, is also the model's major contribution, since none of models proposed previously, considered the environmental factors to be significant during debriefing.

Regardless which model of debriefing someone follows, debriefing is influenced by several factors (Peters and Vissers, 2004), such as the:

- Simulation game design (design decisions, game mechanics etc.).

- Diversity of participants (professional and educational background, professional role etc.).

- Type of simulation (constructive, virtual, or live) Morrison and Meliza (1999).

- Domain (transportation, healthcare, military etc.).

- Purpose (learning (Garris et al., 2002), checking competences (Berns et al., 2013), testing concepts (Zhou and Venkatesh, 1999), acceptance tests (Klemeš, 1986), situation awareness (Lo and Meijer, 2014), decision making (Meijer, 2009), design (Meijer, 2015) etc.).

11.3. FINAL REMARKS ON THE METHODS

In this chapter, the different methods and factors that influence debriefing were illustrated. By now, even non-experienced on debriefing readers have a notion of what debriefing is about. In the next chapter, several pitfalls while debriefing and proposed solutions, which were identified through a literature review, are presented.

12

Pitfalls while Debriefing

This chapter aims at identifying factors that inhibit debriefing in the literature and categorise them using experts' opinion. The goal is not just to create a check-list of pitfalls, in order for the debriefer to use and avoid while debriefing, but to propose an efficient and effective, science based methodology for structuring debriefing. This chapter adopts two categorisations, one from a design perspective and one from an executive perspective: 1. To what extent a pitfall occurs due to problems on the design of the debriefing, and to what extent due to problems on the execution of the debriefing, and 2. Does a pitfall mostly occur in rule-driven games/closed simulations (hereinafter referred to as 'closed games'), in free-play games/open-simulations (hereinafter referred to as 'open games'), or both.

12.1. Pitfalls in Literature

In this section, literature on factors that inhibit debriefing is reviewed, resulting in a list of pitfalls. This list is then used as an input for the proposed methodology in Section 12.2. The literature study revealed 12 pitfalls relating to debriefing:

1. The debriefers' level of involvement and style is not appropriate (Dismukes and Smith, 2000).

2. Debriefers have a lack of understanding of the debriefing process (Arafeh et al., 2010), which can lead into providing "easy" solutions (Ursano and Rundell, 1990) and/or violating the debriefing process Der Sahakian et al. (2015). This might occur due to lack of training and/or interest to improve.

3. Lack of plan (Der Sahakian et al., 2015) and/or rules (Sawyer and Deering, 2013).

4. The allocated time for the debriefing is short (Husebø et al., 2013) and/or the complexity of the simulated scenarios, occurring during the debriefing, may require a

Parts of this chapter have been published in Roungas et al. (2018a).

repetition of the game, or lead into violating the planned time of the debriefing (Der Sahakian et al., 2015).

5. Ineffective use of audiovisual (A-V) material (Crookall, 2010), which can lead to interruptions in finding relevant video segments (Seymour, 2012).

6. Lack of emotional safety of the participant, probably revealed because of (a) different levels of experience between the participants, (b) a difference in education (c) various other psychological reasons Seymour (2012) or due to the fact that debriefers might not take into account emotions (Der Sahakian et al., 2015).

7. Factors related to the actual physical environment, where the debriefing takes place (Decker et al., 2013, Seymour, 2012).

8. Choosing the appropriate structure for debriefing (Kolfschoten, 2007).

9. The tendency of the participants to assign blame (Der Sahakian et al., 2015) and antagonise each other.

10. Lack of trust of the participants towards the debriefers (Dyregrov, 1997).

11. The simulation is not organised in a personal basis, which inhibits the effectiveness of debriefing (Petranek et al., 1992).

12. Inappropriate timing/scheduling of the debriefing (Neill and Wotton, 2011).

As the above list shows, pitfalls while debriefing vary significantly. By no means can this list be deemed complete as it is drawn out of specific contexts, but it is the product of an extensive literature review, and as such, it contains the majority of the most important factors that inhibit debriefing. In Section 12.2, a methodology for identifying the pitfalls that contribute to an inefficiently and ineffectively designed debriefing is proposed.

12.2. RESEARCH APPROACH & METHODOLOGY

In this section, the pitfalls presented in Section 12.1 are categorised, aiming at identifying those pitfalls responsible for inefficiently and ineffectively designed debriefings. The focus of this study is on the design aspects of debriefing; its purpose to develop a broad methodology for debriefing and not just a checklist for avoiding pitfalls. Given the list identified in Section 12.1, which already has a considerable amount of underlying complexity, the list could grow quite lengthy. Therefore, a methodology is needed both for practical and scientific use.

The methodology for debriefing might need to be different for the different types of games and is influenced by factors like the participants, the facilitator, and the variations in the context of the game. The conglomeration of all these criteria would probably render a complex and difficult to analyse result, which is why researching games is a challenge (Kriz, 2003, Leigh and Spindler, 2004). Therefore, it is a logical first step to look at factors that a facilitator is able to influence, such as the design and the execution of the debriefing. Gaming is a design science, and therefore looking at design is crucial (Klabbers, 2009). The design of the game influences the behavior, which influences the learning during the game, and hence reflection during debriefing. Kolfschoten (2007) made a distinction between the person that designs a collaboration, which in our case is the debriefing, and the one who executes it. Despite the fact that in a game these persons might be one and the same, this distinction is an indication of the different tasks and skills the design of a debriefing requires as compared to the execution. Hence, this criterion could enable the distinction and focus on those pitfalls that are mostly influenced by the design of the debriefing.

The second criteria used is the division between open (free play) and closed (rule based) games (Klabbers, 2009). In other words, whether a pitfall is more likely to occur in a closed or in an open game. The design of these games, both at the opposite ends of a continuum, is fundamentally different. In a rule based game, rules define the actions in the gameplay, whereas in an open game there are nearly no rules other than rules to guide ethical behavior like “do not hurt each other” and “you may stop playing the game at all times after deliberation with the facilitator” (Klabbers, 2009). During debriefing, looking back at the game play is a crucial part. As such, using this criterion allows to identify whether, and to what extent, the rules on a closed game, which constrain behavior during gameplay, influence the debriefing. Therefore, the hypothesis is that the rules governing a game affect the debriefing methodology. The aim is to verify this hypothesis by conducting case studies, which will help capture this process in practice, where both the debriefing of open and closed games is qualitatively studied.

In order to verify which pitfalls contribute in an ineffective and inefficient debriefing, and whether, and to what extent, the rules of the game affect the debriefing methodology, an online questionnaire was built. The questionnaire construction involved the following decisions:

1. Questionnaire type (Lethbridge et al., 2005). The online questionnaire is semi-structured, which is evident by the use of a pre-defined set of questions rated on a Likert scale or through Boolean values, followed by the possibility for the interviewees to comment on their response. This protocol combines the advantages of the structured, comparable results provided by scale-based questions with the flexibility and richness of feedback from the comments. Moreover, the web-based nature of the questionnaire facilitates approaching people without any geographical restrictions. Additionally, examples of different open and closed games are offered, in order to serve as reference point.

2. Selection of subjects. The questionnaire was sent to facilitation experts. A facilitation expert is defined as a professional, who has been active in the gaming field more than 15 years and who has been designing and facilitating game for different groups, cultures, and domains. The reason for that choice is twofold. On the one hand, the lack of the ability to be present and explain any quandary regarding the questionnaire, required for the subjects to be particularly experienced, in order to minimise any threat to the validity of the results. On the other hand, the high expertise of the subjects, even when their number is relatively small, ensures that more plausible and valid conclusions will be made.

The structure of the questionnaire, for each pitfall, is shown below.

To what extent does this occur due to the design of the debriefing?

- ○ 1
- ○ 2
- ○ 3
- ○ 4
- ○ 5

To what extent does this occur due to the execution of the debriefing?

- ○ 1
- ○ 2
- ○ 3
- ○ 4
- ○ 5

Mostly occurs on:

- ○ Rule-driven Games - Closed Sims
- ○ Free-play Games - Open Sims
- ○ Both

From an epistemological point of view, this research is a combination of interpretivism and post-positivism. On the one hand, interpretivism enables the incorporation of the complexity that arises from human behavior (Bredo and Feinberg, 1982), which is the building block of debriefing (Kriz, 2003). On the other hand, reality is better interpreted through post-positivism (Denzin and Lincoln, 1994). In other words, the combination of these seemingly opposed research philosophies allows to capture the complexity of games (interpretivism), through the pitfalls. One may subsequently deduct patterns and propose a methodology (post-positivism), by using the questionnaire both with quantitative and qualitative aspects so as to verify the initial hypothesis.

In the next section, the results from the questionnaire are analysed in detail.

12.3. RESULTS

In this section, the results from the questionnaire are presented. In total, from the 20 subjects that were approached, only 8 fully completed the questionnaire. Hence, the results presented are only based the answers of experts that completed the questionnaire in full. The experts' demographics are:

- They are all ISAGA members, and some also members of Digra and Absel.
- They all have approximately 15 years of experience in designing and facilitating games for different groups, cultures, and domains.

- 5 of them are males and 3 females.
- 4 of them are academics and 4 are professionals in the commercial market of games.
- They come from Germany, Switzerland, Australia, Sweden, USA, and The Netherlands.

The number of experts does not allow for a statistically reliable generalization of the results. Nevertheless, due to their high expertise in facilitating games, the results provide interesting information.

With regards to issues occurring due to the design of the debriefing, pitfalls 3, 7, 8, 11, and 12 seem to be affected the most, since they have an average larger or equal to 4. Moreover, with the exception of pitfall 11, all these pitfalls have standard deviation less than 1, showing a convergence in expert's opinion.

With regards to issues occurring due to the execution of the debriefing, pitfalls 2, 4, 6, 9, and 10 seem to be affected the most, since they have an average larger than 4. Again here, with the exception of pitfall 10, all the rest have standard deviation less than 1, showing a convergence in expert's opinion.

Two noticeable facts derived from Table 12.1, where all the results are shown in detail, are the following:

1. None of the pitfalls in either of the 2 options (design and execution) scored less than 3 in average, which shows that while a pitfall might occur mostly due to the design or the execution of the debriefing, in the end both the design and the execution contribute towards its existence; and
2. All pitfalls with average less than 4 have standard deviation larger than 1, which shows a conflicting opinion among the experts. This might be due to the nature of the pitfalls in reference or due to a misunderstanding on behalf of the experts.

Table 12.2 shows the results of whether a pitfall occurs mostly in closed games, in open games, or in both. The results do not show any noticeable distinction between these 2 types of games, but the outcome should be considered inconclusive. The reasons for this are:

- The controversy of the rendered results when compared with some of the comments the experts gave, in which they made a clear distinction between closed games and open games.
- The fact that the question is in a Boolean form and does not have any granularity, like the questions revolving around the design and the execution of the debriefing.

One of the advantages of the questionnaire is the ability for the experts to comment freely on each pitfall. Herebelow the most important comments and their contribution of each pitfall are mentioned and discussed further.

12.3.1. PITFALL 1: LEVEL OF INVOLVEMENT

Pitfall: The debriefers' level of involvement and style is not appropriate

TABLE 12.1. RESULTS ON PITFALLS OCCURRING DUE TO THE DESIGN OR THE EXECUTION OF THE DEBRIEFING

#	Pitfall	Design	Execution
1	The debriefers' level of involvement and style is not appropriate	3.62 (SD: 1.18)	3.5 (SD: 1.41)
2	Debriefers have lack of understanding of the debriefing process	3.25 (SD: 1.16)	4.75 (SD: 0.7)
3	Lack of plan and/or rules	3.87 (SD: 0.83)	3.37 (SD: 1.3)
4	The allocated time for the debriefing is short	3.25 (SD: 1.38)	4.37 (SD: 0.91)
5	Ineffective use of audiovisual (A-V) material	3.87 (SD: 1.64)	3.75 (SD: 1.48)
6	Lack of emotional safety of the participants	3.37 (SD: 1.4)	4.37 (SD: 0.91)
7	Factors related with the actual physical environment, where the debriefing takes place	4.12 (SD: 0.99)	3.75 (SD: 1.16)
8	Choosing the appropriate structure for debriefing	4.5 (SD: 0.75)	3.87 (SD: 1.45)
9	The tendency of the participants to assign blame and antagonise each other	3.37 (SD: 1.5)	4.5 (SD: 0.75)
10	Lack of trust of the participants towards the debriefers	3.25 (SD: 1.28)	4.37 (SD: 1.06)
11	The simulation is not organised in a personal basis	4.12 (SD: 1.45)	3.75 (SD: 1.38)
12	Inappropriate timing/scheduling of the debriefing	4.25 (SD: 0.88)	3.75 (SD: 1.16)

TABLE 12.2. RESPONSES OF EXPERTS ON APPLICABILITY OF PITFALLS ON OPEN/CLOSED GAMES (NA STANDS FOR 'NOT ANSWERED')

#	Pitfalls	Experts							
		1	2	3	4	5	6	7	8
1	The debriefers' level of involvement and style is not appropriate	Closed	Both	NA	Both	Both	Open	Both	Both
2	Debriefers have lack of understanding of the debriefing process	Both	Both	NA	Both	Both	Both	Both	Both
3	Lack of plan and/or rules	Both	Both	NA	Both	Both	Both	Both	Both
4	The allocated time for the debriefing is short	Open	Both	NA	Both	NA	Both	NA	Both
5	Ineffective use of audiovisual (A-V) material	Both	Open	NA	Both	Both	Both	Both	Close
6	Lack of emotional safety of the participants	Both	Open	NA	Both	Both	Both	Both	Both
7	Factors related with the actual physical environment, where the debriefing takes place	Both	Both	NA	Both	NA	Both	Both	Close
8	Choosing the appropriate structure for debriefing	Both	Both	NA	Both	Both	Both	Both	Close
9	The tendency of the participants to assign blame and antagonise each other	Both	Open	NA	Both	Both	Both	Both	Close
10	Lack of trust of the participants towards the debriefers	Both	Both	NA	Both	Both	Both	Both	Both
11	The simulation is not organised in a personal basis	Open	Both	NA	Both	Both	Both	NA	Both
12	Inappropriate timing/scheduling of the debriefing	Both	Both	NA	Both	NA	Both	NA	Both

EXPERTS' COMMENTS

12

"My experience in debriefing practices is that the extent and significance of the debriefing is very much dependent on the instructor's/debriefer's own attitudes towards and skills in debriefing."

"The execution is dependent on the random composition of the group in terms of temperament and group dynamics. When you design a low involvement and nothing happens, you have to do something. And vice versa; then you design high involvement in a spirited group, a simple procedure will suffice."

"The wrong level of debriefers involvement could destroy the free play character. In the rule driven games, there are much more situations fixed by the designer. So it should be clearer for the debriefer, how to act."

"In a rule driven sim, the facilitator is basically the TEACHER. The game is a tool that helps transmit the message. The teacher is highly involved."

"When the debriefers' level of involvement is not appropriate, this is mostly because:

a The design mis-directs the debriefer.

b The debriefer implements the design improperly.

c Both.

The link between debriefer error and open/closed sims is most likely to be linked to the expertise of the person, rather than to the type of simulation itself. That is if I am inexperienced, I will be no more - or less - likely to do the debriefing well because of the nature of the game itself. There is some likelihood that I will make fewer errors in debriefing a closed game early in my learning, but that may simply mean that I miss the potential for exploring learning that was present but not in line with the 'closed' debriefing format."

DISCUSSION

Experts believe that the attitude and the skill of the facilitator are important factors when determining the level of involvement and the style. Additionally, they consider the composition of the group, and the dynamics formed within, to be directly linked to the target audience, and as such influencing the level of involvement. Moreover, experts point out that the role of the facilitator is also influenced by the rules of the game, hence the distinction between closed games and open games makes sense.

12.3.2. PITFALL 2: UNDERSTANDING OF THE DEBRIEFING PROCESS

Pitfall: Debriefers have lack of understanding of the debriefing process

EXPERTS' COMMENTS

"When a facilitator is not up to the job, it is the facilitator's fault. Nothing in the design or type of game can make that change."

"It has a lot to do with the 'Haltung', in which you start to prepare a debriefing."

"Idem to pitfall 1, my experience in debriefing practices is that the extent and significance of the debriefing is very much dependent on the instructors'/debriefers' own attitudes towards and skills in debriefing."

DISCUSSION

Experts' opinion converges towards the attitude and the skill of the facilitator to be factors that strongly influence pitfall 2, indicating also a correlation with pitfall 1. This convergence of opinions potentially also shows the importance of these factors.

12.3.3. PITFALL 3: PLAN AND RULES

Pitfall: Lack of plan and/or rules

Experts' comments

> "If you like to work according to a plan, you should prepare it in advance - in the execution it would be too late. I have difficulties with the word *plan*. It could both mean theory or methodological ideas."
>
> "At first glance, a 'lack of plan/rules' could be primarily due to flaws in the design that flow over into the debriefing. Similarly, it could be due to the debriefer not thinking through what might happen in each enactment of a design. A closed game is more likely to have more rules/plans in place but this is not necessarily going to help a debriefer at all times."

Discussion

Experts perceive *'plan'* as a term that could be referring both to the methodology and the structure of debriefing. Additionally, *'plan'* can also be considered as the application of theory, and as such it is more applicable to closed games, since rules and theory are important components of such games. An important outcome of the comments is that experts believe that keeping a strict plan could potentially inhibit debriefing, since the debriefer might not be able to connect and identify what participants need.

12.3.4. Pitfall 4: Allocated time for debriefing

Pitfall: The allocated time for the debriefing is short

Experts' comments

> "A good designer will provide a clear statement about notional debriefing time, based on their beta testing regime. This may take longer or shorter. And it may - or may not - include conditions similar to the one being experienced by any particular debriefer. Thus, the designer may over/underestimate the time required, and therefore cause a debriefer using the activity for the first time to fall into error. This might not occur for a second time for that debriefer but might occur again each time a new debriefer follows the designer's directions. Conversely, the debriefer may over/underestimate the time required simply because of their own prior experiences and encounter error of their own making. For example, I recently used a game that usually takes 30 minutes to play, and generates competition and conflict. In this occasion, it took only 15 and the players collectively deciphered the key design factors and bypassed the errors they would otherwise have fallen into. I allowed too much time for the debriefing, because I was 'secure' in my knowledge about the time this procedure usually takes."
>
> "In a closed game, this pitfall can occur the first times the game is played. Because these games are quite straightforward, the pitfall is likely to be solved quickly."
>
> "This could relate to complexity of simulated scenarios - during debriefing."
>
> "This happens frequently to facilitators that have little or no experience with the game at hand, and therefore are missing the capability to speed up or slow down the game when desired."

> "Lack of time is a real threat to the effectiveness of debriefing. This factor also leads to problems when researching the effectiveness of debriefing, since it affects the amount of time that the participants are subjected to the content of the game. I don't know how much it differs between different types of games, but my guess is that an open game will be more affected by a lack of sufficient debriefing."

> "Many times there is an objective for the game. By walking around, asking questions, showing participants what is happening, afterwards walking through a course of actions, etc. a facilitator is able to generate discussions that will be fruitful in the debriefing. When you do this in a right way - *c'est le ton qui fait la musique* - people will not experience this behavior as manipulating. It is just curiousness."

Discussion

Experts point out that reality can differ from what has been planned, and this should be taken into account. Their extensive experience indicates that debriefing closed games usually take less time and such games are relatively easier to plan because they are more predictable. Often, the complexity of game scenarios causes the debriefing to take more time, which is more applicable for open games.

12.3.5. Pitfall 5: Audiovisual material

Pitfall: Ineffective use of audiovisual (A-V) material

Experts' comments

> "This is also a pitfall I have seen in my own studies. Without replays, the instructor can sometimes be contradicted or challenged, lessening the impact of the game."

> "It sounds like a lack of preparation. The facilitator must have the presentation organised, if this is not the case, it is due to both the design and the execution."

Discussion

Experts' opinion, which is also supported by the literature Grant et al. (2014), Savoldelli et al. (2006), Scherer et al. (2003), shows contradicting beliefs and practices. Several consider the use of audio-visual material to be essential for an effective debriefing, and the effective use of it should be taken into account while preparing the debriefing. Other experts perceive audio-visual material to be time consuming while debriefing and less effective for handling emotions.

12

12.3.6. Pitfall 6: Emotional safety

Pitfall: Lack of emotional safety of the participants

Experts' comments

> "I have learned to choose activities with care to provide for a safe learning context, and still get it wrong sometimes. And each time that the 'wrongness' emerges during the debriefing point, I can trace it back to the choice

> of activity and what it brings out for one/some players. I have not yet chosen an activity that is wrong for everyone, (I remain alert to that possibility) but do know that when the debriefing reveals issues of emotional safety, it is because of internal factors unique to the individual(s) and not primarily because of either the design or the debriefing process."
>
> "Guarding the safety is a typical pedagogical skill one may expect from a facilitator/teacher. Even when the debriefing is ill designed, a good facilitator can keep it safe."
>
> "Emotional safety is such a multifaceted issue, that it is hard to pinpoint which part of the debriefing affects it the most. It depends on what type of emotional safety we are talking about."

DISCUSSION

Experts' opinion show that guarding the emotional safety of the participants can be related to the chosen activities during debriefing, but it can also be related to individual/group traits that cannot be influenced neither by the design nor the execution of the debriefing. It takes a lot of skill from the facilitator to address this well.

12.3.7. PITFALL 7: PHYSICAL ENVIRONMENT

Pitfall: Factors related with the actual physical environment, where the debriefing takes place

EXPERTS' COMMENTS

> "Mostly you will have to work with a given situation and cannot plan the perfect environment. So you will have to improvise during execution."

DISCUSSION

Despite the fact that it was perceived as a pitfall affected mostly by the design of the debriefing (scoring of 4.29 (SD: 0.95) and 3.71 (SD: 1.25) on design and execution respectively), albeit not with statistical significance, the only comment that we received showed that, many times, the securing of an appropriate environment for debriefing requires improvisation on behalf of the facilitator.

12.3.8. PITFALL 8: STRUCTURE OF DEBRIEFING

Pitfall: Choosing the appropriate structure for debriefing

EXPERTS' COMMENTS

> "A facilitator can adjust the briefing process accordingly."
>
> "It depends on random group dynamics."

DISCUSSION

Experts' comments show that the structure of the debriefing depends on different circumstances during the debriefing.

12.3.9. PITFALL 9: BLAME AND ANTAGONISATION

Pitfall: The tendency of the participants to assign blame and antagonise each other

EXPERTS' COMMENTS

> "This is very seldom due to problems in the debriefing - but has much to do with the design and management of the activity itself in the first instance. Where the debriefing is concerned, problems of blame/counter blame that arise will less often be caused by the design of the debriefing, but will be influenced by the implementation."

> "When this is foreseen, a wise design can help prevent it. A facilitator can also manage the process and try to stop it. When it does occur in a strong way, I would say that it is due to the design of the game itself. When a game triggers fierce emotional reactions, the game design is too competitive or confrontational. It is not analogous enough I would say."

> "One of the roles of the debriefer is and should be to counteract these behaviours, such as setting up rules for the debriefing session (or rules of conduct for serious gaming overall)."

DISCUSSION

Experts believe that not only the design of the debriefing can cause this pitfall but also the design of the game itself. An example of this is the Fish Banks game Whelan (1994).

12.3.10. PITFALL 10: TRUST TOWARDS THE DEBRIEFERS

Pitfall: Lack of trust of the participants towards the debriefers

EXPERTS' COMMENTS

> "This is another 'more than/less than' question and is very dependent on the quality and capability of the debriefer. It is also dependent on the quality of the management of the activity itself. If the game manager makes changes or interacts in a manner that is not intended by the designer this can generate dis-trust that will inevitably carry over into the debriefing."

> "This can be due to many factors, including participants being forced to play the game by their superiors. A good facilitator can reduce the resistance, but cannot make it go away completely. Undoubtedly, if the players do not feel safe in the vicinity of the facilitator, the facilitator is the problem, which is also possible. Then the facilitator is to blame."

> "Many times I have seen facilitators taking over the problems the manager has expressed. This can devastate trust completely."

> "In my experience, this is fairly rare, especially for experienced facilitators."

DISCUSSION

The skills and the personal traits of the facilitators are considered by the experts to play a significant role, since they can determine for example whether the facilitators act independently or as an extension of the management, or whether they are in-line with the game designer's intentions.

12.3.11. PITFALL 11: ORGANISATION IN A PERSONAL BASIS

Pitfall: The simulation is not organised in a personal basis

EXPERTS' COMMENTS

> "This is not something that 'occurs due to a problem' with the debriefing in any way. It has happened before the debriefing begins and will have an impact on the quality of the debrief."
>
> "Hosking and Morley state that people are both products of the context and participants in the shaping of the context. In my opinion it is not the individual that should be the central element, but - instead - the relation between individual and the group. Where did you assimilate? Where did you accommodate? People can be accommodators in one setting and assimilators in another."
>
> "Depending on the game, the participants and the purpose of the serious game (i.e. the learning goal) this isn't necessarily something negative. In some cases, it is better to focus on the group rather than the individual, for instance when teamwork is evaluated or when targeting an individual can have negative social and/or emotional effects on that individual."

DISCUSSION

Experts state that also this pitfall depends on context factors as well as on the purpose of the game. Games in essence are always social constructs in which individuals have their role Klabbers (2009).

12.3.12. PITFALL 12: TIMING/SCHEDULING OF DEBRIEFING

Pitfall: Inappropriate timing/scheduling of the debriefing

EXPERTS' COMMENTS

> "I have heard from colleagues about all of these problems in every combination of simulation types, and originating in both planning and execution. The biggest problem that I have witnessed over the years is a persistent belief that games do not need to be debriefed for the students to learn what was intended by the designer/facilitator. This belief seems to be tied up with a feeling that debriefing is a waste of time. In my personal experience, this is very much not true."
>
> "Choices about timing of a debriefing schedule are influenced by a large number of factors: the time available in the overall/larger program, the preceding/following activities, the design of the debriefing as verbal only/verbal-written or written first verbal later etc."

DISCUSSION

Experts confirm this is an important pitfall that is related to many underlying factors, ranging from the debriefing methodology to the importance a facilitator attaches to the debriefing.

12.4. Final Remarks on the Pitfalls

In this chapter, a methodology is proposed for identifying the factors that inhibit debriefing because of problems on the design of the debriefing and for ascertaining whether these pitfalls depend on the different types of games. The methodology includes a questionnaire, which is answered by game facilitation experts, and results in both quantitative and qualitative results. The answers complement each other so the quantitative results could be interpreted and placed into the perspective in which an answer is given. As shown in the analysis of the results in Section 12.3, the conclusions that can be drawn are:

1. On the one hand, the pre-defined questions showed that most of the experts consider all pitfalls to be relevant to both closed and open games. On the other hand, the comments showed that some pitfalls (Pitfalls 1, 3, and 4) are more relevant either to closed or open games. This contradiction can characterise the results, with regards to this categorisation, as inconclusive. If pitfalls prove to be independent of the rules of games, they will disprove the initial hypothesis. Thus, it is important and interesting to research this relationship until the point that it would be possible to support or disprove the initial hypothesis with statistical significance.

2. Despite the fact that some pitfalls seem to occur mostly due to the design and others due to the execution of the debriefing, all pitfalls had an average of 3.25 or higher on both categories, showing that to some extent, both the design and the execution of the debriefing influence all pitfalls. This result came as a surprise, since it was expected that the pitfalls [or at least some of them], were independent either from the design or the execution of the debriefing. Therefore, it will be interesting in the future to validate these results and understand the underlying reason for the above.

3. Both the experts that filled the questionnaire, and the ones that did not, reported that they perceive debriefing as a complex event due to the multiple, context and game-related factors it depends on. Nevertheless, their comments gave insight on the relationships among pitfalls and context factors, which in the future can help to model debriefing by abstracting it, the same way a game abstracts a real-world system.

4. The personal traits of the facilitators, such as their skills, experience, attitude, style, and overall personality, influence in multiple ways the effectiveness of debriefing. Researching further when, where, and how a facilitator influences debriefing is both important and fascinating, since it introduces new aspects from different scientific fields to our analysis, such as psychology, education, and management.

Additionally, with regards to future work, the list of pitfalls can be enhanced, by further researching the state-of-the-art in the gaming field and conducting more interviews with facilitation experts. Getting responses from additional experts will help establish a statistical significance on the results, and fine-tune the questionnaire wherever applicable. Moreover, decomposing the pitfalls into their core constituents, will help identify where and when a pitfall occurs, as well as its origin. Finally, constructing a relationship model of these core constituents can assist in the identification of patterns and propose a methodology for a more effective and efficient debriefing. Throughout this process and in order to verify and build up every step of the methodology, case studies shall be necessarily conducted.

13

The Tacit Knowledge in Games

Game sessions consist of three phases: briefing, gameplay, and debriefing, with the latter being considered the most important feature of games (Crookall, 2010). Nevertheless, their almost completely synthetic nature raises the question: are game sessions in general and debriefing in particular performed and analysed in a rigorous scientific way? In other words, are they consistently structured, given the different characteristics of games, and is it also clear what would constitute a successful game session and debriefing? The answer to all these questions is no. The reason for this negative outcome is that expertise regarding game sessions and debriefing resides almost entirely in the tacit knowledge spectrum. As a result, knowledge and best practices on how to conduct fruitful game sessions and debriefing are either disseminated without understanding the reason for which a particular decision would be beneficial or not disseminated at all. Hence, the aim of this chapter is to shed some light in this tacit knowledge possessed by experts and to gain understanding on why certain practices are more prone to success than others as well as bring into the surface other practices that have remained well hidden. In order to accomplish this goal, two rounds of interviews were conducted.

13.1. 1st Round of Interviews

The first round of interviews was with 19 experts of which 7 game designers, 6 project leaders, 4 participants, and 2 department managers. The primary tool for analysis was Q-methodology, while at a later stage the Principal Component Analysis (Abdi and Williams, 2010) and K-means clustering (Likas et al., 2003) were used for further validating the results. In the Q-methodology, the results from the first four interviewees were used to build the q-sort statements, which the remaining 15 interviewees used. Results, showing in Table 13.1, revealed several factors that either boost or inhibit games' success.

Parts of this chapter have been published in Roungas et al. (2019a) and Roungas et al. (2019c)

TABLE 13.1. RESULTS FROM FIRST ROUND OF INTERVIEWS USING THE Q-METHODOLOGY (ANGELETTI, 2018).

Factor	Impact	Comments
Presence of a game manager	+	A person that would attend all game-related procedures was found to be beneficial. These procedures involve choosing participants, make these participants available the day of the game, managing missing players, taking care of the space and the infrastructure for the gaming session, to name a few.
Managerial guidance and involvement	+	The involvement of mid/high level managers made the participants feel that what they are doing during the game session matters and it is not just a game.
Structured and concrete results	+	While the limitation of analytical sciences have been pinpointed in this thesis, complete absence of it is also detrimental. Apart from the lack of robust scientific methods for evaluating certain results, the absence of quantifiable results was found to be diminishing the credibility of the game itself.
Strict rules	+	Stricter rules were perceived by the interviewees as an insurance of higher validity of results.
High variety of roles involved in game design	+	Involvement of stakeholder not just during the game but also during the design process was appreciated by the interviewees, especially from operational personnel.
Simulator validated beforehand	+	Not properly validated software has created frustration among the stakeholders and negative opinion about the game overall.
Structured debriefing	+	Particularly for games for P&DM, an unstructured open discussion after the game was found to often distract from the goal of the game.
High complexity of the game's scope	-	Due to time and budget restrictions, over-complex games should be avoided, in order for results to be obtained in an affordable and timely manner. Moreover, often complex environments tend to overwhelm the participants causing the opposite effect from the desired one.
Unexpressed and/or conflicting stakeholders' interests	-	Unexpressed interests and expectations were found to severely increase the risk of unanswered research questions and unclear results.
Time pressure	-	Time pressure was recognised as a factor that forces untested or not well tested simulators to be used in game session that often causes crashes in the software leading to negative appreciation on behalf of the participants and potentially invalid results.
Pressure from external actors (for obtaining a solution suitable to their interests)	-	Some stakeholders might put pressure for obtaining results that fit their interests and agenda, which in turn can cause conflicts among all the stakeholders and potentially invalid results.

13

The application of Principal Component Analysis was inconclusive, while the K-means clustering showed similar results with the Q-methodology, thus further validating the findings.

13.2. 2^{nd} ROUND OF INTERVIEWS

The second round of interviews was with 21 game facilitation experts, which was mainly characterised by the contradicting answers in almost all questions. This result translates in a non-unified approach towards games in general and debriefing in particular. The complexity characterising modern systems, as it was examined in Section 1.1.1, immediately excludes pure analytical methods as the absolute and only solution, as the probability for ludic fallacy (Taleb, 2004) increases significantly. Therefore, these interviews aim to provide insights on how facilitation experts approach debriefing, hence tap into their tacit knowledge.

The questions these interviews intend to address are:

1. Given the limitation of analytical methods to provide clear criteria for success of game sessions, how should success be defined?

2. What is the level of knowledge of clients regarding their goal using games and how should they be prepared prior to the game session?

3. How do facilitators adapt their approach to the game session based on the players' characteristics?

The first question yielded perhaps the most answers with regards to how experts define success. 21 interviews resulted in more than 10 different answers, confirming the lack of consistency in the field. Nevertheless, three answers were far more common than the others. Freedom and feeling safe to share your experience from the game was considered a factor of paramount importance by six experts. The second most frequent criteria for success was the degree to which players would actually implement in their work, what they have learned during the game. Finally, a factor acknowledged particularly from game designers, that could determine success, was the level of involvement of players and their desire to play the game again.

The first part of the second question was initially expected to be answered overwhelmingly positively, yet more often than not, clients want to build a game but without knowing the actual goals. In the second part of the question, in order for facilitators to manage the varying levels of awareness of clients, the former inform the latter about the possible, unpredictable results of open games, like games for P&DM.

The third question relates back to theory, where it was introduced the idea of the interchanging roles that facilitators can, and should, take during a game (Kriz, 2010). The first step for facilitators is to identify any knowledge gap of the players with regards to the game they will participate in. Then, when the participants feel safe enough during the debriefing, the facilitator should capitalise that by taking the conversation into a deeper level. It should be noted that the interviewees acknowledged the influence of particular debriefing methods but none stood out as more effective or preferred.

13.3. CONCLUSION

The two rounds of interviews, analysed in this chapter, provide "inside" information on best practices when conducting game sessions and subsequently for debriefing. While analytical methods can provide invaluable insights when quantitative variables are available, the kind of knowledge provided in this chapter can only be attained by interviewing experts and then properly interpreting the results. While conducting such interviews, it is inevitable to encounter conflicting opinions from the interviewees; that was especially the case in the second round of interviews, with the facilitation experts, which is a strong indicative of the various profiles of facilitators and on the different ways the interpret certain situations. This characteristic is neither positive nor negative, it is a fact that needs to be acknowledged and taken into account while conducting such research.

IV

Knowledge Management

All men by nature desire to know.

Aristotle

14

Theory of Knowledge Management

All knowledge is not created equally (Dixon, 2000). The same applies to the knowledge gained from games. This part of the thesis therefore explores how knowledge from games can be acquired, stored, and disseminated. In this first chapter, the state of the art on knowledge management in general and on the elements of the games that are relevant to knowledge management is explored. A knowledge management framework is then proposed and applied in three case studies in Chapter 15.

14.1. Introduction

Knowledge management and reuse is not, and should not be, of academic interest only. The effectiveness of a corporation depends heavily on how it manages and reuses knowledge (Markus, 2001), or in layman terms, how in the first place it obtains and thereafter maintains the so-called "Know-how". As a corporation acquires and builds up on knowledge, it improves its know-how, and thus sustains or even increases its competitive advantage (Dixon, 2000).

In order for the reader to be able to easily follow the rationale in this chapter, certain terms need to be clarified in advance. Particularly terms including "knowledge management". Namely, these terms are:

> **Knowledge Management.** While there is not one widely accepted definition of knowledge management (KM) (Ragab and Arisha, 2013), it can be seen as a set of processes for storing, transforming, and transporting knowledge throughout an organisation (Gold et al., 2001). KM is considered to be a way of thinking that has, or at least should have, its roots deep in an organisation's culture, rather than being a project-specific one-time process (Wiig, 1997).

Parts of this chapter have been published in Roungas et al. (2018d) and Roungas et al. (2018c).

Knowledge Management Framework. From a practical point of view, KM should be translated in a system, i.e. a knowledge management system. In order to build such a system, the backbone structure should first be set, or in other words, a knowledge management framework (KMF) should be defined. A KMF defines the major KM elements, their relationships and the principles that define the way in which these elements interact. In this way, it provides the reference for decisions about the implementation and application of a KM system within an organisation (Metaxiotis et al., 2005). In this study, KMF as a term is used to describe a consistent set of guidelines for the instantiation of a KM system for games.

Knowledge Management System. Knowledge management systems (KMS) are a class of information systems, which aim is the creation, transfer, and application of knowledge in organisations (Alavi and Leidner, 2001).

What differentiates games from the ordinary activities of an organisation, and subsequently their knowledge management, is their abstracted nature. By default, a game environment is a simplified version of reality, thus using a game effectively requires proper abstraction during the design process and appropriate transfer of the "abstracted" knowledge in a real-world setting upon the completion of the game session. Therefore, while theory on knowledge management can be used in games, it cannot be adopted without carefully considering the distinct characteristics of games, as those were identified in Section 1.2.

Moreover, KM of games should ideally not be seen as an independent artefact that an organisation needs to foster; it should rather be seen as a module that can greatly enhance the KM process of an organisation that utilises games extensively. This means that such an organisation should already have adopted certain KM practices in general, which can then be complemented with a framework targeted specifically in games. Alternatively, in an organisation that has a poor, or even a complete lack of, KM culture, a KMF for games should have the capability to help those parts of the organisation that are involved with games even if that is only accomplished for a narrow small part of the organisation.

14.1.1. MOTIVATION

Despite the fact that games have proven to be cost effective, in multiple occasions, they still involve a substantial financial cost (Michael and Chen, 2005). Depending on several factors, like the degree of realism or the intended audience, the cost of developing a game may vary significantly. For instance, board games are considered to be a low-cost solution, whereas high-fidelity simulators usually bear a significant cost. In addition to development costs, there are costs associated with game sessions which, more often than not, are not trivial. A game session might require expensive hardware, an appropriate space to take place, and most importantly participants, who are compensated for participating in the game. Moreover, time is required to process the game outcomes and come with the best possible business decision. This additional time does not only increase the accrued costs but also delays decisions that sometimes are time-sensitive.

All of the above combined with the lack of a comprehensive methodology for managing and reusing knowledge acquired through games, lead companies, researchers, and game practitioners to “reinvent the wheel” by conducting consecutive and (almost) identical game sessions, accompanied by data analysis. The motivation for this study is therefore triggered by our strong belief that the capturing, compilation, maintenance, and dissemination of knowledge requires a methodology that will maximise the game outcomes concurrently with the minimisation of the associated costs and risks.

14.2. Background Work

In this section, the state of the art on KM frameworks is explored (Section 14.2.1) and the building blocks of the framework are identified (Section 14.2.2). Due to gaming being a relatively young and immature field, characterised by a lack of literature in KM, existing literature in the general areas of knowledge capturing, compilation, maintenance, and dissemination is used, which creates a pathway towards KM of games.

14.2.1. KM Frameworks

Various KM frameworks have been proposed during the course of the last two decades. A large number of these frameworks are either broad, high level, or relatively old (Leonard-Barton, 1995, Nonaka, 1994, Petrash, 1996, Van der Spek and Spijkervet, 1997, Wiig, 1993). There is though a significant number of recent frameworks that are specialised and could be interesting to explore for gaming. Below, some of these specialised frameworks are analysed, in order to extract lessons for the gaming field.

Maier (2007) proposed a framework that is structured in three different levels: strategic, design, and organisational. The framework is based on well-thought classifications and categorisations of its elements and identifies key aspects of KM. However, users and in general key stakeholders are mostly absent, which in gaming play an important role.

Pawlowski and Bick (2012) proposed a global knowledge management system that is well structured and very detailed. The core of the framework is the processes, which are divided in three levels: knowledge, business, and external processes. Everything else, like the stakeholders, the company culture, the technical infrastructure, revolves around those processes. The framework is currently one of the most complete and comprehensive KM frameworks and the gaming field can benefit greatly by this approach.

Pirró et al. (2010) developed a framework for the digital implementation of a KMS, which would be distributed and P2P. While from a theoretical point of view, the paper does not make significant contributions, the technical analysis is interesting and potentially beneficial for gaming.

Kavitha and Ahmed (2011) focused on developing a framework that would bring together the area of KM and that of agile software development. Agile methodologies have become quite common also in gaming, thus an approach like this one could provide insights on how the development of a game, and not just the gameplay, could contribute towards the organisational knowledge.

14.2.2. KMF Components

In this subsection, the building blocks of the KMF are defined. Each building block refers to a different aspect of the KMS and/or the organisation. Namely, these building blocks are the *KM Strategy* (Section 14.2.2), *Purpose of KMS* (Section 14.2.2), *Type of Knowledge*

(Section 14.2.2), and *Users of the KMS* (Section 14.2.2). An important factor absent from the framework is an organisation's culture.

organisational culture and knowledge management (KM) have a reciprocal relationship. On the one hand, the cultural values within an organisation influence the way people experience the KM outcomes and force the underlined KMS to evolve (Alavi et al., 2006). On the other hand, KM shapes the organisational values (Alavi et al., 2006) and improves the organisational performance through the development of human capital (Hsu, 2008). Therefore, the organisational culture is an important building block of a KMS and to a considerable degree it influences the success or failure of a KMS. Yet, as it has already been mentioned, a KMS for games is not a standalone artefact but a module that can only flourish in an organisation that has adopted KM, not just as a practice but deep in their culture. Thus, the organisational culture as an element of the KMS is not taken into account for this particular instance.

KM Strategy

By looking into management consulting firms, Hansen et al. (1999) distinguished two KM strategies, which in turn heavily influence the final implementation of the KMS. These strategies are called *Codification* and *Personalization*. *Codification* stores and makes available for reuse any acquired knowledge, which is in reality isolated from its source. On the other hand, *Personalization* is the exchange of knowledge that has been acquired in the past through one-to-one conversations and brainstorming sessions; it is a way to promote discussion and exchange of ideas and knowledge between people in a more personal manner.

Codification should be preferred when people want to learn from past projects and apply this knowledge in the future (secondary knowledge miners) (Markus, 2001), thus they would rather consult a documented and detailed record of these past projects. *Personalization* should be preferred when people inquire on experts' opinion but do not want to acquire their knowledge (expertise-seeking novices) (Markus, 2001), thus they would rather consult an expert in a one-to-one conversation.

Purpose of KMS

There are various reasons for which an organisation would want to build a KMS. Moreover, a single KMS might be built for more than one reason. Namely, these reasons are i. own-project improvement, ii. cross-project improvement and organisational learning, and iii. network improvement. While there is a possibility that building a KMS for one purpose would automatically exclude it from being used for another, the constrains that each purpose might impose are context- and organisation-specific.

Own Project Improvement: Own project improvement refers to the utilisation of the knowledge acquired during the lifecycle of a game to improve the game itself and the project for which the game was built. With regards to the game itself, unlike the waterfall methodology (Kasser, 2002), in which the game's development lifecycle follows a linear sequential approach, modern agile methodologies allow for an iterative approach that enables the requirements, the design, and the final solution to evolve continuously (Cockburn, 2000). In such a setup, knowledge management is required to improve the game at any given moment.

With regards to the project for which the game was built, KM can help it two ways. On the one hand, it allows to capture, organise, maintain, and disseminate the knowledge from the game and through that build evidences that would steer stakeholders towards the solution of the problem under examination (Roungas et al., 2018d). On the other hand, KM can be used as a root cause analysis (RCA) tool. In general, RCA is a collective term used to describe a wide range of approaches, tools, and techniques used to uncover causes of problems (Andersen and Fagerhaug, 2006). In other words, RCA refers to the investigation of why a certain event happened, and particularly which are the root factors that caused this event (Rooney and Van den Heuvel, 2004). Particularly in games, RCA requires strict and precise protocols to be in place, in order to help with the examination of problems or failures that might occur throughout the lifecycle of a game or due to decisions made based on a game (Latino et al., 2016). There is a vast selection of tools for implementing RCA in a KMS. On top of these tools, a significant addition is the implementation of digital forensics that bears many similarities with RCA and have been largely used in organisations (Grobler et al., 2010, von Solms et al., 2006).

Cross Project Improvement: Cross project improvement refers to the utilisation of the knowledge acquired during the lifecycle of a game to improve other projects, current or future. This is the most common usage of a KMS. A KMS can influence a current or future project either explicitly or implicitly. The term explicitly means that the knowledge generated or the skills acquired by the involved stakeholders during a game are directly used in another project.

The term implicitly, also known as organisational learning (Jarrar, 2002), means that the knowledge generated during a game creates added value (Spender, 2008), perhaps even a paradigm shift, within the organisation, which changes its modus operandi, and consequently influences any project developed thereafter. This kind of knowledge utilisation is considered to be a more long term investment, albeit one that significantly affects the competitive advantage of an organisation (Easterby-Smith and Lyles, 2011). The learning aspect of a KMS is one of the most frequent reasons for which organisations build a KMS. Yet, since the study is more focused on the contextual type of knowledge and on the principal as a beneficiary, learning is not explored further but acknowledged due to its significance.

Network Improvement: Network improvement refers to using a KMS to strengthen the relationships of individuals and teams within an organisation, especially in large organisations, by bringing awareness of the totality of knowledge possessed within. This is accomplished by using a *Personalization*-type of KMS and thus, utilising experts' knowledge in an organisational or even inter-organisational setup (Ahuja et al., 2012). The utilisation of a KMS as a mean to improve an organisation's network should be trod carefully. A KMS will not succeed in an organisation with a limited or non-existent network; but within an organisation with strong interconnections, it can further strengthen the inter-organisational structures and bring together experts from different professional background and geographic locations.

Users of KMS

Regardless of its type and purpose, the primary function of a KMS is to manage and disseminate knowledge to people, i.e. users. Therefore, users are the epicentre of a KMS and any frameworks aiming at building a KMS should put users first.

Markus (2001) identified three categories of users:

1. Knowledge producers are the people that contribute their knowledge into the KMS. A KMS can only be successful if people contribute and use it. Therefore, a KMS framework should provide incentives fork the knowledge producers to frequently and effectively share their knowledge and expertise. Moreover, knowledge producers should be experts in their respective field, since the success of a KMS depends heavily on the credibility of the primary source of information. A person who aims at using knowledge previously acquired shall be confident of the expertise of the knowledge producers, and thus trust their respective findings (Watson and Hewett, 2006). Usually, knowledge producers are people that one way or another have been involved either in the game development or the game sessions.

2. Knowledge intermediaries are the people that manage the knowledge, by indexing, summarising, and to the extent that possible and appropriate objectify it. One of the most important goals of a KMS should be the adaptation of methodologies for mitigating knowledge bias that derives from human subjectivity, which is an inseparable part of the contributed knowledge (Musen, 1992). Knowledge producers need to feel confident that the time required to contribute to a KMS is not wasted time and that their input has high chances to be easily accessible and used.

3. Knowledge consumers are the end users of the KMS, thus the ones that benefit from it. Depending on the type and the purpose of the KMS, knowledge consumers can be the game designers, project managers, investigators, researchers, or even the participants of a game, to name a few.

TYPE OF KNOWLEDGE

Knowledge can be defined in a number of ways. One of the most widely used definition is the distinction between explicit and implicit, the latter also known as tacit, knowledge (Smith, 2001). According to this classification, explicit knowledge is considered to be data or information that is communicated in a formal language and/or digitally or printed information that can be shared, such as manuals. On the other hand, tacit knowledge focuses on the cognitive features of humans, such as mental models, beliefs, insights, and perceptions. Spender (1996) proposed a categorisation of knowledge not just as explicit or implicit but also in juxtaposition with whether it is possessed individually by a person or within a social environment, like an organisation. Table 14.1 shows the categorisation proposed by Spender (1996).

TABLE 14.1. THE DIFFERENT TYPES OF ORGANISATIONAL KNOWLEDGE (SPENDER, 1996).

Knowledge Type	Individual	Social
Explicit	Conscious	Objectified Scientific
Implicit	Automatic	Collective Cultural

In cognitive sciences, explicit and implicit knowledge have conceptually different definitions. Explicit knowledge refers to conscious knowledge that resides in the working memory, whereas implicit knowledge refers to unconscious knowledge that is difficult to retrieve and verbalise (Dienes and Perner, 1996). This study proposes a mixture of these two definitions, hence acknowledging the implicit knowledge that is consciously possessed by individuals or groups as opposed to the knowledge residing in the unconscious part of the human brain. Figure 14.1 shows an illustration of the explicit (Type 1) and tacit (Type 3) knowledge, as well as the newly defined implicit-explicit knowledge (Type 2).

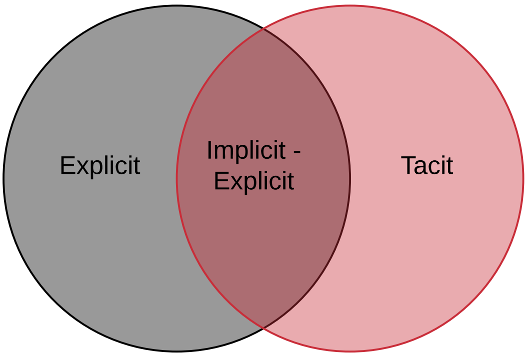


Figure 14.1. An illustration of the explicit, tacit, and implicit-explicit knowledge.

Below, an analysis of the explicit and implicit knowledge in games is presented.

Explicit knowledge

Explicit knowledge produced in and from games can be of quantitative or qualitative nature. In each phase of the game explicit knowledge is produced that serves one or more of the purposes identified therein Section 14.2.2.

There are five phases that intertwine in the lifecycle of a game. Namely, these phases are: 1. Requirements elicitation, 2. Game design, 3. Validation and Verification, 4. Game session, which also includes debriefing, and 5. Data analysis. All these phases are analysed in more detail below.

Requirements are the first step towards developing a game. There are two sources from where requirements should be elicited, the client and the real system the game imitates (Roungas et al., 2018d). The latter might be obsolete in case of open games, where there are (almost) no rules or restrictions.

In order for one to keep track of the progress and make sure that all the features are implemented as planned, eliciting and documenting the game requirements remains equally important throughout the development of the game. Although requirements are usually considered to be relevant only for the game they are elicited for, according to Zave (1997), requirements engineering is also concerned with the evolution of the relationships among the several factors of a system across software families. As such, requirements immediately become a tool for knowledge reuse, as they provide common ground for comparing different systems and pointing similarities. These similarities can be used either to improve future game development, as domain specific knowledge (Callele et al., 2005), or in order to avoid building new games and reuse the outcome of previously created games to analyse a current issue.

Game design is the art and science of actually designing a game. Numerous methodologies have been proposed on how to design a game but this study is only concerned with the knowledge management of games, thus only with the aspects of game design that affect the KMS. Therefore, when it comes to game design, what is crucial in terms of knowledge management, is the proper structure and documentation. There have been several approaches towards the adaptation of game design documents from the entertaining industry to more "serious" games, as well as approaches towards structuring the latter in a model-driven way (Rodríguez Rocha and Faron-Zucker, 2015, Roungas and Dalpiaz, 2016, Tang and Hanneghan, 2011). Regardless of the approach one chooses to adopt, there are certain game elements that need to be taken into account:

- Rules, which reveal information about the real system and the fidelity level of the game.

- Scenario, which shows the particular challenges and tasks of the game as well as its relevant setup details.

- Stakeholders/Actors, who are actively engaged with the game, such as participants, facilitator(s), game designer(s), and any other interested party.

- Purpose, which is the reason the client wants to build the game, e.g. to investigate a certain scenario that assists the organisation in their decision-making.

Documenting the above indicated information would allow for a comparison between a previously created game and a potential new game, which, in turn, can determine whether the new game is actually required or not (Roungas et al., 2018d). In case the similarities between the two games are enough for the results of the old game to be used in the current occasion, the new game is obsolete.

Validation and Verification (V&V) of games are two relatively different approaches that almost always go hand by hand. Game validation deals with the assessment of the behavioral or representational accuracy of the game and addresses the question of whether we are creating the "right game" (Balci, 2003). On the other hand, game verification deals with the assessment of the transformational accuracy of the game and addresses the question of whether we are creating the game in the right way (Balci, 2003).

Despite the fact that V&V usually require formal methods and quantitative data, they are almost always subjective. Moreover, games cannot be absolutely validated (Martis, 2006) and verified, and successful V&V can only be claimed for an instantiation of a game (a specific game session) and for a specific use (purpose). Therefore, meticulously documenting every detail associated with V&V is of paramount importance; it can also reveal a twofold benefit to potential knowledge consumers: i) they can ascertain, with rather minimal effort, whether the results of the game can be used for the intended purpose, and ii) they can, again with much less effort than without the V&V details, perform their own V&V study and hence, use the game for slightly or completely different purposes.

A V&V study incorporates two types of data; i) metadata associated with the V&V study, and ii) information describing the V&V process along with raw data and final results.

The metadata that should be stored are:

- The date, time, & location (if relevant) of the V&V study.
- The version of the game, in case there are multiple versions available.
- The purpose for which the game has been used.

With regards to the actual V&V, the information that should be stored is:

- The methods used for the V&V study and the justification on preferring these methods for the different phases of the game (Roungas et al., 2018f), like game requirements, game design, and game results (Balci, 2003).
- All the input and output data, both quantitative and qualitative ones, that were used during the V&V study.

Game session is where the actual gameplay takes place and it includes the preparation for the gameplay (briefing), the actual gameplay, and the debriefing. A game session can also be seen as a game instantiation. In object oriented programming terms (Rentsch, 1982), the game can be seen as a class with the rules and general guidelines of how the game works, whereas the game session can be seen as an object of this class. A game is usually designed once (involving several iterations) but can be played multiple times with a similar or a completely different setup. In other words, a game session is the application of a game with a specific scenario, stakeholders, and purpose. Therefore, this characterization helps to understand how an actual KMS can be built to support a game.

Every game is different, hence it requires a different approach with regards to how the knowledge produced can be captured and stored. Nevertheless, all games have the same main pillars: metadata, input and output data (quantitative & qualitative), and debriefing (Roungas et al., 2018d).

The metadata that should be stored are:

- The time, date, & location (if relevant) the game session took place.
- Detailed information regarding all stakeholders (participants, facilitator(s), and any other interested party).
  - Professional & educational background.
  - Age, sex, and any other relevant information.
- The annotations to all the data (quantitative & qualitative), which can be in the form of textual description, path to figures and audiovisual material on a server etc.

Input data usually include the game design decisions captured in the requirements and documented in the game design document, and they are common (but not exactly the same) for every instantiation of the game. Some input data might differ for each game session, where variations in the rules or the scenarios can be introduced (Roungas et al., 2018d). Input data can be both quantitative and qualitative.

In turn, output data might include quantitative data produced during the gameplay, audiovisual material captured during the game session or the debriefing, notes from the participants and/or the facilitator(s), game-specific artifacts, and a textual analysis of the lessons learned from the game. Same as with input data, output data can be both quantitative and qualitative.

With regards to quantitative data:

- Raw datasets and any further quantitative analysis of these data should be stored in separate tables in a database, and any textual description or analysis should be included in the metadata.

- Figures produced from quantitative data should be stored on a server, and the path and any other information associated with them should be included in the metadata.

With regards to qualitative data:

- Audiovisual and any other relevant material should be stored on a server. Whilst the corresponding material should be annotated for ease of use in the future, these annotations should be included in the metadata.

- Any quantification of qualitative variables should be stored in a database or on a server, depending on their format. Similarly to the textual description and the quantification methodology, the relevant material shall contain annotations, which shall be included in the metadata.

Data analysis is the final phase of a game, in which the data gathered during the game are analysed aiming at deriving meaningful insights, whether these insights concern a decision that has to be made or a hypothesis that was put into testing. While *data analysis* can be embedded in the *game session* through automated procedures, especially if the data are purely quantitative, in this study, it is treated separately. The reason for this distinction is that by using a KMS the aim is to be able in the future, whenever it is appropriate, to skip the *game session*, and use data gathered from a previous session to perform a new analysis.

Based on the expected output of a game, *data analysis* should incorporate the appropriate methodologies for handling the produced data. In the actual implementation of the KMS, different technologies like Python or the R statistical language, combined with quantitative and qualitative research methods, like the Q-methodology, can be used.

Implicit-Explicit/Tacit knowledge

With regards to the non-explicit knowledge, based on the definition of the different types of knowledge proposed in Section 14.2.2, this study is only concerned with the implicit-explicit (Type 2) knowledge. The reason for excluding tacit knowledge (Type 3) from this analysis is that due to its unconscious nature, it needs a completely different approach with regards to the methods required to capture and disseminate it.

Tacit knowledge in general, and as a result the implicit-explicit knowledge as well, is considered to be a tremendous resource for all activities within an organisation (Leonard and Sensiper, 1998). Particularly in games, implicit-explicit knowledge can correspond to pieces of knowledge or skills related to the game discipline (e.g. best game design practices, best game facilitation practices etc.), to the professional domain the game is applied in (e.g. experience, being mentored etc.), or the individual traits of people (e.g. talent in leadership or in creativity etc.) Yet, unlike explicit knowledge, implicit-explicit knowledge is not so straightforward to capture and manage. A database and a filesystem most probably would not be adequate to tackle the underline challenges. Therefore, different methodologies, which might also result in different approaches with regards to the implementation of the KMS, are needed. These methodologies should aim at converting the implicit-explicit knowledge to some sort of explicit knowledge (Nonaka and von Krogh, 2009), which can then be disseminated in a more unequivocal way. While, the literature is not exhaustive on how to capture and convert such knowledge, several approaches have been proposed, particularly from the perspective of tacit knowledge.

One of the most common techniques for capturing tacit knowledge is "cognitive maps", which facilitate the representation of individuals' view of reality (Eden et al., 1981). There are different types of cognitive maps, one of which is causal maps (Ambrosini and Bowman, 2001). Causal maps are interpretations of individuals' or groups' beliefs about causal relationships (Markíczy and Goldberg, 1995). Causal maps have been proven to be an effective tool for the elicitation of tacit knowledge for a variety or reasons, e.g. allowing to focus on action, eliciting context dependent factors etc. (Ambrosini and Bowman, 2001).

Semi-structured interviews are another tool that can help elicit tacit knowledge. While the purpose and structure of such an interview is predetermined, the essence of the "semi-structure" lies on the fact that interviewees are encouraged to answer questions by telling stories (Ambrosini and Bowman, 2001). The story telling nature of these interviews allows people to manage the collective memory of an organisation (Boje, 1991), frame their experiences (Wilkins and Thompson, 1991), and reflect on the complex social web of an organisation (Brown and Duguid, 1991).

Tacit knowledge encompasses a large amount of subjectivity and a research method to study it is the Q-methodology (Stephenson, 1953). In a nutshell, in Q-methodology the interviewee sorts a series of items/statements throughout a continuum (e.g. from strongly disagree to strongly agree) that is approximately normally distributed, in the sense that more of these statements are placed close to the neutral area than in the two edges of the continuum.

Various scholars argue that the use of metaphors can serve to transmit tacit knowledge (Ambrosini and Bowman, 2001, Martin, 1982) and since metaphors allow different ways of thinking, people may be able to explain complex organisational phenomena (Tsoukas, 1991). The term metaphors connotes the transfer of information from a relatively familiar domain to a relatively unknown domain (Tsoukas, 1991).

The most ancient form for exchanging knowledge in general (Gurteen, 1998) and tacit knowledge in particular is dialog. This is perhaps why social media have become prominent on how people interact not only at a personal but also at a professional level. While research is still relatively poor in this area, the use of social media sounds indeed promising in tacit knowledge sharing, since they encompass interactive and collaborative technologies (Panahi et al., 2012).

The aforementioned approaches for eliciting implicit-explicit knowledge can be used independently or in combination with each other. In the case studies presented in Section 15.2, the method of semi-structured interviews was used.

14.2.3. CONCLUSION OF THE LITERATURE

In this section, the literature on KM was examined. In Section 14.2.1, KM frameworks were researched and briefly described. While the list of frameworks is certainly not exhaustive, it is representative of the state of the art. Gaming as a field has a lot to learn from frameworks like the one proposed by Pawlowski and Bick (2012), yet even a comprehensive framework like this does not take into account the KM strategy nor the reason for which the organisation wants to develop the KMS; two factors that later in this study are shown to be important in KM.

15

A Knowledge Management Framework

In this chapter, based on the literature and subsequently the components associated with knowledge, identified in Chapter 14, a knowledge management framework is proposed and then tested with three case studies from the Dutch railways.

15.1. The KM Framework for Games

In this section, the composition of the KMF, based on the components identified in the literature, is described. The primary aim of the framework is to provide the guidelines for implementing a KMS module for games in the form of a knowledge infrastructure (Davenport and Prusak, 1998). An illustration of the KMF is shown in Figure 15.1. Knowledge is the preeminent component of any KMS, hence it holds the lion's share when analysing games; yet, the centre of the KMF is the KMS strategy. The KMS strategy an organisation adopts determines how the acquired knowledge is disseminated and it is a decision that should be made even before eliciting any knowledge. Choosing between a *Codification*, a *Personalization*, or a hybrid strategy depends on:

- the organisational culture, which gives a strong indication on the quality of the contributed knowledge from the knowledge producers and its utilisation from the knowledge consumers.

- its current technical infrastructure, in which case for an organisation that has a strong technical background and infrastructure, the implementation and maintenance of a *codified* KMS would be efficient and affordable. Whereas, for an organisation with a strong social factor among its employees, a KMS based on a *Personalization* strategy could be a more prudent solution.

Parts of this chapter have been published in Roungas et al. (2018c).

- the type of knowledge, in which case the more explicit a piece of knowledge is the easier it is to store it in a database-based system (*Codification*), as opposed to a more tacit piece of knowledge that is more easily disseminated through conversation (*Personalization*).

- the purpose for which the organisation wants to build a KMS, which is directly shown in Figure 15.1.

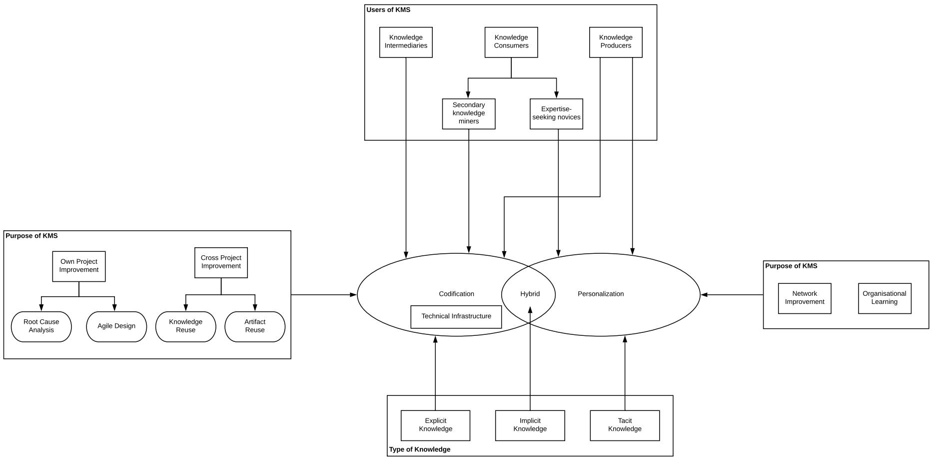


FIGURE 15.1. AN ILLUSTRATION OF THE KNOWLEDGE MANAGEMENT FRAMEWORK.

While everything seems to revolve around the KMS strategy, the type (Section 14.2.2), purpose (Section 14.2.2), and potential users (Section 14.2.2) of the KMS should also be understood and taken into account. In effect, the purpose and the users of the KMS heavily influence how knowledge from games is captured, stored, and disseminated. Moreover, the purpose of the KMS defines its potential users and particularly the knowledge consumers. The latter is illustrated with two examples. $1^{st}example$: the knowledge consumers of a KMS developed solely for *own project improvement* could be any stakeholder involved in the project but not individuals or groups outside the project. $2^{nd}example$: the knowledge consumers of a KMS developed for *root cause analysis* are going to be either specific individuals or groups involved in the project or investigators and/or digital forensic analysts.

15.2. APPLICATION OF THE FRAMEWORK

In order to validate and substantiate the structure and the components of the framework, three games conducted by the Dutch railway infrastructure manager (ProRail) are used as case studies. In Section 15.2.1, a summary of the three games is provided, while in the subsequent subsections, the unique components of these games are presented following the structure of the proposed KMF.

15.2.1. The Games

Since 2009, ProRail is using games to test future changes in the infrastructure, timetable, and processes, amongst other things (Meijer, 2012a). These games vary from testing with and training single train traffic operators to large scale multi-actor games that also involve operators from the principle train operating company, Nederlandse Spoorwegen (NS). The games used as case studies in this chapter were selected based on the type of question asked by the principle stakeholders and the technological characteristics of the game (analogue/digital). Namely, these games are called OV-SAAL, A2 chain simulation, and ERTMS. In Table 15.1, the three games are described based on their design requirements.

OV-SAAL

The OV-SAAL corridor between Schiphol-Amsterdam-Almere-Lelystad is part of the Program High-frequent Transport (PHS, in Dutch, Programma Hoogfrequent Spoor). One of the aims of PHS is to increase the number of trains per hour in the Randstad (The Dutch financial centre consisting of the four largest cities, Amsterdam, Rotterdam, The Hague, and Utrecht). In order to increase the capacity, i.e. run more trains on this corridor, and increase the quality of public transport in this region, several measurements need to be made, one of which is the doubling of tracks around station Weesp.

The project team for OV-SAAL requested a game session to support their decision-making on which infrastructure investment to make. The conclusions that were summarised and discussed in the game session were used to inform the project team about the outcomes; the project team members were present during the concluding part of the game session.

A2 chain simulation

Similar to OV-SAAL, the A2 chain simulation, which focuses on the A2 Corridor between the cities Amsterdam, Utrecht and Eindhoven, is also part of PHS. The A2 chain simulation is a large-scale participatory simulation, in which in four different game sessions (days), disruptions were simulated on the A2 corridor (the main corridor in the Netherlands connecting Amsterdam with Eindhoven) in combination with the new high-frequency timetable for 2018. The aim of the simulation was to 1. test the feasibility of the implementation and management of the new timetable, 2. measure the bottlenecks and learn how to mitigate them in case of disruptions, and 3. get the operational and management personnel familiar with the new high-frequency timetable.

The results from testing the new timetable and identifying bottlenecks were presented in a report to the project team of the A2 chain simulation. In turn, the project team analysed which results and recommendations needed to be further investigated during a pilot in real-time operations or to find solutions for bottlenecks that were deemed as urgent.

ERTMS

ERTMS stands for European Railway Traffic Management System and it is a major industrial project that aims at replacing the different national train control and command safety systems in Europe. The game currently aims at testing a selection of 10 out of 66

TABLE 15.1. GAME DESIGN REQUIREMENTS OF THE THREE CASE STUDIES

Core Aspect	**OV-SAAL**	**A2 Chain Simulation**	**ERTMS**
Purpose	Exploring the impact of different infrastructural expansions	Exploring the impact of different scenarios on a high frequency timetable	Testing the impact of newly created user processes for ERTMS
Scenarios	Five scenarios, such as 1. No infrastructural expansion, 2. Four additional tracks at Almere station, 3. Additional haul tracks at Weesp station etc.	Seven scenarios spread over four days (1-2 scenarios each game day). such as 1. freight train malfunction between Oud Zaltbommel and Geldermalsen, 2. Fire alarm at train tunnel in Best, 3. power failure at station Den Bosch etc.	Five scenarios to test ten newly created user processes, such as turning train direction at Lage Zwaluwe, entering a non-automatic management shunting area, combining two trains etc.
Simulated world	Railway infrastructure on two trajectories: Amsterdam Central Station - Lelystad and Amsterdam Zuid - Hilversum, co-location of operators occurred by seating arrangements (each table was a control centre). Current time table.	Railway infrastructures on the trajectory between Amsterdam Central and Eindhoven (A2 corridor). Co-location of operators per NS or ProRail control centre and national control centre (OCCR) separately. Timetable of the following year.	Future railway infrastructure with ERTMS safety system, co-location of operators in same room
No. of participants	8	Between 17 and 23 operators per session	3
Roles	Train traffic controller (TTC) (2), regional network controller (RNC) (2), national network controller (NNC) (1), regional passenger traffic monitor (RPTM) (2), national passenger traffic controller (NPTC) (1)	Train traffic controller (TTC) (6-9), regional network controller (RNC) (2-3), national network controller (NNC) (1), regional passenger traffic monitor (RPTM) (2-3), regional passenger traffic junction coordinator (RPTJC) (2-3), regional passenger traffic material and passenger coordinator (RPTMPC) (3), national passenger traffic controller (NPTC) (1)	Train traffic controller (TTC) (1), train driver (2)
Type of role	Similar to own (5), prior experience in role (3)	All operators played a role similar to their own in the operational environment, facilitators (ca. 6 per session) covered for other operational roles	All operators played a role similar to their own in the operational environment, one facilitator covered other operational roles
Objectives	Determining own decisions for the next 15 minutes given the status of the system at paused moment	Conducting operator tasks as done in the 'real' operational environment	Conducting operator tasks as done in the 'real' operational environment
Constraints	Separation of train traffic regions: one regional train (2) and passenger traffic centre (2) each versus other remaining regional train traffic centre (2), exclusion of roles outside the defined infrastructure area, exclusion of train driver	Limitation of a number of simulator functionalities, solved through workarounds	Limitation of a number of simulator functionalities, solved through workarounds
Load	Four types of disruptions: 1. Local train delay (+5 min), 2. Freight train delay (+10 min), 3. Corridor train (intercity) delay (+10 min), 4. Disruption as chosen by participants themselves	Scenarios included a various sorts of disruption with a major impact on train traffic operations. This required collaboration between operational roles in applying a disruption mitigation procedure	Scenarios include normal train traffic conditions with no disruption
Situation (external factors)	Presence of observers and video cameras. At the end of the day results were discussed with invited stakeholders	Occasionally presence of observers. A debriefing was held at the end of every scenario.	Presence of an operator acting as observer. A debriefing was held at the end of XX.
Time model	Step-wise (per time periods of 15 min)	Real-time/ continuous	Real-time/ continuous

newly developed user processes for ERTMS. The project is still on-going as the implementation of ERTMS should be operational by 2028 in the Dutch railway sector. Two game sessions were conducted, which had a combined focus of testing the technical development of the simulator and a number of new user processes. Unlike OV-SAAL and the A2 chain simulation, the governance of the ERTMS program has been a separate organisational structure and is only recently brought under the lead of ProRail. Multiple organisations, such as the Dutch ministry of infrastructure and the principle train operating company Nederlandse Spoorwegen (NS) are still strongly involved.

Table 15.2. Explicit and tacit knowledge in the three games.

Game	Type of Knowledge	
	Explicit	Implicit-Explicit
OV-SAAL	Audiovisual data from the game session and debriefing	Conducted interviews with 6 people in total, 4 game designers, 1 project team member, and 1 stakeholder
	A questionnaire measuring collaboration and decision-making	Game designers: Insights on the game design processes and the importance to involve stakeholders and how to balance validity and game design
	Notes taken by the observers on noticeable events or statements	Game designers with academic purpose: Knowledge on the operational activities of the Dutch railways
	Results of the game session documented in a spreadsheet	Project team member and the stakeholder: Insights on the value and use of games for the particular research questions
A2 Chain Sim.	Simulation's logs (e.g. train punctuality, replay of train traffic flow etc.), which were stored after each game scenario	Conducted interviews with 8 people in total, 5 game designers, 2 facilitators, and 1 participant
	Audio-video recordings during each group debriefing session	Game designers: Insights on the organisation of a large scale multi-actor game and the subsequent technical challenges
	Questionnaires regarding the game experience and the simulator's validity	Facilitators and participant: Insights on understanding why certain decisions are taken and what other operational roles entail
	The communication between the different operators, which were logged through a specific template	
	Events of the game session, such as the start and end time, technical issues etc., which were registered by the facilitators through a specific template	
	Reports comprised of the summaries of each game day and a report on the overall conclusions from the game sessions	
ERTMS	Reports (e.g. summary outcomes of the debriefing with regards to the validity of the simulator, gained insights on testing the new user processes, using the current simulator as a testing tool etc.)	Conducted 4 interviews, all of which with game designers, who gained insights on the processes and technical aspects of ERTMS, as well as the technological requirements

15.2.2. Organisational culture

The culture of an organisation is key in building, using, and maintaining a KMS. For an expert organisation such as ProRail, which has knowledge ingrained in different departments and levels, it would be a logical step to capture and structure knowledge from games, which can serve as a knowledge hub. For example, as the case studies are all part of a large scale infrastructural change program, knowledge that is developed during the design comes together in these game sessions. Project teams are interested in testing their assumptions, while operators are able to illustrate and share their knowledge on the viability of the designs. However, the urgency of KM as a process and of a KMS implementation must first be recognised within the organisation and supported. Currently, the implementation of a KMS is only considered for research purposes at ProRail, thus the proposed framework needed to be structured in such a way as to accommodate an implementation in a narrow, e.g. departmental, level.

15.2.3. KM strategy and purpose of KMS

The absence of urgency to establish a KMS system within ProRail might be explained by the purpose of games and adoption of the results. Each game has its own unique characteristics, yet some components are common between them. Currently, ProRail is interested in building a system that would accommodate the knowledge produced from games as opposed to strengthening its culture and internal network; thus, the KM strategy governing all games is that of *Codification*. With regards to the purpose of the KMS, *Cross project improvement* is the "usual suspect" for which a KMS is built, and in this case there is no exception. The knowledge from all games examined in this chapter is meant to improve future projects, yet the knowledge from some games is also meant for more than that. However, in the case of both OV-SAAL and the A2 chain simulation games, there has been no follow-up or request to retrospectively provide additional investigation, except for academic research on games. For the ERTMS game there is a potential use of *Root-Cause Analysis*, specifically in the case of investigations related to safety issues.

15.2.4. Type of knowledge

This component of the KMF can also be seen as the actual output of a KMS, in which the KMS obtains its physical structure. Table 15.2 summarises the produced and collected explicit knowledge for each game.

15.2.5. Users of KMS

Table 15.3. Potential KMS users for each game

Users	OV-SAAL	A2	ERTMS
Knowledge producers	Operators Game Designers	Operators Game Designers Observers	Operators Game Designers
Knowledge intermediaries	Not defined yet	Not defined yet	Not defined yet
Knowledge consumers	Researchers	Project team	Researchers

15

In the three games examined in this chapter, knowledge producers are mostly similar as the games are designed by a pool of experienced game designers and the games always include the knowledge of railway operators. Table 15.3 summarises the potential KMS users.

In the case of knowledge intermediaries there have been and there currently are no active stakeholders. However, knowledge transfer has been increasing the past years within the growing team of experienced game designers. Whereas mere four game designers were involved in the OV-SAAL game, this number increased to seven with the A2 chain simulation game. Also knowledge about game design has been transferred to other departments, in which the aim is integrate research games into the organisational strategy. As such, the urgency and need to develop a KMS may be aligned with a maturing organisational strategy on gaming.

Finally, the knowledge consumers have been predominantly researchers or game designers so far. Whereas the results of both the OV-SAAL game and the A2 chain simulation game resulted in advice to the project team, the project team of the A2 chain simulation game also included these findings in a database which was used for live-tests. As ERTMS is a program that runs over a number of years and has close implications with operational safety, it is expected that the number of knowledge consumers will grow when games will be fully integrated in the program.

15.2.6. Final Remarks on the Application of the Framework

The analysis of the games reveals that the proposed KMF is able to help the management and dissemination of knowledge derived from games, and the KMF itself is a proof of concept on the feasibility of developing a KMS module for games. Therefore, compared to other frameworks, as those were examined in Section 14.2.1, the proposed framework does not claim that it can contribute towards a standalone KMS but rather towards enriching an existing KMS that could then also target games. Still, the integration of such a KMF into a KMS seems, and usually is, labour intensive. Nevertheless, it is evident both from theory and from the games examined in this chapter that games produce different types and quality of knowledge. Particularly, the games examined in this chapter are designed for testing changes in the railway infrastructure, resulting in a strong focus on the debriefing after each game session. In turn, debriefing becomes the primary source of knowledge, especially for knowledge of Type 2 and 3. Hence, capturing all knowledge from games gives new opportunities for validity assessments at a higher level of detail, which both compliments and puts pressure to the current sense-making approaches (van den Hoogen et al., 2014).

Finally, the purpose for which an organisation builds, and then enhances, a KMS depends heavily on its maturity with regards to KM. In this context, mature means that the organisation has the "know-how" of managing knowledge, which allows it to follow a top-down approach on the design of the KMS, thus starting by first defining the purpose and then gathering the required data. On the contrary, immature means that the organisation follows a bottom-up approach on the design of the KMS, thus starting by first gathering data, and then defining the purpose of the KMS based on the quality of the knowledge produced from the gathered data.

15.3. CONCLUSION

The theoretical contribution of this part of the thesis is on how to approach the different levels of the engineering process, in order to improve the management and dissemination of the knowledge produced by and in these games, and particularly this type of knowledge that could be described as transferable tacit knowledge, i.e. Implicit-Explicit Knowledge. As a result, this study proposed a knowledge management framework and used three design and/or research games from the Dutch railways in order to ascertain that the framework can capture the knowledge in and around games and to contribute towards its integration within a knowledge management system. Moreover, the games provided additional insights on how such a framework can both enhance the validation process as well as help the organisation to grow. All in all, the framework proposes general guidelines on the components to consider for the development of a module that would accommodate the particularities of games, in order to integrate the knowledge acquired through games in a knowledge management system. Specific details on how to develop the knowledge management system in general and the game module in particular are dependent on the organisation culture itself and the users that support and use the knowledge management system.

With regards to future work, the final implementation of the knowledge management system would provide additional insights and yet another validation layer for the underlying framework. New technologies integrated in the system, like xAPI (Bakharia et al., 2016), could allow for dynamic evaluation and tracking of activities. Moreover, another major issue not analysed in this thesis is the privacy of data. Especially within the same organisation, ethical and privacy concerns arise with regards to data and knowledge obtained from certain projects, either due to the employees involved or the client for whom the project was developed. Finally, companies should nurture an organisational culture that would favour the dissemination of knowledge and further research towards that direction, and more specifically the factors that either positively or negatively influence the exchange of knowledge within an organisation, should be explored.

16

Conclusion

Man: a being in search of meaning.

Plato

In this final chapter, the conclusions from all four parts of this thesis are summarised and the research questions for both levels, as those were posed in Section 2.3, are answered. Moreover, the results from the questionnaire, which attempts to generalise the results beyond the relatively narrow scope of the Dutch railways, are presented. Finally, the limitations of this study and the future work are identified.

16.1. Summary of Part I through Part IV: The First Level

This study is comprised of four parts, each of which deals with an aspect of games. For each one of the aforementioned aspects, this thesis contributes in new methodologies or conceptual frameworks, or simply promotes a new way of approaching and thinking about games. In the four subsections below, the individual conclusions separately for each area of games are summarised.

16.1.1. Design

Game design is the focus of Part I. This was chosen to be the first part because it is actually the first step in a game's lifecycle. The predominant finding whilst researching the design of games was the severe limitation in formal methods. This limitation led us to explorr the formalisation of game design in an, at first, unrelated field, game theory. Whereas game theory is widely considered a branch of mathematics lying in the intersection with economics - which is true - it also bears an important resemblance with games, evidenced in the word "game". What initially seemed just a coincidence, it was actually what determined this seemingly unorthodox line of research, that is the utilisation of game theory for formalising game design.

Indeed, for complex systems with multiple actors, game theory is deemed successful in abstracting and subsequently modelling these systems. As a result, models derived from game theory can be used in game design in a much more formal way than before. While no method or tool can be absolutely precise and successful in designing games, the answer to the first research question, i.e. what aspects from game theoretical analysis can be translated to game design, and in what way, shows the connection between the game theoretical and the game design elements, which proved to be insightful on how certain game elements can be designed.

Moreover, the advantages of the framework are not only related to games but also extend to the general analysis of systems. The second research question, i.e. to what extent can the design of a meaningful game be determined from game theoretical analysis, reveals several advantages of using game theory with regards, but not limited, to game design, like the identification of the worst-case scenario and the prediction of how a situation characterised by a specific game concept can evolve in the future.

As a concluding remark, game theory, despite its drawbacks and limitations, was shown to be a powerful tool for the formalisation of game design.

16.1.2. VALIDATION

Part II focuses on validation, not just of games but also of simulations. The reason for not solely focusing on game validation was that the field of simulation validation is mature, incorporating multiple methods both formal and informal, and it seemed promising at offering insights that would also benefit games. As a result, four out of the five research questions were related to simulations and only one to games, though the answers to all research questions were designed as to potentially benefit also games.

In Chapter 7 and Chapter 8, the first research question, i.e. How can the selection of validation and verification methods and applicable statistical techniques given the simulation and the real-world system at hand be optimised as to be more time efficient and rigorous, is answered. While initially this line of research was focused purely on simulations, later it became evident that games could also benefit, which resulted in adding games in the framework, The proposed framework aims to optimise the simulation (and game) validation method selection and the case study shows that it can indeed be both time efficient and rigorous in selecting the most suitable validation methods.

In Chapter 9, the second research question, i.e. how can the operational validation of a simulation model be automated or semi-automated, in order to reduce the time, cost, and human error associated with it, is answered. One of the most important paradigm shifts of the last couple of decades has been agile methodologies (Schwaber, 2004). These methodologies have expanded also in games (Keith, 2010) and allow for game development phases like design, implementation, and validation to take place concurrently. While they have several advantages, one of their main characteristic which is the multiple iterations has some disadvantages, one of which is the need to validate the underline simulation model of the game after every change. As a result, the need for at least some level of automation in the validation process has become paramount. The methodology proposed in Chapter 9 uses web technologies and the R statistical language and it was shown that it can, to a great extent, mitigate several problems associated with validation, including the regular need to perform the full validation cycle.

In Chapter 10, the third research question, i.e. which factors play a critical role in the success of a simulation validation study, is answered. The methodology proposed in Chapter 9 was extensively used in two case studies, resulting in several lessons learned. While at a first glance the correspondence to games might seem elusive, it becomes relevant due to the fact that the two simulation models in this study have been also part of games, under a high-level architecture (Kuhl et al., 1999).

Finally, in Chapter 6, the fourth research question, i.e. which are the implications with regards to the validity of games and which lessons can we learn from pure simulations, is answered. The purpose of this chapter is not to propose a particular methodology with regards to game validation but rather to promote dialogue and understanding on the benefits and the limitations of using formal methods in game validation. In other words, a balance should be struck and this balance can only occur if the analytical and design communities involved in gaming recognise the need of one another and collaboratively bridge the gap between them.

16.1.3. GAME SESSIONS

Part III focuses on game sessions in general, with debriefing having the lion's share. Debriefing is considered to be the most important part of games (Crookall, 2010), since its aim is to bridge the abstracted world of games with reality, and thus transfer the acquired knowledge in a real world setting. In this part, debriefing was not the only subject examined; the best practices in conducting game sessions, through a series of interviews with debriefing experts and other game stakeholders, were presented.

Chapter 11 and Chapter 12 are solely dedicated to debriefing, where an introduction and some of the most well established methods on debriefing are described in the former. In the latter, a literature review initially revealed the most common pitfalls while debriefing games and then interviews with debriefing experts pinpointed whether, and to which extent, these pitfalls occur mostly in closed or open games, and whether there are most likely occurring due to the initial planning or the execution of the debriefing.

Finally, in Chapter 13, the research question, i.e. How can game sessions in general, and debriefing in particular, be performed and analysed in a rigorous scientific way, is answered. Two different rounds of interviews were performed; the first was with various game stakeholders (Angeletti, 2018), like game designers and project managers, whilst the second was only with game and debriefing experts. The first set of interviews identified several factors that either inhibit or improve the gaming experience and, subsequently, the obtained results. The second set of interviews provided insights into the tacit knowledge of game experts, an area of knowledge that has been to date left rather untouched.

16.1.4. KNOWLEDGE MANAGEMENT

In Part IV, knowledge management as a discipline and its application in the gaming field are examined. This thesis is concerned with games in which knowledge is the cornerstone. Thus, this knowledge should be properly captured and managed. The application of games in areas other than education has recently gained popularity and as such is still immature. The management of knowledge of such games is therefore almost completely neglected, as of now.

In Chapter 14, the general theory in knowledge management is presented, which is paramount in order to first understand the field and then apply this knowledge in games. In Chapter 15, a framework for a knowledge management system that supports games is proposed, which is then applied and validated in three case studies from the Dutch railways. What differentiates games from other sources of knowledge, and thus requires particular attention, are certain distinct characteristics that games have, like their abstractive nature, the variety of knowledge beneficiaries, to name a few.

While knowledge management is examined in a separate part, it should be mainly considered as an umbrella that covers, but is not limited to, all the other areas. It is not only through the game sessions that we learn through games; it is during the whole lifecycle of a game that knowledge is created and therefore can potentially be captured. Knowledge management should be therefore considered more of a continuous effort rather than a single phase during the game development.

16.2. SYNTHESIS OF INDIVIDUAL CONCLUSIONS

In this thesis, four areas of games were analysed in four distinct parts. They should not however be seen as four separate and unrelated entities. Being part of games, they often seem to intertwine and influence one another. In this subsection, the relationship between these four areas is explored, which subsequently answers the research question posed in the end of Section 2.3.1, i.e. How do the four areas intertwine and influence one another.

Choices regarding the design of games can positively affect or inhibit validation, depending on how the real system is translated into a game. This connection between design and validation becomes apparent upon moving from a more synthetic design approach to a more formal one, in which case the formalisation enables the utilisation of formal validation methods.

Validation increases a game's credibility in the eyes of the stakeholders, as it was shown in Chapter 13, which in turn enables a game session to be more effective. Equivalently, a game session that is considered "successful" increases the games credibility, usability, and applicability, thus making it more valid.

Knowledge management has been advocated as the "umbrella" that can potentially cover all aspects of games. As such, the methodologies associated with it should be, on the one hand, tailored in accordance with the design, validation, and game sessions, and on the other hand, in accordance with the culture of the users and the organisation.

16.3. GENERAL CONTRIBUTIONS: THE SECOND LEVEL

Apart from the conclusions in each of the four areas, there are contributions that the thesis as a whole aimed - and managed - to fulfil These contributions are based on three research themes in juxtaposition with the four areas of games, shown in Table 16.1. The three research themes are the *Analytical vs. Design sciences*, *Complexity*, and *Quality of Decision Making*, which are analysed in the following subsections. The added value of not just exploring each area separately is twofold. On the one hand, games should also be seen as a complete artefact, in order to understand how as a whole they contribute towards solving the problem at hand. On the other hand, by viewing games more holistically, the relative, and absolute, importance of each area in the big picture of a game is more easily determined and further explored.

TABLE 16.1. GENERAL CONTRIBUTIONS OF THESIS.

Game Area	Research Themes		
	Analytical vs Design	Complexity	Quality of DM
Design	Section 16.3.1	Section 16.3.2	Section 16.3.3
Validation	Section 16.3.1	Section 16.3.2	Section 16.3.3
Game Session	Section 16.3.1	Section 16.3.2	Section 16.3.3
Knowledge Management	Section 16.3.1	Section 16.3.2	Section 16.3.3

16.3.1. ANALYTICAL VS. DESIGN SCIENCES

Starting from the introduction (Chapter 1) and throughout the course of this thesis, it has been mentioned numerous times that games lie in the intersection of the analytical and design sciences. In this subsection, the distinct contributions of each school of thought in the four areas of games are summarised.

In Part I it was shown that the analytical sciences can contribute towards overcoming technical challenges, introduce step-wise formal approaches, as well as identify and model to a certain extent some contextual aspects of systems in game design. However, most of the contextual elements that games need to capture, that would enable them to be valid representations of the system under study, can only be approached from a design point of view.

In games with a dominant simulation layer, analytical methods can almost solely be used during the validation study. However, the more dominant the game layer becomes, the less these analytical methods can ascertain the validity of games. Games with a dominant game layer would benefit more from the design perspective, which could enable the capturing and potentially the quantification of terms like *psychological reality* and *usability*.

Briefing, gameplay, and debriefing can hardly benefit from analytical methods, although certain KPIs can be implemented to assess the progress and success. A game session should primarily be seen from a design point of view. From the space and time in and during which the session takes place, to the specific scenarios that run as part of the game play, all the way to the final debriefing process, it is mainly aspects of the design science that are utilised by game experts either consciously or unconsciously.

While knowledge management can benefit from the structure provided by analytical methods, it is primarily the design science that can enable one to approach it as holistically as possible. This way, the sharing and reusing of knowledge is maximised through the capturing of knowledge and the provision of a platform, both mental and physical.

16.3.2. COMPLEXITY

Systems, and particular the ones pertaining to this thesis, as it was identified in Section 1.1.1, are characterised by an increased complexity. This complexity is not just the result of systems' increased size but is mainly caused by the numerous interdependencies among their different aspects. In turn, while these interdependencies are abstracted to a certain degree, they still bear a significant amount of complexity, which needs to be translated into game design choices. The result is artefacts, i.e. games, characterised by numerous and complex structures with limited knowledge on behalf of researchers and practitioners on how to understand and model them. In Table 16.2, the four phases of games are shown vis-à-vis the three levels of complexity of systems.

Table 16.2. The different complexities in each game phase.

Game Area	Complexities		
	Technical	Actor	Context
Design	Analytical science	Behavioural science	Design science
	⟵ Game theory (See Chapters 4 and 5) ⟶		
Validation	Simulation layer (See Chapters 7 to 10)	Game layer (See Chapter 6)	Simulation & Game layer for the specific context (See Chapter 6)
Game Session	Open & Closed games	Participants & Principals Different background	Contextual & Generalisable knowledge
Knowledge Management	Explicit knowledge (See Chapter 14)	Tacit knowledge (See Chapter 14)	Context communicated through Personalisation (See Chapter 14)

16.3.3. Quality of Decision Making

It has been indicated from the beginning, in Chapter 1, that this thesis is concerned with games for decision making. As such, each area analysed and the associated findings should contribute to an increased quality of the decision-making capabilities of games. In this subsection, the contribution of this thesis on the improvements and the implications with regards to the quality of the decision-making of games are summarised.

The formalisation of the design phase ensures that the translation of the real-system in a game is done in a consistent and formal way, which then enables the utilisation of formal methods in the validation phase. On the other hand, the validation phase ensures that the game is a good representation of the system under study thus the derived decisions have a high probability of yielding the expected outcome. While this statement might initially appear to be obvious, it becomes more ambiguous the less realistic a game is; hence, the less formal methods can be used. In other words, the more abstract and metaphorical a game is, the more uncertain its validity and the relevant assessment therof becomes.

Regardless of how the design has been formalised and the success of the validation study, it is during the game session that the game is actually tested as to the extent to which it provides reasonable and applicable answers on the questions asked. Thus in practice, the outcome of the game session is the only true metric of quality of the decisions made based on the game.

Knowledge management ensures that the game results can be revisited in case they did not yield the expected outcome (root cause analysis) and reused for future reference, reducing thus any future risk and cost. Moreover, a decision making process comprises, more often than not, from multiple steps that all together build evidences on the system under study, and lead in turn to a decision. Hence, knowledge management is crucial in this “building” process, since it ensures that every step is documented in the appropriate manner, and as a result contributes positively to the quality of current and future decisions.

16.4. Generalisation of Results

Whereas the literature used to develop the various methodologies in this thesis is not domain-specific, most of the case studies are from the railway sector and from the Netherlands, which could create bias towards the generalisability of the results. It was therefore deemed necessary to seek the answer on whether, and the extent to which, the results presented in this thesis are generalisable beyond the narrow scope of the Dutch railways by asking 8 game experts of their opinion on the matter. The findings of this thesis were presented in the form of propositions, where the experts could answer in a Likert scale from *Strongly Disagree* to *Strongly Agree*, as follows:

> From Part I: Game theory has the ability to support game designers by first providing for a formal methodology for modelling the real system under study and then enabling the translation of certain game theoretical elements, like strategies and pay-offs, in game design decisions.
>
> From Part II: Validation methods from pure simulations can provide rigour in game validation for improving games' usability and credibility. However, the use of analytical methods in game validation should be done with caution and a balance should be struck between empirical and analytical methods.
>
> From Part III: With regards to game sessions, we have identified several factors that have either a positive or negative impact. *Positive:* Presence of a game manager, managerial guidance and involvement, structured and concrete results, strict rules, high variety of roles involved in the game design, game validated beforehand, structured debriefing. *Negative:* High complexity of the game's scope, unexpressed and/or conflicting stakeholders' interests, time pressure, pressure from external factors (for obtaining a solution suitable to their interests).
>
> From Part IV: Knowledge management can contribute towards extended usability and credibility of games by providing for a platform in which our knowledge about certain systems and games can gradually be built, As a result, important limitations with regards to analytically approaching game validity, which is the usually low number of participants, and partial coverage of the system of study are overcome.

The results, which can be found in more detail in Appendix D, were quantified with 1 representing the *Strongly Disagree* and 5 the *Strongly Agree*. All of the propositions were on average assessed positively, i.e. average above 3, which means that indeed the findings of this thesis are for their most part generalisable beyond the Dutch railways.

Needless to say that some factors raised conflicting responses even from the same interviewee, especially regarding Part III. Some examples include the conflicting stakeholders' interests, which apart from their below average score, they were found to have a negative impact and foster discussion in certain occasions; and the time pressure, which can also create eustress thus having a positive impact. Finally, it was indicated that while some factors might have a positive impact on their own, in combination with other factors, they might create a negative impact for the overall game session.

16.5. LIMITATIONS

The methodologies proposed in this thesis do have, as expected, some limitations. It is the multidisciplinary nature of the field of gaming that prohibits methodologies from being universally and unquestionably accepted.

The most apparent limitation of this thesis is the relatively narrow domain of the case studies, where with the exception of one case, they all have been from the Dutch railway sector. Though while this holds true for the case studies, this limitation is to a degree mitigated from the broad areas literature is originated from, which was heavily used to develop the proposed methodologies. The disadvantage of the narrow domain of the case studies is further mitigated from the positive results presented in Section 16.4.

Apart from the generic limitation identified above, each of the four areas this thesis delved with has its own limitations. In design, the limitations of the methodology coincide with the limitations of game theory. Game theory assumes that, with no exceptions, actors always behave and choose rationally, which has been shown not to be true (Simon, 1957). Moreover, game theory has only a restricted ability to reveal the reasoning behind certain choices made by actors; in other words it is a good tool for answering the What but not the Why. In validation, the nature of games prevented this study from diving deeper into proposing a specific methodology for game validation and instead promoted the adaptation of certain methods from simulations. In game sessions, the conducted interviews provided a wealth of information, a fraction of which was successfully decoded rendering interesting results. Given more time, more pieces of tacit knowledge could have been revealed that would interest the gaming community as a whole. In knowledge management, the greatest limitation of the proposed methodology is that it requires an already mature organisation, which employs and is familiar with knowledge management methods and, which is willing to invest resources for extending the current knowledge management system.

16.6. FUTURE WORK

This thesis is by all means not the end of the subjects explored, not even of the methodologies proposed. Several improvements and enhancements could take place in the future starting by addressing the most important limitation, the limited scope of the case studies. The methodologies proposed could be further validated in domains other than the railways or even transportation and logistics as well as in different geographical areas, i.e. outside the Netherlands.

With regards to each of the four areas of games, the methodology based on game theory, proposed in Part I, could been further validated by using it in a real-world design situation. In validation, a closer cooperation of the analytical and design communities could enable the application of formal/analytical methods in game validation as well as the quantification of terms like psychological reality and usability (Sauro and Kindlund, 2005). In game sessions, an abundance of information could be extracted even from the already obtained transcripts but the field could also benefit from more interviews with the proposed methodology. Finally, in knowledge management, the actual knowledge management system proposed could be implemented, further tested, and eventually become operational.

References

Abdi, H. and Williams, L. J. (2010). Principal Component Analysis. In *Wiley Interdisciplinary Reviews: Computational Statistics*, pages 433–459. Hoboken: Wiley Publishing Inc.

Abt, C. C. (1970). *Serious games: The art and science of games that simulate life.* Viking Press, New York, USA.

Adrion, R. W., Branstad, M. A., and Cherniavsky, J. C. (1982). Validation, verification, and testing of computer software. *ACM Computing Surveys (CSUR)*, 14(2):159–192.

Ahuja, G., Soda, G., and Zaheer, A. (2012). The genesis and dynamics of organizational networks. *Organization Science*, 23(2):434–448.

Alavi, M., Kayworth, T. R., and Leidner, D. E. (2006). An empirical examination of the influence of organizational culture on knowledge management practices. *Journal of Management Information Systems*, 22(3):191–224.

Alavi, M. and Leidner, D. E. (2001). Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly*, 25(1):107–136.

Albers, A., Hettel, R., Behrendt, M., Düser, T., and Schwarz, A. (2012). Extended flexible environment and vehicle simulation for an automated validation. In *Proceedings of the FISITA 2012 World Automotive Congress*, pages 1263–1273, London, United Kingdom. Springer.

Alessi, S. M. (1988). Fidelity in the design of instructional simulations. *Journal of Computer-Based Instruction*, 15(2):40–47.

Ambrosini, V. and Bowman, C. (2001). Tacit knowledge: Some suggestions for operationalization. *Journal of Management Studies*, 38(6):811–829.

Andersen, B. and Fagerhaug, T. (2006). *Root cause analysis: Simplified tools and techniques.* American Society for Quality, Milwaukee, WI, USA.

Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 23(2):193–212.

Angeletti, R. (2018). *Managing knowledge in the era of serious games and simulations: An exploratory study on the elicitation of serious games' requirements for the generation and reuse of knowledge.* PhD thesis, Delft University of Technology, Delft, The Netherlands.

Apperley, T. H. (2006). Genre and game studies: Toward a critical approach to video game genres. *Simulation & Gaming*, 37(1):6–23.

Arafeh, J. M. R., Hansen, S. S., and Nichols, A. (2010). Debriefing in simulated-based learning: Facilitating a reflective discussion. *The Journal of Perinatal & Neonatal Nursing*, 24(4):302–309.

Aristotle (1916). Book VII: 1336a. In Jowett, B., editor, *Aristotle's Politics (Translated in English)*, page 264. Oxford University Press, Oxford, United Kingdom.

Aronson, E., Wilson, T. D., Akert, R. M., and Sommers, S. R. (2018). *Social psychology.* Pearson Education Limited, London, United Kingdom, 9th edition.

Asuka, M. and Komaya, K. (1996). A simulation method for rail traffic using microscopic and macroscopic models. *WIT Transactions on The Built Environment*, 18:287–296.

Axelrod, R. and Cohen, M. D. (1999). *Harnessing complexity: Organizational implications of a scientific frontier.* The Free Press, New York, USA.

Backhouse, R. C. (1986). *Program construction and verification.* Prentice-Hall International, Upper Saddle River, NJ, USA.

Bakharia, A., Kitto, K., Pardo, A., Gašević, D., and Dawson, S. (2016). Recipe for success: Lessons learnt from using xAPI within the connected learning analytics toolkit. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pages 378–382, Edinburgh, Scotland. ACM.

Balci, O. (1988). The implementation of four conceptual frameworks for simulation modeling in high-level languages. In Abrams, M. A., Haigh, P. L., and Comfort, J. C., editors, *Proceedings of the 20th Conference on Winter Simulation*, pages 287–295, San Diego, CA, USA. ACM.

Balci, O. (1990). Guidelines for successful simulation studies (tutorial session). In Balci, O., Sadowski, R. P., and Nance, R. E., editors, *Proceedings of the 22nd Conference on Winter Simulation*, pages 25–32, New Orleans, LA, USA. IEEE Press.

Balci, O. (1994). Validation, verification, and testing techniques throughout the life cycle of a simulation study. *Annals of Operations Research*, 53(1):121–173.

Balci, O. (1997). Verification validation and accreditation of simulation models. In Andradóttir, S., Healy, K. J., Withers, D. H., and Nelson, B. L., editors, *Proceedings of the 29th Conference on Winter Simulation*, pages 135–141, Atlanta, GA, USA. IEEE.

Balci, O. (1998). Verification, validation, and testing. In Banks, J., editor, *Handbook of Simulation*, chapter 10, pages 335–393. John Wiley & Sons, Inc., New York, USA.

Balci, O. (2003). Verification, validation, and certification of modeling and simulation applications. In Chick, S., Sánchez, P. J., Ferrin, D., and Morrice, D. J., editors, *Proceedings of the 35th Conference on Winter Simulation*, pages 150–158, New Orleans, LA, USA. IEEE Press.

Balci, O. (2004). Quality assessment, verification, and validation of modeling and simulation applications. In Ingalls, R. G., Rossetti, M. D., Smith. J. S., and Peters, B. A., editors, *Proceedings of the 36th Conference on Winter simulation*, pages 122–129, Washington, D.C., USA. Association for Computing Machinery.

Balci, O., Adams, R. J., Myers, D. S., and Nance, R. E. (2002). A collaborative evaluation environment for credibility assessment of modeling and simulation applications. In Yucesan, E., Chen, C.-H., Snowdon, J. L., and Charnes, J. M., editors, *Proceedings of the 2002 Winter Simulation Conference*, pages 214–220, San Diego, CA, USA. IEEE.

Balci, O. and Sargent, R. G. (1984). Validation of simulation models via simultaneous confidence intervals. *American Journal of Mathematical and Management Sciences*, 4(3-4):375–406.

Banks, J., Carson, J., Nelson, B. L., and Nicol, D. (1984). *Discrete-event system simulation*. Pearson Education, London, United Kingdom, 4th edition.

Barendregt, H. P. (1984). *The Lambda Calculus: Its syntax and semantics*. Elsevier B.V., Amsterdam, The Netherlands.

Barlas, Y. (1994). Model validation in system dynamics. In *Proceedings of the 1994 International System Dynamics Conference*, pages 1–10, Sterling, Scotland. System Dynamics Society.

Barreteau, O., Le Page, C., and Perez, P. (2007). Contribution of simulation and gaming to natural resource management issues: An introduction. *Simulation & Gaming*, 38(2):185–194.

Beizer, B. (1990). *Software testing techniques*. Van Nostrand Reinhold Company Limited, London, United Kingdom, 2nd edition.

Bekius, F., de Bruijn, H., Cunningham, S. W., and Meijer, S. (2016). Structuring the multi-issue and hub-spoke games found in public administration. In *PICMET 2016, Portland International Conference on Management of Engineering and Technology*, Portland, OR, USA. IEEE.

Bekius, F. and Meijer, S. (2018). Selecting the right game concept for social simulation of real-world systems. In *14th Social Simulation Conference*, Stockholm, Sweden.

Bekius, F., Meijer, S., and de Bruijn, H. (2018). Collaboration patterns in the Dutch railway sector: Using game concepts to compare different outcomes in a unique development case. *Research in Transportation Economics*, 69:360–368.

Bennett, P. G. (1987). Beyond Game Theory - where? In Bennett, P. G., editor, *Analysing Conflict and its Resolution: Some Mathematical Contributions*, chapter 3, pages 43–70. Oxford University Press, London, United Kingdom.

Bergmann, S. and Strassburger, S. (2010). Challenges for the automatic generation of simulation models for production systems. In *Proceedings of the 2010 Summer Computer Simulation Conference*, pages 545–549, Ottawa, Canada.

Berns, A., Palomo-Duarte, M., Dodero, J. M., and Valero-Franco, C. (2013). Using a 3D online game to assess students' foreign language acquisition and communicative competence. In *Scaling up Learning for Sustained Impact*, pages 19–31. Springer, Berlin, Germany.

Bethke, E. (2003). *Game development and production*. Wordware Publishing, Inc., Plano, TX, USA.

Bever, B. and Collofello, J. S. (2002). An investigation of techniques for addressing software affordability. In *Aerospace Conference Proceedings*, volume 5, pages 2577–2585, Big Sky, MT, USA. IEEE.

Bhargava, H. K., Power, D. J., and Sun, D. (2007). Progress in web-based decision support technologies. *Decision Support Systems*, 43(4):1083–1095.

Binmore, K. G. (1987). Why game theory "doesn't work". In Bennett, P. G., editor, *Analysing Conflict and its Resolution: Some Mathematical Contributions*, chapter 2, pages 23–42. Clarendon Press, Oxford, United Kingdom.

Blomqvist, E. (2014). The use of semantic web technologies for decision support - a survey. *Semantic Web*, 5(3):177–201.

Boje, D. M. (1991). Consulting and change in the storytelling organisation. *Journal of Organizational Change Management*, 4(1):7–17.

Bolton, G. E. (2002). Game theory's role in role-playing. *International Journal of Forecasting*, 18(3):353–358.

Bolton, M. L. (2013). Automatic validation and failure diagnosis of human-device interfaces using task analytic models and model checking. *Computational and Mathematical Organization Theory*, 19(3):288–312.

Bond, W. F., Lammers, R. L., Spillane, L. L., Smith-Coggins, R., Fernandez, R., Reznek, M. A., Vozenilek, J. A., and Gordon, J. A. (2007). The use of simulation in emergency medicine: A research agenda. *Academic Emergency Medicine*, 14(4):353–363.

Bovee, M., Srivastava, R. P., and Mak, B. (2003). A conceptual framework and belief-function approach to assessing overall information quality. *International Journal of Intelligent Systems*, 18(1):51–74.

Bredo, E. and Feinberg, W. (1982). Part two: The interpretive approach to social and educational research. In *Knowledge and Values in Social and Educational Research*, pages 115–128. Temple University Press, Philadelphia, PA, USA.

Brown, J. S. and Duguid, P. (1991). Organizational learning and communities-of-practice: Toward a unified view of working, learning, and innovation. *Organization Science*, 2(1):40–57.

Burnstein, I. (2006). *Practical software testing: A process-oriented approach*. Springer-Verlag New York.

Byrne, J., Heavey, C., and Byrne, P. J. (2010). A review of web-based simulation and supporting tools. *Simulation Modelling Practice and Theory*, 18:253–276.

Byun, J. H., Choi, C. B., and Kim, T. G. (2009). Verification of the DEVS model implementation using aspect embedded DEVS. In *Proceedings of the 2009 Spring Simulation Multiconference*, San Diego, CA, USA. Society for Computer Simulation International.

Callele, D., Neufeld, E., and Schneider, K. (2005). Requirements engineering and the creative process in the video game industry. In *RE '05 Proceedings of the 13th IEEE International Conference on Requirements Engineering*, pages 240–252. IEEE.

Callele, D., Neufeld, E., and Schneider, K. (2006). Emotional requirements in video games. In Glinz, M. and Lutz, R., editors, *14th IEEE International Conference on Requirements Engineering*, pages 299 – 302, Minneapolis/St. Paul, MN, USA. IEEE.

Cantrell, M. A. (2008). The importance of debriefing in clinical simulations. *Clinical Simulation in Nursing*, 4(2):e19–e23.

Casti, J. L. (1986). On system complexity: Identification, measurement, and management. In Casti, J. and Karlqvist, A., editors, *Complexity, Language, and Life: Mathematical Approaches. Biomathematics, vol 16*, chapter 6, pages 146–173. Springer, Berlin, Germany.

Chakravarty, I. M., Roy, J. D., and Laha, R. G. (1967). *Handbook of methods of applied statistics*. John Wiley & Sons, New York, USA.

Chandrasekaran, S., Silver, G., Miller, J. A., Cardoso, J., and Sheth, A. P. (2002). Web service technologies and their synergy with simulation. In Yucesan, E., Chen, C.-H., Snowdon, J. L., and Charnes, J. M., editors, *Proceedings of the 2002 Winter Simulation Conference*, pages 606–615, San Diego, CA, USA. IEEE.

Cockburn, A. (2000). *Agile software development: The cooperative game*. Addison-Wesley, Boston, MA, USA, 2nd edition.

Cooper, J. (2008). *Automatic validation and optimisation of biological models*. Phd thesis, University of Oxford.

Copi, I. M., Cohen, C., and Flage, D. E. (2016). *Essentials of logic*. Taylor & Francis, Didcot, United Kingdom.

Cramér, H. (1928). On the composition of elementary errors. *Scandinavian Actuarial Journal*, 1928(1):13–74.

Cronbach, L. J. and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302.

Crookall, D. (2010). Serious games, debriefing, and simulation/gaming as a discipline. *Simulation & Gaming*, 41(6):898–920.

Darling, D. A. (1957). The Kolmogorov-Smirnov, Cramer-von Mises tests. *The Annals of Mathematical Statistics*, 28(4):823–838.

Davenport, T. H. and Prusak, L. (1998). *Working knowledge: How organizations manage what they know.* Harvard Business School Press, Boston, MA, USA.

de Bruijn, H. and ten Heuvelhof, E. (2008). *Management in networks: On multi-actor decision making.* Routledge, London, United Kingdom.

De Weck, O. L., Roos, D., and Magee, C. L. (2011). Life-cycle properties of engineering systems: The ilities. In *Engineering Systems: Meeting Human Needs in a Complex Technological World*, chapter 4, pages 65–96. Massachusetts Institute of Technology.

Decker, S., Fey, M., Sideras, S., Caballero, S., Rockstraw, L. R., Boese, T., Franklin, A. E., Gloe, D., Lioce, L., Sando, C. R., Meakim, C., and Borum, J. C. (2013). Standards of best practice: Simulation standard VI: The debriefing process. *Clinical Simulation in Nursing*, 9(6):S26–S29.

DeLaurentis, D. A. (2005). Understanding transportation as a system-of-systems design problem. In *43rd AIAA Aerospace Sciences Meeting and Exhibit*, pages 1–14, Reno, Nevada, USA. American Institute of Aeronautics and Astronautics.

Denzin, N. and Lincoln, Y. (1994). *Handbook of qualitative research.* Sage Publications, Inc., New York, USA, 1st edition.

Der Sahakian, G., Alinier, G., Savoldelli, G., Oriot, D., Jaffrelot, M., and Lecomte, F. (2015). Setting conditions for productive debriefing. *Simulation & Gaming*, 46(2):197–208.

Dewey, J. (1966). *Democracy and education: An introduction to the philosophy of education.* Free Press, New York, USA.

Dewey, J. (1986). Experience and education. *The Educational Forum*, 50(3):241–252.

Di Benedetto, M. D., Di Gennaro, S., and D'Innocenzo, A. (2007). Diagnosability verification for hybrid automata. In Bemporad, A., Bicchi, A., and Buttazzo, G., editors, *Hybrid Systems: Computation and Control*, pages 684–687. Springer, Berlin, Heidelberg.

Dienes, Z. and Perner, J. (1996). Implicit knowledge in people and connectionist networks. In Underwood, G. D. M., editor, *Implicit cognition*, pages 227–255. New York, NY, US: Oxford University Press.

Dijkstra, E. W. (1976). Guarded commands, non-determinacy and a calculus for the derivation of programs. In Bauer et al., F., editor, *Language Hierarchies and Interfaces. Lecture Notes in Computer Science, vol 46*, pages 111–124. Springer, Berlin, Heidelberg.

Dismukes, K. and Smith, G. M. (2000). *Facilitation and debriefing in aviation training and operations.* Routledge, London, United Kingdom, 1st edition.

Dixon, N. M. (2000). *Common knowledge: How companies thrive by sharing what they know.* Harvard Business School Press, Boston, MA, USA.

DoD-5000.59-M (1997). *DoD Modeling and Simulation (M&S) Glossary.* U.S. Department of Defense.

Dufrene, C. and Young, A. (2014). Successful debriefing - Best methods to achieve positive learning outcomes: A literature review. *Nurse Education Today*, 34(3):372–376.

Duke, R. D. (1974). *Gaming: The future's language.* Sage Publications, New York, USA.

Duke, R. D. (1980). A paradigm for game design. *Simulation & Gaming*, 11(3):364–377.

Duke, R. D. and Geurts, J. L. A. (2004). *Policy games for strategic management.* Dutch University Press, Amsterdam, The Netherlands.

Dunn, R. H. (1987). The quest for software reliability. In Schulmeyer, G. G. and McManus, J. I., editors, *Handbook of Software Quality Assurance*, pages 342–384. Van Nostrand Reinhold Company Limited, New York, USA.

Durbin, J. and Watson, G. S. (1950). Testing for serial correlation in least squares regression: I. *Biometrika*, 37(3/4):409–428.

Dyregrov, A. (1997). The process in psychological debriefings. *Journal of Traumatic Stress*, 10(4):589–605.

Easterbrook, S. (1991). *Elicitation of requirements from multiple perspectives.* PhD thesis, Imperial College of Science, Technology and Medicine, University of London.

Easterby-Smith, M. and Lyles, M. A. (2011). *Handbook of organizational learning and knowledge management.* John Wiley & Sons, New York, USA, second edition.

Eden, C., Jones, S., Sims, D., and Smithin, T. (1981). The intersubjectivity of issues and issues of intersubjectivity. *Journal of Management Studies*, 18(1):37–47.

Eisenhart, M. and Howe, K. (1992). Validity in educational research. In LeCompte, M., Millroy, W., and Preissle, J., editors, *The Handbook of Qualitative Research in Education*, chapter 14, pages 642–680. San Diego: Academic Press.

Emshoff, J. R. and Sisson, R. L. (1970). *Design and use of computer simulation models.* MacMillan, New York, USA.

Fanning, R. M. and Gaba, D. M. (2007). The role of debriefing in simulation-based learning. *Simulation in Healthcare*, 2(2):115–125.

Feld, S. L. (1997). Simulation games in theory development. *Sociological Forum*, 12(1):103–115.

Feller, W. (2015). The fundamental limit theorems in probability. In *Selected Papers I*, pages 667–699. Springer International Publishing, New York, USA.

Fernandez, A., Insfran, E., Abrahão, S., Carsí, J. Á., and Montero, E. (2012). Integrating usability evaluation into model-driven video game development. In *International Conference on Human-Centred Software Engineering*, pages 307–314. Springer, Berlin, Heidelberg.

Fisher, R. A. (1924). The conditions under which χ^2 measures the discrepancey between observation and hypothesis. *Journal of the Royal Statistical Society*, 87(3):442–450.

Fitzsimmons, J. A. (1974). The use of spectral analysis to validate planning models. *Socio-Economic Planning Sciences*, 8(3):123–128.

Forrester, J. W. (1971). *World dynamics*. Wright-Allen Press, Lawrence, KS, USA.

Forrester, J. W. and Senge, P. M. (1980). Tests for building confidence in system dynamics models. In *System Dynamics, TIMS Studies in Management Sciences, 14*, pages 209–228. North-Holland Publishing Company, Amsterdam, The Netherlands.

Fox, C., Levitin, A., and Redman, T. (1994). The notion of data and its quality dimensions. *Information Processing & Management*, 30(1):9–19.

Fullerton, T. (2004). *Game design workshop: A playcentric approach to creating innovative games*. Morgan Kaufmann Publishers, Burlington, MA, USA, 2nd edition.

Fumarola, M., van Staalduinen, J.-P., and Verbraeck, A. (2012). A ten-step design method for simulation games in logistics management. *Journal of Computing and Information Science in Engineering*, 12(1):1–6.

Garris, R., Ahlers, R., and Driskell, J. E. (2002). Games, motivation, and learning: A research and practice model. *Simulation & Gaming*, 33(4):441–467.

Gass, S. I. (1983). Decision-aiding models: Validation, assessment, and related issues for policy analysis. *Operations Research*, 31(4):603–631.

Glejser, H. (1969). A new test for heteroskedasticity. *Journal of the American Statistical Association*, 64(325):316.

Goeree, J. K. and Holt, C. A. (1999). Stochastic game theory: For playing games, not just for doing theory. *Proceedings of the National Academy of Sciences*, 96(19):10564–10567.

Goguen, J. A. (1994). Requirements engineering as the reconciliation of technical and social issues. In Jirotka, M. and Goguen, J. A., editors, *Requirements Engineering: Social and Technical Issues*, pages 1–31. Academic Press Professional, Inc., San Diego, CA, USA.

Gold, A. H., Malhotra, A., and Segars, A. H. (2001). Knowledge management: An organizational capabilities perspective. *Journal of Management Information Systems*, 18(1):185.

Grant, J. S., Dawkins, D., Molhook, L., Keltner, N. L., and Vance, D. E. (2014). Comparing the effectiveness of video-assisted oral debriefing and oral debriefing alone on behaviors by undergraduate nursing students during high-fidelity simulation. *Nurse education in practice*, 14(5):479–84.

Greenberg, S. and Buxton, B. (2008). Usability evaluation considered harmful (some of the time). In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 111–120. ACM.

Grobler, T., Louwrens, C., and von Solms, S. H. (2010). A framework to guide the implementation of proactive digital forensics in organisations. In *2010 International Conference on Availability, Reliability and Security*, pages 677–682, Krakow, Poland. IEEE.

Grogan, P. T. and Meijer, S. A. (2017). Gaming methods in engineering systems research. *Systems Engineering*, 20(6):542–552.

Guardiola, E. and Natkin, S. (2005). Game theory and video game, a new approach of game theory to analyze and conceive game systems. In Mehdi, Q., Gough, N., and Natkin, S., editors, *CGAMES'05, International Conference on Computer Games*, pages 166–170. Wolverhampton : School of Computing and Information Technology.

Gurteen, D. (1998). Knowledge, creativity and innovation. *Journal of Knowledge Management*, 2(1):5–13.

Hahs-Vaughn, D. L. and Lomax, R. G. (2013). *An introduction to statistical concepts*. Routledge, London, United Kingdom.

Hansen, M. T., Nohria, N., and Tierney, T. (1999). What's your strategy for managing knowledge? In Woods, J. A. and James, C., editors, *The Knowledge Management Yearbook 2000-2001*, pages 1–10. Butterworth-Heinemann, Oxford.

Harteveld, C. (2011). *Triadic game design: Balancing reality, meaning and play*. Springer Publishing Company, New York, USA, 1st edition.

Helbing, D. (1994). A mathematical model for the behavior of individuals in a social field. *The Journal of Mathematical Sociology*, 19(3):189–219.

Helbing, D. and Balietti, S. (2011). From social simulation to integrative system design. *The European Physical Journal Special Topics*, 195(1):69–100.

Henry, S. L., Abou-Zahra, S., and Brewer, J. (2014). The role of accessibility in a universal web. In *Proceedings of the 11th Web for All Conference*, pages 1–4, New York, New York, USA. ACM.

Hermann, C. F. (1967). Validation problems in games and simulations with special reference to models of international politics. *Behavioral Science*, 12(3):216–231.

Herz, J. C. and Macedonia, M. R. (2002). Computer games and the military: Two views. *Defense Horizons*, 11-12.

Heubusch, K. (2006). Interoperability: What it means, why it matters. *Journal of American Health Information Management Association*, 77(1):26–30.

Hofmann, M. A. (2015). Reasoning beyond predictive validity: The role of plausibility in decision-supporting social simulation. In Yilmaz, L., Chan, W. K. V., Moon, I., Roeder, T. M. K., Macal, C., and Rossetti, M. D., editors, *Proceedings of the 2015 Winter Simulation Conference*, pages 2728–2739, Huntington Beach, CA, USA. IEEE.

Hofstede, G. J. and Meijer, S. A. (2007). Collecting empirical data with games. In Mayer, I. S. and Mastik, H., editors, *Organisation and Learning Through Gaming and Simulation*, pages 111–121.

Holland, J. H. (1992). Complex adaptive systems. *Daedalus*, 121(1):17–30.

Holland, J. H. (1995). *Hidden order: How adaptation builds complexity*. Basic Books, New York, USA.

Holland, J. H. (2006). Studying complex adaptive systems. *Journal of Systems Science and Complexity*, 19(1):1–8.

Hotelling, H. (1992). The generalization of Student's ratio. In *Breakthroughs in Statistics*, pages 54–65. Springer, New York, USA.

Hsu, I.-C. (2008). Knowledge sharing practices as a facilitating factor for improving organizational performance through human capital: A preliminary test. *Expert Systems with Applications*, 35(3):1316–1326.

Huang, Y. (2013). *Automated simulation model generation*. PhD thesis, Delft University of Technology.

Huizinga, J. (1949). *Homo ludens: A study of the play-element in culture*. Routledge & Kegan Paul, Ltd., London, United Kingdom.

Husebø, S. E., Dieckmann, P., Rystedt, H., Søreide, E., and Friberg, F. (2013). The relationship between facilitators' questions and the level of reflection in postsimulation debriefing. *Simulation in Healthcare*, 8(3):135–142.

IEEE (2011). Systems and software engineering – Vocabulary.

ISO/IEC-9241-11 (2018). Ergonomics of human-system interaction. *Part 11: Usability: Definitions and Concepts*.

Issenberg, S. B., Mcgaghie, W. C., Petrusa, E. R., Gordon, D. L., and Scalese, R. J. (2005). Features and uses of high-fidelity medical simulations that lead to effective learning: A BEME systematic review. *Medical Teacher*, 27(1):10–28.

Jahangirian, M., Eldabi, T., Naseer, A., Stergioulas, L. K., and Young, T. (2010). Simulation in manufacturing and business: A review. *European Journal of Operational Research*, 203(1):1–13.

Janssen, M. and De Vries, B. (1999). Global modelling: Managing uncertainty, complexity and incomplete information. In van Dijkum, C., DeTombe, D., and van Kuijk, E., editors, *Validation of Simulation Models*, pages 45–69, Amsterdam, The Netherlands. SISWO.

Jarrar, Y. F. (2002). Knowledge management: Learning for organisational experience. *Managerial Auditing Journal*, 17(6):322–328.

Jenkins, G. M. and Watts, D. G. (1969). *Spectral analysis and its applications*. Holden-Day, San Fransisco, CA, USA.

Jo, J., Yoon, S., Yoo, J., Lee, H. Y., and Kim, W.-T. (2012). Case study: Verification of ECML model using SpaceEx. In *Korea-Japan Joint Workshop on ICT*, pages 1–4, Pohang, South Korea.

Johnson, R. B. and Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, 33(7):14–26.

Jolliffe, I. T. (1986). *Principal Component Analysis and Factor Analysis*. Springer, New York, USA.

Kasser, J. (2002). The cataract methodology for systems and software acquisition. In *Systems Engineering, Test & Evaluation Conference*, Sydney, Australia.

Katsaliaki, K. and Mustafee, N. (2012). A survey of serious games on sustainable development. In Laroque, C., Himmelspach, J., Pasupathy, R., Rose, O., and Uhrmacher, A., editors, *Proceedings of the 2012 Winter Simulation Conference*, pages 300–312, Berlin, Germany. IEEE.

Kavitha, K. R. and Ahmed, I. M. S. (2011). A knowledge management framework for agile software development teams. In *2011 International Conference on Process Automation, Control and Computing (PACC)*, pages 1–5, Coimbatore, India. IEEE.

Keith, C. (2010). *Agile game development with Scrum*. Addison-Wesley Professional, Massachusetts, USA.

Kimbrough, S. O., Wu, D., and Zhong, F. (2002). Computers play the beer game: Can artificial agents manage supply chains? *Decision Support Systems*, 33(3):323–333.

King, J. C. (1976). Symbolic execution and program testing. *Communications of the ACM*, 19(7):385–394.

Klabbers, J. H. G. (2009). *The magic circle: Principles of gaming & simulation*. Sense Publishers, Rotterdam, The Netherlands, 3rd edition.

Klabbers, J. H. G. (2018). On the architecture of game science. *Simulation & Gaming*, 49(3):207–245.

Kleijnen, J. P. (1995). Verification and validation of simulation models. *European Journal of Operational Research*, 82(1):145–162.

Klemeš, V. (1986). Operational testing of hydrological simulation models. *Hydrological Sciences Journal*, 31(1):13–24.

Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development*. Prentice Hall, Inc., Upper Saddle River, NJ, USA.

Kolfschoten, G. (2007). *Theoretical foundations for collaboration engineering*. PhD thesis, Delft University of Technology, Delft, The Netherlands.

Koppenjan, J. and Klijn, E.-H. (2004). *Managing uncertainties in networks*. Routledge, London, United Kingdom, 1st edition.

Kortmann, R. and Sehic, E. (2010). The Railway Bridge Game - usability, usefulness, and potential usage for railways management. In *Proceedings of the International Simulation and Gaming Association (ISAGA)*.

Kriz, W. C. (2003). Creating effective learning environments and learning organizations through gaming simulation design. *Simulation & Gaming*, 34(4):495–511.

Kriz, W. C. (2010). A systemic-constructivist approach to the facilitation and debriefing of simulations and games. *Simulation & Gaming*, 41(5):663–680.

Kuhl, F., Weatherly, R., and Dahmann, J. (1999). *Creating computer simulation systems: An introduction to the high level architecture*. Prentice Hall PTR, Upper Saddle River, NJ, USA.

Kuiper, N. H. (1960). Tests concerning random points on a circle. *Indagationes Mathematicae*, 63:38–47.

Kum, D.-h., Son, J., Lee, S.-b., and Wilson, I. (2006). Automated testing for automotive embedded systems. In *2006 SICE-ICASE International Joint Conference*, pages 4414–4418, Bexco, Busan, Korea. IEEE.

Landriscina, F. (2013). *Simulation and learning*. Springer, New York, USA.

Landry, M., Malouin, J.-L., and Oral, M. (1983). Model validation in operations research. *European Journal of Operational Research*, 14(3):207–220.

Latino, R. J., Latino, K. C., and Latino, M. A. (2016). *Root cause analysis: Improving performance for bottom-line results*. Taylor & Francis Group, Didcot, United Kingdom, 4th edition.

Law, A. M. (2008). How to build valid and credible simulation models. In Mason, S. J., Hill, R. R., Mönch, L., Rose, O., Jefferson, T., and Fowler, J. W., editors, *Proceedings of the 2008 Winter Simulation Conference*, pages 39–47, Miami, FL, USA. IEEE.

Lederman, L. C. (1991). Differences that make a difference: Intercultural communication, simulation, and the debriefing process. In Crookall, D. and Arai, K., editors, *Proceedings of the 22nd International Conference of the International Simulation and Gaming Association (ISAGA)*, pages 15–19, Tokyo, Japan. Springer.

Lederman, L. C. (1992). Debriefing: Toward a systematic assessment of theory and practice. *Simulation & Gaming*, 23(2):145–160.

Leigh, E. and Spindler, L. (2004). Simulations and games as chaordic learning contexts. *Simulation & Gaming*, 35(1):53–69.

Leininger, M. (1994). Evaluation criteria and critique of qualitative research studies. In Morse, J., editor, *Critical Issues in Qualitative Research Methods*, chapter 6, pages 95–115. Sage Publications, Inc., London, United Kingdom.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, 55(3):187–193.

Leonard, D. and Sensiper, S. (1998). The role of tacit knowledge in group innovation. *California Management Review*, 40(3):112–132.

Leonard-Barton, D. (1995). *Wellsprings of knowledge*. Harvard Business School Press, Boston, MA, USA.

Lethbridge, T. C., Sim, S. E., and Singer, J. (2005). Studying software engineers: Data collection techniques for software field studies. *Empirical Software Engineering*, 10(3):311–341.

Likas, A., Vlassis, N., and J. Verbeek, J. (2003). The global k-means clustering algorithm. *Pattern Recognition*, 36(2):451–461.

Lincoln, Y. S. and Guba, E. G. (1985). *Naturalistic inquiry*. SAGE Publications Inc., New York, USA.

Liu, D., Macchiarella, N. D., and Vincenzi, D. A. (2008). Simulation fidelity. In Hancock, P. A., Vincenzi, D. A., Wise, J. A., and Mouloua, M., editors, *Human Factors in Simulation and Training*, chapter 4, page 453. CRC Press, Florida, USA.

Lo, J., van den Hoogen, J., and Meijer, S. (2013). Using gaming simulation experiments to test railway innovations: Implications for validity. In Pasupathy, R., Kim, S.-H., Tolk, A., Hill, R., and Kuhl, M. E., editors, *Proceedings of the 2013 Winter Simulation Conference*, pages 1766–1777, Washington, D.C., USA. IEEE Press.

Lo, J. C. and Meijer, S. (2014). Gaming simulation design for individual and team situation awareness. In *Frontiers in Gaming Simulation*, pages 121–128. Springer International Publishing, New York, USA.

Lo, J. C., van den Hoogen, J., and Meijer, S. A. (2014). Testing changes in the railway system through gaming simulation: How different types of innovations affect operators' mental models. *Advances in Human Aspects of Transportation: Part III*, 9:291–302.

Locke, J. (1712). *Some thoughts concerning education*. National Society's Depository, London, United Kingdom.

Lomax, R. G. and Hahs-Vaughn, D. L. (2013). *Statistical concepts: A second course*. Taylor & Francis Group, Didcot, United Kingdom.

Louie, M. A. and Carley, K. M. (2008). Balancing the criticisms: Validating multi-agent models of social systems. *Simulation Modelling Practice and Theory*, 16(2):242–256.

Mader, S., Natkin, S., and Levieux, G. (2012). How to analyse therapeutic games: The player/game/therapy model. In Herrlich, M., Malaka, R., and Masuch, M., editors, *International Conference on Entertainment Computing*, pages 193–206, Bremen, Germany. Springer Berlin Heidelberg.

Maier, R. (2007). *Knowledge management systems: Information and communication technologies for knowledge management.* Springer, New York, USA, 3rd edition.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60.

Markíczy, L. and Goldberg, J. (1995). A method for eliciting and comparing causal maps. *Journal of Management*, 21(2):305–333.

Markus, L. M. (2001). Toward a theory of knowledge reuse: Types of knowledge reuse situations and factors in reuse success. *Journal of Management Information Systems*, 18(1):57–93.

Martin, J. (1982). Stories and scripts in organizational settings. In Hastorf, A. H. and Isen, A. M., editors, *Cognitive Social Psychology*, pages 255–305. New York: Elsevier.

Martis, M. S. (2006). Validation of simulation based models: A theoretical outlook. *Electronic Journal of Business Research Methods*, 4(1):39–46.

Mavin, A. and Maiden, N. (2003). Determining socio-technical systems requirements: Experiences with generating and walking through scenarios. In *Proceedings of the 11th IEEE International Conference on Requirements Engineering*, pages 213–222. IEEE Comput. Soc.

Mayo, E. (1946). *The human problems of an industrial civilization.* Macmillan Co., New York, USA, 2nd edition.

McCausland, L. L., Curran, C. C., and Cataldi, P. (2004). Use of a human simulator for undergraduate nurse education. *International Journal of Nursing Education Scholarship*, 1(1).

Meadows, D. H., Meadows, D. L., Randers, J., and Behrens III, W. W. (1972). *The limits to growth.* Universe Books, New York, USA.

Meadows, D. L., Behrens III, W. W., Meadows, D. H., Naill, R. F., Randers, J., and Zahn, E. (1974). *Dynamics of growth in a finite world.* Wright-Allen Press, Cambridge, MA, USA.

Meijer, S. (2009). *The organisation of transactions.* PhD thesis, Wageningen University, Wageningen, The Netherlands.

Meijer, S. (2012a). Gaming simulations for railways: Lessons learned from modeling six games for the Dutch infrastructure management. In Perpinya, X., editor, *Infrastructure Design, Signaling and Security in Railway*, chapter 11, pages 275–294. InTech.

Meijer, S. (2012b). Introducing gaming simulation in the Dutch railways. *Procedia - Social and Behavioral Sciences*, 48:41–51.

Meijer, S. (2015). The power of sponges: Comparing high-tech and low-tech gaming for innovation. *Simulation & Gaming*, 46(5):512–535.

Meng, R., Ye, Y., and Xie, N.-g. (2010). Multi-objective optimization design methods based on game theory. In *8th World Congress on Intelligent Control and Automation (WCICA)*, pages 2220–2227, Jinan, China. IEEE.

Merriam-Webster (2018). *Definition of system.* Springfield, MA, USA.

Metaxiotis, K., Ergazakis, K., and Psarras, J. (2005). Exploring the world of knowledge management: Agreements and disagreements in the academic/practitioner community. *Journal of Knowledge Management*, 9(2):6–18.

Michael, D. R. and Chen, S. L. (2005). *Serious games: Games that educate, train, and inform.* Thomson Course Technology PTR, Massachusetts, USA.

Middelkoop, D., Meijer, S., Steneker, J., Sehic, E., and Mazzarello, M. (2012). Simulation backbone for gaming simulation in railways: A case study. In Laroque, C., Himmelspach, J., Pasupathy, R., Rose, O., and Uhrmacher, A., editors, *Proceedings of the 44th Conference on Winter Simulation*, pages 3262–3274, Berlin. IEEE.

Middelkoop, D. A. and Bouwman, M. (2001). Simone: Large scale train network simulations. In Peters. B. A., Smith, J. S., Medeiros, D. J., and Rohrer, M. W., editors, *Proceedings of the 33nd Conference on Winter Simulation*, pages 1042–1047, Arlington, Virginia, USA. ACM.

Middelkoop, D. A. and Loeve, L. (2006). Simulation of traffic management with FRISO. *WIT Transactions on the Built Environment*, 88:501–509.

Miller, L. A., Groundwater, E., and Mirsky, S. M. (1993). Survey and assessment of conventional software verification and validation methods. Technical report, Electric Power Research Institute, Palo Alto, CA, USA.

Morgan, P. J., Cleave-Hogg, D., DeSousa, S., and Tarshis, J. (2004). High-fidelity patient simulation: Validation of performance checklists. *British Journal of Anaesthesia*, 92(3):388–392.

Morrison, J. E. and Meliza, L. L. (1999). Foundations of the after action review process. Technical report, United States Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA, USA.

Moss, S. (2001). Game theory: Limitations and an alternative. *Journal of Artificial Societies and Social Simulation*, 4(2).

Musen, M. A. (1992). Dimensions of knowledge sharing and reuse. *Computers and Biomedical Research*, 25(5):435–467.

Myers, G. J., Badgett, T., Thomas, T. M., and Sandler, C. (2011). *The art of software testing*. John Wiley & Sons, Inc., New York, USA.

Myerson, R. B. (1997). *Game theory: Analysis of conflict*. Harvard University Press, Boston, MA, USA.

Mylopoulos, J., Chung, L., and Yu, E. (1999). From object-oriented to goal-oriented requirements analysis. *Communications of the ACM*, 42(1):31–37.

Nash, A. and Huerlimann, D. (2004). Railroad simulation using OpenTrack. *WIT Transactions on The Built Environment*, 74:45–54.

Naylor, T. H. and Finger, J. M. (1967). Verification of computer simulation models. *Management Science*, 14(2):B–92–B–101.

Nehring, W. M. and Lashley, F. R. (2004). Current use and opinions regarding human patient simulators in nursing education: An international survey. *Nursing Education Perspectives*, 25(5):244–248.

Neill, M. A. and Wotton, K. (2011). High-fidelity simulation debriefing in nursing education: A literature review. *Clinical Simulation in Nursing*, 7(5):e161–e168.

Nemani, R. R. and Running, S. W. (1989). Testing a theoretical climate-soil-leaf area hydrologic equilibrium of forests using satellite data and ecosystem simulation. *Agricultural and Forest Meteorology*, 44(3-4):245–260.

Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 231:289–337.

Nielsen, J. (2003). Usability 101: Introduction to usability.

Nilsson, F. and Darley, V. (2006). On complex adaptive systems and agent-based modelling for improving decision-making in manufacturing and logistics settings. *International Journal of Operations & Production Management*, 26(12):1351–1373.

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization Science*, 5(1):14–37.

Nonaka, I. and von Krogh, G. (2009). Perspective - Tacit knowledge and knowledge conversion: Controversy and advancement in organizational knowledge creation theory. *Organization Science*, 20(3):635–652.

Nordhaus, W. D. (1973). World dynamics: Measurement without data. *The Economic Journal*, 83(332):1156–1183.

Osborne, M. J. and Rubinstein, A. (1994). *A Course in game theory*. The MIT Press, Cambridge, MA, USA.

Overstreet, M. C. and Nance, R. E. (2004). Characterizations and relationships of world views. In Ingalls, R. G., Rossetti, M. D., Smith, J. S., and Peters, B. A., editors, *Proceedings of the 36th Conference on Winter Simulation*, pages 279–287, Washington D.C., U.S.A. ACM.

Pace, D. K. (2004). Modeling and simulation verification and validation challenges. *Johns Hopkins APL Technical Digest*, 25(2):163–172.

Panahi, S., Watson, J., and Partridge, H. (2012). Social media and tacit knowledge sharing: Developing a conceptual model. In *World Academy of Science, Engineering and Technology (WASET)*, pages 1095–1102, Paris, France.

Pawlowski, J. and Bick, M. (2012). The global knowledge management framework: Towards a theory for knowledge management in globally distributed settings. *Electronic Journal of Knowledge Management*, 10(1):92–108.

Pearson, M. and Smith, D. (1985). Debriefing in experience-based learning. In Boud, D., Keogh, R., and Walker, D., editors, *Reflection: Turning Experience Into Learning*, chapter 4, pages 69–84. Routledge, London, United Kingdom.

Perry, W. E. (2007). *Effective methods for software testing.* Wiley Publishing Inc., New York, USA.

Peters, V., Vissers, G., and Heijne, G. (1998). The validity of games. *Simulation & Gaming*, 29(1):20–30.

Peters, V. M. and Vissers, G. N. (2004). A simple classification model for debriefing simulation games. *Simulation & Gaming*, 35(1):70–84.

Petranek, C. F. (1994). A maturation in experiential learning: Principles of simulation and gaming. *Simulation & Gaming*, 25(4):513–523.

Petranek, C. F., Corey, S., and Black, R. (1992). Three levels of learning in simulations: Participating, debriefing, and journal writing. *Simulation & Gaming*, 23(2):174–185.

Petrash, G. (1996). Dow's journey to a knowledge value management culture. *European Management Journal*, 14(4):365–373.

Piaget, J. (1952). *Play, dreams and imitation in childhood.* W. W. Norton & Co., New York, USA.

Pipino, L. L., Lee, Y. W., and Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, 45(4):211–218.

Pirró, G., Mastroianni, C., and Talia, D. (2010). A framework for distributed knowledge management: Design and implementation. *Future Generation Computer Systems*, 26(1):38–49.

Plato (1866). Book 7: 537a. In Davies, J. L. and Vaughan, D. J., editors, *Republic (Translated into English)*. London : Macmillan, 3rd edition.

Porzycki, J., Lubaś, R., Mycek, M., and Wąs, J. (2015). Dynamic data - driven simulation of pedestrian movement with automatic validation. In Chraibi, M., Boltes, M., Schadschneider, A., Seyfried, A., editor, *Traffic and Granular Flow'13*, pages 129–136. Springer, Cham, Switzerland.

Power, D. J. (2004). Specifying an expanded framework for classifying and describing decision support systems. *Communications of the Association for Information Systems Volume*, 13:158–166.

Pressman, R. S. (2015). *Software engineering: A practitioner's approach (8th edition)*. McGraw-Hill, New York, USA.

Price, R. and Shanks, G. (2005). A semiotic information quality framework: Development and comparative analysis. *Journal of Information Technology*, 20(2):88–102.

Raemer, D., Anderson, M., Cheng, A., Fanning, R., Nadkarni, V., and Savoldelli, G. (2011). Research regarding debriefing as part of the learning process. *Simulation in Healthcare*, 6(7):S52–S57.

Ragab, M. A. F. and Arisha, A. (2013). Knowledge management and measurement: A critical review. *Journal of Knowledge Management*, 17(6):873–901.

Raghothama, J., Hanchi, H., and Meijer, S. (2017). Hybrid, composable approach to simulations in healthcare operations and management. In Chan, W. K. V., D'Ambrogio, A., Zacharewicz, G., Mustafee, N., Wainer, G., and Page, E., editors, *Proceedings of the 2017 Winter Simulation Conference*, pages 2857–2868, Las Vegas, NV, USA. IEEE.

Raser, J. R. (1969). *Simulation and society: An exploration of scientific gaming'*. Allyn and Bacon, Boston, MA, USA.

Rasmusen, E. (2007). *Games and information: An introduction to game theory (4th edition)*. Blackwell Publishing, Hoboken, NJ, USA.

Rattray, C. (1990). *Specification and verification of concurrent systems*. Springer-Verlag, London, United Kingdom.

Reis, H. T. and Judd, C. M. (2014). *Handbook of research methods in social and personality psychology*. Cambridge University Press, Cambridge, United Kingdom, 2nd edition.

Rentsch, T. (1982). Object oriented programming. *ACM SIGPLAN Notices*, 17(9):51–57.

Rinaldi, S. M., Peerenboom, J. P., and Kelly, T. K. (2001). Identifying, understanding, and analyzing critical infrastructure interdependencies. *IEEE Control Systems*, 21(6):11–25.

Ritterfeld, U., Cody, M., and Vorderer, P. (2009). *Serious games: Mechanisms and effects*. Routledge, London, United Kingdom.

Rodríguez Rocha, O. and Faron-Zucker, C. (2015). Ludo: An ontology to create linked data driven serious games. In *ISWC 2015-Workshop on LINKed EDucation, LINKED 2015*, Bethlehem, Pennsylvania, United States.

Rooney, J. J. and Van den Heuvel, L. N. (2004). Root Cause Analysis For Beginners. *Quality Progress*, 37(7):45–53.

Ropohl, G. (1999). Philosophy of Socio-Technical Systems. *Techné: Research in Philosophy and Technology*, 4(3):186–194.

Roungas, B., Bekius, F., Meijer, S., and Verbraeck, A. (2019a). Improving the decision-making qualities of gaming simulations. *Journal of Simulation. Under Review.*

Roungas, B., Bekius, F. A., and Meijer, S. (2019b). The game between game theory and gaming simulations: Design choices. *Simulation & Gaming*, 50(2).

Roungas, B. and Dalpiaz, F. (2016). A model-driven framework for educational game design. In De Gloria, A. and Veltkamp, R., editors, *4th International Conference, GALA 2015, Revised Selected Papers*, volume 9599, pages 1–11, Rome, Italy. Springer International Publishing.

Roungas, B., De Wijse, M., Meijer, S., and Verbraeck, A. (2018a). Pitfalls for debriefing games and simulation: Theory and practice. In Naweed, A., Wardaszko, M., Leigh, E., and Meijer, S., editors, *Intersections in Simulation and Gaming*, pages 101–115. Springer, Cham, Switzerland.

Roungas, B., Lo, J. C., Angeletti, R., Meijer, S. A., and Verbraeck, A. (2018b). Eliciting requirements of a knowledge management system for gaming in an organization: The role of tacit knowledge. In *49th International Conference of International Simulation and Gaming Association*, Bangkok, Thailand.

Roungas, B., Lo, J. C., Meijer, S., and Verbraeck, A. (2018c). Guidelines for the management and dissemination of knowledge from gaming simulations. *Simulation & Gaming. Under Review.*

Roungas, B., Meijer, S., and Verbraeck, A. (2017). A framework for simulation validation & verification method selection. In *SIMUL 2017: The Ninth International Conference on Advances in System Simulation*, pages 35–40, Athens, Greece.

Roungas, B., Meijer, S., and Verbraeck, A. (2018d). Knowledge management of games for decision making. In Lukosch, H., Bekebrede, G., and Kortmann, R., editors, *Simulation Gaming. Applications for Sustainable Cities and Smart Infrastructures. ISAGA 2017. Lecture Notes in Computer Science, vol 10825*, Cham, Switzerland. Springer.

Roungas, B., Meijer, S., and Verbraeck, A. (2018e). Validity of railway microscopic simulations under the microscope: Two case studies. *International Journal of System of Systems Engineering*, 8(4):346–364.

Roungas, B., Meijer, S., and Verbraeck, A. (2019c). The tacit knowledge in games: From validation to debriefing. In *50th International Conference of International Simulation and Gaming Association*, Warsaw, Poland.

Roungas, B., Meijer, S. A., and Verbraeck, A. (2018f). A framework for optimizing simulation model validation & verification. *International Journal On Advances in Systems and Measurements*, 11(1 & 2).

Roungas, B., Meijer, S. A., and Verbraeck, A. (2018g). Harnessing Web 3.0 and R to mitigate simulation validation restrictions. In *International Conference on Simulation and Modeling Methodologies, Technologies and Applications*, Porto, Portugal.

Roungas, B., Meijer, S. A., and Verbraeck, A. (2018h). The future of contextual knowledge in gaming simulations: A research agenda. In Rabe, M., Juan, A. A., Mustafee, N., Skoogh, A., Jain, S., and Johansson, B., editors, *Proceedings of the 2018 Winter Simulation Conference*, Gothenburg, Sweden. IEEE.

Rouse, W. B. (2008). Health care as a complex adaptive system: Implications for design and management. *Bridge-Washington-National Academy of Engineering*, 38(1):17–25.

Rousseau, J.-J. (1979). *Emile or on education. Trans. Allan Bloom.* Basic Books, New York, USA.

Rudolph, J. W., Simon, R., Dufresne, R. L., and Raemer, D. B. (2006). There's no such thing as "nonjudgmental" debriefing: A theory and method for debriefing with good judgment. *Simulation in Healthcare*, 1(1):49–55.

Rudolph, J. W., Simon, R., Raemer, D. B., and Eppich, W. J. (2008). Debriefing as formative assessment: Closing performance gaps in medical education. *Academic Emergency Medicine : Official Journal of the Society for Academic Emergency Medicine*, 15(11):1010–6.

Saadawi, H. and Wainer, G. (2009). Verification of real-time DEVS models. In *Proceedings of the 2009 Spring Simulation Multiconference*, page 143, San Diego, CA, USA. Society for Computer Simulation International.

Salen, K. and Zimmerman, E. (2004). *Rules of play: Game design fundamentals.* The MIT Press, Cambridge, MA, USA.

Sargent, R. G. (1996). Verifying and validating simulation models. In Charnes, J. M., Morrice, D. J., Brunner, D. T., and Swain, J. J., editors, *Proceedings of the 28th Conference on Winter Simulation*, pages 55–64, Coronado, California, USA. IEEE Computer Society.

Sargent, R. G. (2000). Verification, validation, and accreditation of simulation models. In Joines, J. A., Barton, R. R., Kang, K., and Fishwick, P. A., editors, *Proceedings of the 32nd Conference on Winter Simulation*, pages 50–59, Orlando, Florida, USA. IEEE.

Sargent, R. G. (2005). Verification and validation of simulation models. In Chick, S., Sánchez, P. J., Ferrin, D., and Morrice, D. J., editors, *Proceedings of the 37th Conference on Winter Simulation*, pages 130–143, Orlando, Florida, USA. Winter Simulation Conference.

Sauro, J. and Kindlund, E. (2005). A method to standardize usability metrics into a single score. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 401–409, Portland, Oregon, USA. ACM.

Savoldelli, G. L., Naik, V. N., Park, J., Joo, H. S., Chow, R., and Hamstra, S. J. (2006). Value of Debriefing during Simulated Crisis Management: Oral versus video-assisted oral feedback. *Anesthesiology*, 105(2):279–285.

Sawyer, T. L. and Deering, S. (2013). Adaptation of the US Army's after-action review for simulation debriefing in healthcare. *Simulation in Healthcare*, 8(6):388–397.

Schach, S. R. (2011). *Classical and object-oriented software engineering (8th edition)*. McGraw-Hill, New York, USA.

Scherer, L. A., Chang, M. C., Meredith, J. W., and Battistella, F. D. (2003). Videotape review leads to rapid and sustained learning. *The American Journal of Surgery*, 185(6):516–520.

Schlesinger, S., Crosbie, R. E., Gagné, R. E., Innis, G. S., Lalwani, C. S., Loch, J., Sylvester, R. J., Wright, R. D., Kheir, N., and Bartos, D. (1979). Terminology for model credibility. *Simulation*, 32(3):103–104.

Schruben, L. W. (1980). Establishing the credibility of simulations. *Simulation*, 34(3):101–105.

Schwaber, K. (2004). *Agile project management with Scrum.* Microsoft Press, Redmond, WA, USA.

Seymour, N. E. (2012). Debriefing after simulation. In Tichansky, D. S., Morton, J., and Jones, D. B., editors, *The SAGES Manual of Quality, Outcomes and Patient Safety*, pages 501–506. Springer, Boston, MA, USA.

Shannon, R. and Johannes, J. D. (1976). Systems simulation: The art and science. *IEEE Transactions on Systems, Man, and Cybernetics*, 6(10):723–724.

Sheckley, B. G., Kehrhahn, M., Bell, S., and Grenier, R. S. (2008). Trio: An emerging model of adult professional learning. In *Proceedings of the 49th Annual Adult Education Research Conference*, St. Louis, MO, USA. University of Missouri.

Shinnick, M. A., Woo, M., Horwich, T. B., and Steadman, R. (2011). Debriefing: The most important component in simulation? *Clinical Simulation in Nursing*, 7(3):e105–e111.

Simon, H. A. (1957). *Models of man: Social and rational.* Wiley, Hoboken, NJ, USA, 1st edition.

Simon, H. A. (1972). Theories of bounded rationality. In McGuire, C. B. and Radner, R., editors, *Decision and Organization*, chapter 8, pages 161–176. North-Holland Publishing Company, Amsterdam, The Netherlands.

Skardi, M. J. E., Afshar, A., and Solis, S. S. (2013). Simulation-optimization model for non-point source pollution management in watersheds: Application of cooperative game theory. *KSCE Journal of Civil Engineering*, 17(6):1232–1240.

Smith, E. A. (2001). The role of tacit and explicit knowledge in the workplace. *Journal of Knowledge Management*, 5(4):311–321.

Sommerville, I. (2004). *Software engineering (9th edition)*. Addison-Wesley, Reading, MA, USA.

Spender, J. (1996). Organizational knowledge, learning and memory: Three concepts in search of a theory. *Journal of Organizational Change Management*, 9(1):63–78.

Spender, J.-C. (2008). Organizational learning and knowledge management: Whence and whither? *Management Learning*, 39(2):159–176.

Stephenson, W. (1953). *The study of behavior: Q-technique and its methodology.* The University of Chicago Press, Chicago, IL, USA.

Sterman, J. D. (1989). Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. *Management Science*, 35(3):321–339.

Sweeney, L. B. and Meadows, D. (2010). *The systems thinking playbook: Exercises to stretch and build learning and systems thinking capabilities.* Chelsea Green Publishing, Hartford, VT, USA, 1st edition.

Taleb, N. N. (2004). No, small probabilities are "not attractive to sell": A comment. *Journal of Behavioral Finance*, 5(1):2–7.

Tanenbaum, A. S., Klint, P., and Bohm, W. (1978). Guidelines for software portability. *Software: Practice and Experience*, 8(6):681–698.

Tang, S. and Hanneghan, M. (2011). Game content model: An ontology for documenting serious game design. In *Developments in E-systems Engineering (DeSE), 2011*, pages 431–436. IEEE.

Teisman, G. R. and Klijn, E.-H. (2008). Complexity theory and public management. *Public Management Review*, 10(3):287–297.

Thatcher, D. C. and Robinson, M. J. (1985). *An introduction to games and simulations in education.* Solent Simulations, Southampton, United Kingdom.

Thavikulwat, P. (2004). The architecture of computerized business gaming simulations. *Simulation & Gaming*, 35(2):242–269.

Theil, H. (1958). *Economic forecasts and policy.* North-Holland, Amsterdam, The Netherlands.

Thiagarajan, S. (1992). Using games for debriefing. *Simulation & Gaming*, 23(2):161–173.

Thompson, K. M. and Badizadegan, N. D. (2015). Valuing information in complex systems: An integrated analytical approach to achieve optimal performance in the beer distribution game. *IEEE Access*, 3:2677–2686.

Trist, E. L. (1981). The evolution of socio-technical systems. *Occasional paper*, 2.

Trist, E. L. and Bamforth, K. W. (1951). Some social and psychological consequences of the Longwall method of coal-getting. *Human Relations*, 4(1):3–38.

Tsoukas, H. (1991). The missing link: A transformational view of metaphors in organizational science. *Academy of Management Review*, 16(3):566–585.

Ursano, R. J. and Rundell, J. R. (1990). The prisoner of war. *Military Medicine*, 155(4):176–180.

van den Hoogen, J. (2019). *The gaming of systemic innovations: Innovating in the railway sector using gaming simulations.* PhD thesis, Delft University of Technology.

van den Hoogen, J., Lo, J., and Meijer, S. (2014). Debriefing in gaming simulation for research: Opening the black box of the non-trivial machine to assess validity and reliability. In Tolk, A., Diallo, S. Y., Ryzhov, I. O., Yilmaz, L., Buckley, S., and Miller, J. A., editors, *Proceedings of the 2014 Winter Simulation Conference*, pages 3505–3516, Savannah, Georgia, USA. IEEE Press.

Van den Hoogen, J. and Meijer, S. (2012). Deciding on innovation at a railway network operator: A grounded theory approach. In *CESUN 2012: 3rd International Engineering Systems Symposium*, Delft University of Technology, The Netherlands.

Van der Spek, R. and Spijkervet, A. (1997). Knowledge management: Dealing intelligently with knowledge. In Liebowitz, J. and Wilcox, L. C., editors, *Knowledge Management and its Integrative Elements*, chapter 2, pages 31–59. CRC Press, Boca Raton, FL, USA, 1st edition.

van Gunsteren, W. F. and Mark, A. E. (1998). Validation of molecular dynamics simulation. *The Journal of Chemical Physics*, 108(15):6109–6116.

van Lankveld, G., Sehic, E., Lo, J. C., and Meijer, S. A. (2017). Assessing gaming simulation validity for training traffic controllers. *Simulation & Gaming*, 48(2):219–235.

Vangheluwe, H., de Lara, J., and Mosterman, P. J. (2002). An introduction to multi-paradigm modelling and simulation. In Barros, F. and Giambiasi, N., editors, *Proceedings of the AIS'2002 Conference (AI, Simulation and Planning in High Autonomy Systems)*, pages 9–20, Lisboa, Portugal.

Veeke, H. P., Ottjes, J. A., and Lodewijks, G. (2008). *The Delft systems approach: Analysis and design of industrial systems.* Springer London.

Vollmer, H. (2013). What kind of game is everyday interaction? *Rationality and Society*, 25(3):370–404.

von Solms, S. H., Louwrens, C., Reekie, C., and Grobler, T. (2006). A control framework for digital forensics. *Advances in Digital Forensics II*, 222:343–355.

Vorobeychik, Y. and Wellman, M. P. (2009). Strategic analysis with simulation-based games. In Rossetti, M. D., Hill, R. R., Johansson, B., Dunkin, A., and Ingalls, R. G., editors, *Proceedings of the 2009 Winter Simulation Conference*, pages 359–372, Austin, TX, USA. IEEE.

Warne, R. T. (2014). A primer on multivariate analysis of variance (MANOVA) for behavioral scientists. *Practical Assessment, Research & Evaluation*, 19(17):1–10.

Watson, S. and Hewett, K. (2006). A multi-theoretical model of knowledge transfer in organizations: Determinants of knowledge contribution and knowledge reuse. *Journal of Management Studies*, 43(2):141–173.

Welch, B. L. (1947). The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28.

Whelan, J. G. (1994). Building the fish banks model and renewable resource depletion. Course material – MIT Sloan School of Management.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838.

Whitner, R. B. and Balci, O. (1989). Guidelines for selecting and using simulation model verification techniques. In MacNair, E., Musselman, K., and Heidelberger, P., editors, *Proceedings of the 21st Conference on Winter Simulation*, pages 559–568, Washington, D.C., USA. ACM.

Wiig, K. M. (1993). *Knowledge management foundations: Thinking about how people and organizations create, represent, and use knowledge.* Schema Press, Arlington, TX, USA.

Wiig, K. M. (1997). Knowledge management: An introduction and perspective. *Journal of Knowledge Management*, 1(1):6–14.

Wilkins, A. L. and Thompson, M. P. (1991). On getting the story crooked (and straight). *Journal of Organizational Change Management*, 4(3):18–26.

Yin, R. K. (2009). *Case study research: Design and methods.* Sage Publications, Inc., Thousand Oaks, CA, USA, 4th edition.

Yuan, J. and Hansen, I. A. (2007). Optimizing capacity utilization of stations by estimating knock-on train delays. *Transportation Research Part B: Methodological*, 41(2):202–217.

Zave, P. (1997). Classification of research efforts in requirements engineering. *ACM Computing Surveys (CSUR)*, 29(4):315–321.

Zeigler, B. P. and Sarjoughian, H. S. (2012). *Guide to modeling and simulation of systems of systems.* Springer, London, United Kingdom.

Zevin, B., Levy, J. S., Satava, R. M., and Grantcharov, T. P. (2012). A Consensus-based framework for design, validation, and implementation of simulation-based training curricula in surgery. *Journal of the American College of Surgeons*, 215(4):580–586.e3.

Zhou, M. C. and Venkatesh, K. (1999). *Modeling, simulation, and control of flexible manufacturing systems: A Petri net approach (Vol. 6).* World Scientific, Singapore.

Zigmont, J. J. (2010). *How paramedics learn: The relationship of experience, mental models, and analogical reasoning.* PhD thesis, University of Connecticut.

Zigmont, J. J., Kappus, L. J., and Sudikoff, S. N. (2011). The 3D model of debriefing: Defusing, discovering, and deepening. *Seminars in Perinatology*, 35(2):52–8.

Zyda, M. (2005). From visual simulation to virtual reality to games. *Computer*, 38(9):25–32.

SUMMARY

GAMING simulations (games) have been closely related with systems, and particularly with systems characterised as being complex. There is a long history on the evolution of systems from simply technical and social, to socio-technical (STS) and all the way to complex (adaptive) systems (CAS). Within this ever-increasing complexity, an important question that can be raised is: What is the role of games? Or in other words, what do games have to offer, that other tools do not, that can help us understand and further explore these complex systems?

Pure simulations (hereinafter referred to as simulations) have been perhaps the most popular tool for analysing systems. But as systems become more complex, and especially when that complexity is attributed to the bounded rationality of humans, it becomes cumbersome for simulations to capture the sometimes unpredictable behaviour of the increasing number of humans within systems. By all means, the field of simulations is not a static one, and advances in agent-based modelling (ABM) have shed light upon areas of human activity that previously was not feasible. Nevertheless, ABM is still an evolving area that has a long way to go until it can fully capture the richness of information that emerges from human activity. This is perhaps the most critical limitation of analytical science with regards to analysing human-enabled systems and a gap that games can fill, and have filled in many occasions. Therefore, the answer to the question posed in the previous paragraph is that games can help us grasp, both on an individual and collective level, the copiousness and complexity of systems through coupling the rigour of analytical science with the social problem-solving nature of design science.

Despite gaming being a relatively mature field with more than 50 years of academic and applied history, it is still lacking systematic and rigorous methodologies for designing, validating and executing a game session, and for managing the knowledge derived from them. While this holds true for the totality of games, this thesis focuses on games for decision making,particularly for engineering systems. Engineering systems are a typical example of CAS, due to their large scale, long lifetime, and in many occasions their close proximity to social systems, which in turn evoke complex features such as adaptation, self-organisation, and emergence.

Unlike games for learning and training, which enjoy a much more widespread adaptation resulting in an extensive body of knowledge in the policy and management domains, games for decision making have a shortage of literature resulting from their limited adaptation in organisations. Hence, the aim of this study is threefold: i. propose solutions to gaps identified in the literature, ii. extend the current relatively limited literature, and iii. start bridging the gap between the design and analytical communities by first acknowledging the struggle between them and then providing a fertile ground for discussion and future research. In order to accomplish that, this thesis explores four areas of games: design, validation, game sessions, and knowledge management.

These four areas are examined independently but are also explored as aspects of

games that intertwine and influence one another. Besides depending heavily on literature, several case studies, i.e. games, are used from the Dutch organisation, ProRail, that takes care of maintaining and extending the national railway network infrastructure, of allocating rail capacity, and of traffic control.

Design

Game design is characterised by a severe limitation in formal methods. Various frameworks and methods for formalising game design have been proposed, each having their advantages and disadvantages. In this thesis, the contributions and limitations of existing literature in game design are described and a novel framework based on game theory is proposed. While game theory has been scarcely used in game design, it is a well-established tool for understanding and modelling the relationships of different actors in an analytical way. There have been supportive studies towards using game theory in game design; game theory is a tool that can be coupled with existing methodologies and can help in bridging the gap between the design and analytical communities. Moreover, game theory can systematically analyse real-world systems and pinpoint those areas within these systems that are problematic and could benefit from the use of games. Finally, game theory can be directly used in game design by formalising game design choices and using optimal strategies wherever this is required.

The proposed framework consists of two phases: The Characterisation and the Links phases. The Characterisation phase is concerned with abstracting and modelling the real system using game theory and it was accomplished through an extensive literature review. The Links phase is concerned with connecting the elements of the model, i.e. the product of the Characterisation phase, with game design elements, thus creating a roadmap towards a formalised game design. This phase was construed first also through an extensive literature review and then by using two cases, i.e. games, from ProRail, in order to validate the connections. Finally, the complete framework was validated through three case studies, two from ProRail and one from the Swedish healthcare system.

Validation

Unlike pure simulations, games have a distinct characteristic, which is human participation, or in other words, games have the Game Layer on top of the Simulation Layer. Game validation, due to its nature of including humans, usually depends more on the subjective opinion of experts, e.g. interviews, than formal methods. This limitation is related to the lack of design methods for games as well as to the usually low number of participants. The first limitation is analysed as part of the research on the design of games (see the previous section). The second one, i.e. the sample size, plays a significant role for the applicability of game results. A small sample size is easy to obtain but has limited possibilities for analytical conclusions and thus limited possibilities for generalising the observations from the game. A large sample size, while solving the analytical problem and the generalisability of the results, is usually expensive to obtain and difficult to coordinate.

Validation of the Simulation Layer has been researched extensively over the course of the last three decades, and numerous formal methods and statistical techniques have been developed. However, this abundance of research has become an actual inhibitor due to the large number of available methods and techniques. In this thesis, a framework for selecting the most appropriate simulation validation methods and statistical techniques, given the different characteristics of simulations, is proposed. Moreover, a web tool that automates validation is developed and the lessons learned from its application in two cases from the railways are derived.

Validation of the Game Layer due to its nature of including uncertainties pertaining to human activity, is usually not as straightforward as the Simulation Layer. On the one hand, the formalisation of game design can provide more structure to game validation. On the other hand, with regards to the sample size, the Game Layer would benefit only through gradually extending the body of knowledge by building upon previous work. This aspect is directly linked with knowledge management in the sense that the more game sessions are conducted the more evidence of a system's behaviour are discovered and the cumulative sample size gradually becomes large enough to generalise the outcome of the game.

The games used in engineering organisations, like the primary case study in this thesis, tend to have a very dominant Simulation Layer, as is the case with all the games used in this thesis. As a result, the part concerning the validation of games is heavily focused, though not solely, on the validation of what appear to be pure simulations. Nevertheless, the implications of the validation of the simulation model in game validation become evident throughout the thesis.

Game Session

A game session consists of three phases: briefing, gameplay, and debriefing, with the last one being considered the most important feature of game sessions. Nevertheless, games' almost completely synthetic nature raises the question: are game sessions in general and debriefing in particular performed and analysed in a rigorous scientific way? In other words, are they consistently structured, given the different characteristics of games, and is it also clear what would constitute a successful game session and debriefing? The answer to all these questions is no. The reason for this negative outcome is that expertise regarding game sessions and debriefing resides almost entirely in the tacit part of the knowledge spectrum. As a result, knowledge and best practices on how to conduct fruitful game sessions and debriefing are either disseminated without understanding the reason why a particular decision would be beneficial, or not disseminated at all. Hence, the aim in this thesis is to elicit this tacit knowledge possessed by experts and to gain understanding on why certain practices are more prone to success than others, as well as bring to the surface other practices that have remained well hidden. In order to accomplish this goal, two rounds of interviews with facilitation experts and other game stakeholders were conducted.

Knowledge Management

Knowledge management (KM) and reuse of games is not, and should not be, of academic interest only. The effectiveness of a corporation depends heavily on how it manages and reuses knowledge. As a corporation acquires and builds up knowledge obtained through games, it improves its know-how, and thus sustains or even increases its competitive advantage.

From a practical point of view, knowledge should be translated in a system, i.e. a knowledge management system. In this thesis, a methodology for defining the backbone structure of such a system is proposed. The proposed methodology is first constructed based on the state of the art and then tested using three cases from ProRail.

Synthesis of Results

The four areas of games mentioned above should not be seen as four separate and unrelated entities. Choices regarding the design of games can positively affect or inhibit validation, depending on how the real system is translated into a game. This connection between design and validation becomes apparent when moving from a more synthetic design approach to a more formal one, in which case the formalisation enables the utilisation of formal validation methods. Validation increases a game's credibility in the eyes of the stakeholders, which in turn enables a game session to be more effective. Equivalently, a game session that is considered âĂIJsuccessfulâĂİ increases the game's credibility, usability, and applicability, thus making it more valid. As such, the methodologies associated with knowledge management should be, on the one hand, tailored to the design, validation, and execution of game sessions, and on the other hand according to the user characteristics and the organisation's culture.

In addition to the influence these areas have on one another, during the course of this thesis, they are also juxtaposed with three research themes: i. the contribution of the analytical and design sciences to each area, ii. the different kinds of complexity that occur in each area, and iii. how these areas contribute to the quality of a decision making process. The added value of not just exploring each area of games separately is twofold. On the one hand, games are seen as the complete artefact that they are, in order to understand how as a whole they contribute towards solving the problem at hand. On the other hand, by viewing games more holistically, the relative, and absolute, importance of each area in the big picture of the game is more easily determined and further explored.

Conclusion

This study is comprised of four parts, each of which deals with an aspect of games. For each of these aspects, this thesis contributes to new methodologies or conceptual frameworks and promotes a new way of approaching and thinking about games. With regards to design, a framework based on game theory was proposed. The framework was shown to be able to translate some elements from the game theoretical concepts into game design decisions directly (e.g., Actors → Characters, Pay-offs → Rewards, Outcome → Purpose), while for other more qualitative elements (e.g., Actions, Strategies) to provide insights to game designers. Moreover, the utilisation of game concepts enabled the framework to be able to predict how a situation characterised by a specific concept might evolve in the future in another concept. Finally, in one particular case, the framework identified the worst-case scenario, thus pinpointing the area that would benefit the most from a game.

With regards to validation, the thesis contributed in two areas of simulations both of which are linked with games for engineering systems. First, a framework for validation and verification methods and statistical techniques selection was proposed. The framework incorporated various properties and characteristics of simulations and systems and it demonstrated that indeed some of these influence the method and technique selection and thus, the final results of a simulation study. Second, the lessons learned from the development and application of a tool for automating a validation study, based on Web technologies and the R statistical language, were presented and their potential benefits in games were illustrated. Some of the critical success factors of validation studies included the need for organisations to develop knowledge management protocols, so as to accommodate the difference in the educational and professional background of all the involved actors, and the necessity for the analysis and visualisation of data to initially include all the available information, which analysts can then reduce, abstract, or simplify in view of fulfilling their respective goals.

With regards to game sessions, several best practices were identified through a series of interviews with debriefing experts and other game stakeholders. In more detail, three rounds of interview were conducted in total. The first was with 8 game facilitation experts, which yielded several pitfalls occurring during the debriefing of games, like the inadequate preparation of the facilitators and the lack of emotional safety on behalf of the participants. The second was with 19 experts of which 7 game designers, 6 project leaders, 4 participants, and 2 department managers, and it used the Q-methodology to determine factors that either boost or inhibit games' success, like the strict rules and time pressure. The third was with 21 game facilitation experts, and it was mainly characterised by their contradicting answers in almost all questions, still the study managed to pinpoint some factors that could be determinants on whether a game session was successful, like the freedom and feel of safety for the participants to share their experience and whether the knowledge obtained from the game was actually used in the real world.

With regards to knowledge management, a framework for a knowledge management system that supports games was proposed, which was then applied and validated in three case studies, i.e. games, from the Dutch railways. The games provided additional insights on how such a framework can both enhance the validation process as well as help the organisation to grow. All in all, the framework proposed general guidelines on the components to consider for the development of a module that would accommodate the particularities of games, in order to integrate the knowledge acquired through games in a knowledge management system.

SAMENVATTING

GAMING simulaties (in het vervolg games) zijn sterk gerelateerd aan systemen, en in het bijzonder aan systemen die gekarakteriseerd zijn als complex. Er is een lange geschiedenis over de evolutie van systemen, van simpel technisch en sociaal, naar sociotechnische systemen (STS) tot complex (adaptieve) systemen (CAS). Door de toenemende complexiteit ontstaat de volgende vraag: wat is de rol van games? Of, met andere woorden, wat kunnen games bijdragen, wat andere tools niet kunnen, om ons te helpen de complexe systemen te begrijpen en verder te onderzoeken?

Pure simulaties (in het vervolg simulaties) zijn wellicht de meest populaire tool voor het analyseren van systemen. Maar wanneer systemen meer complex worden, en wanneer de complexiteit is toegeschreven een de begrensde rationaliteit van mensen, wordt het lastig voor simulaties om het soms onvoorspelbare gedrag van een toenemend aantal actoren in het systeem te vangen. In ieder geval, het veld van simulaties is niet statisch, en ontwikkelingen in agent-gebaseerde modellering (ABM) heeft mogelijkheden gebracht op het gebied van menselijke activiteiten die eerder niet mogelijk waren. Aan de andere kant, ABM is nog steeds een gebied in ontwikkeling dat nog een lange weg te gaan heeft voor het de volledige rijkheid van informatie die ontstaat door menselijke activiteit kan vangen. Dit is waarschijnlijk de meest kritische limitatie van de analytische wetenschap met betrekking tot het analyseren van systemen met mensen. Games zijn in staat, en hebben dat in veel gevallen bewezen, om deze limitatie te overkomen. Het antwoord op de vraag in de eerste alinea is dat games ons kunnen helpen, zowel op een individueel als op een collectief niveau, om de overvloedigheid, en daaruit volgende complexiteit van systemen, te begrijpen. Zij doen dit door de analytische wetenschap met het sociaal probleem oplossende vermogen van ontwerp wetenschap met elkaar te verbinden.

Ondanks dat games een wetenschappelijke en toegepaste geschiedenis hebben van ruim 50 jaar missen er methoden voor het systematisch ontwerpen, valideren en uitvoeren van game sessies, en het organiseren van de hieruit verkregen informatie. Het voorgaande geldt voor alle games, maar in deze thesis is gekozen voor een focus op games voor besluitvorming, en in het bijzonder, voor engineering systemen. Engineering systemen zijn voorbeelden van CAS vanwege hun grote schaal, lange levensduur, en nabijheid met sociale systemen die complexe kenmerken zoals aanpassing, zelforganisatie en emergentie oproepen.

Anders dan games voor leren en training, die een veel grotere toepassing hebben, wat vervolgens resulteert in een grote kennis in het beleid en management domein, is er over games voor besluitvorming een beperkte hoeveelheid literatuur beschikbaar. Dit resulteert in een beperkte toepassing van dit soort games binnen organisaties. Dus, het doel van deze studie is driezijdig: i. het voorstellen van kennistekorten geïdentificeerd in de literatuur, ii. het verbreden van de huidige relatief gelimiteerde literatuur, en iii. het overbruggen van de kenniskloof tussen onderwerp en analytische gemeenschappen door eerst te onderkennen dat er een strijd is tussen hen, en daarmee een vruchtbare basis te creëren voor discussie en verder onderzoek. Om dat te bereiken onderzoekt deze thesis vier gebieden van games: ontwerp, validatie, game sessies, en kennismanagement.

De vier gebieden zijn onafhankelijk van elkaar bestudeerd, maar ook onderzocht als aspecten van games die met elkaar vervlochten zijn en elkaar beïnvloeden. Naast dat we ons baseren op literatuur zijn er meerdere casussen, i.e., games, van de Nederlands infrastructuur manager ProRail gebruikt. ProRail is verantwoordelijk voor het onderhoud en de uitbreiding van het spoor, het verdelen van de capaciteit op het spoor, en de verkeersleiding.

Ontwerp

Het ontwerpen van games is gekarakteriseerd door een sterke limitatie in het gebruik van formele methoden. Verschillende frameworks en methoden voor het formaliseren van het ontwerpen van games zijn voorgesteld, elk met hun eigen voor- en nadelen. In deze thesis worden de bijdragen en beperkingen van bestaande literatuur in het ontwerp van games beschreven, en een nieuw framework gebaseerd op speltheorie is ontwikkeld. Ondanks dat speltheorie weinig is gebruikt in het ontwerpen van games is het wel een gevestigde tool voor het begrijpen en modeleren van de relaties tussen actoren op een analytische manier. Er zijn studies die speltheorie voor het ontwerpen van games gebruiken; speltheorie is een tool die gekoppeld kan worden aan bestaande methoden en kan helpen in het overbruggen van ontwerp en analytische gemeenschappen. Bovendien kan speltheorie op een systematische manier de systemen in de echte wereld analyseren en benadrukken waar deze systemen problematisch zijn en waar zij kunnen profiteren van het gebruik van games. Speltheorie kan direct worden gebruikt voor het ontwerpen van games door ontwerpkeuzes te formaliseren en optimale strategieën te gebruiken waar wenselijk.

Het voorgestelde framework omvat twee fasen: de Karakterisatie en de Verbinding. De Karakterisatie fase gaat over het abstraheren en modeleren van het echte systeem gebruik makend van speltheorie en is verkregen door een uitgebreid literatuur review. De Verbinding fase gaat over het verbinden van de elementen van het model, i.e., het product van de Karakterisatie fase, met de game ontwerp elementen. Het maakt dus een stappenplan voor het formaliseren van het ontwerpen van games. Ook deze fase is geconstrueerd door een uitgebreid literatuur review en voor de validatie van de verbindingen zijn twee casussen, i.e., games van ProRail, gebruikt. Uiteindelijk is het complete framework gevalideerd met drie casussen, twee van ProRail en een casus over het Zweedse gezondheidszorgsysteem.

Validatie

Anders dan pure simulaties hebben games hebben een onderscheidend kenmerk, namelijk de deelname van mensen. In andere woorden, games hebben een Game Level bovenop een Simulatie Level. Het valideren van games hangt daardoor meer af van de subjectieve mening van experts, bijvoorbeeld door middel van vragenlijsten, dan van formele methoden. Deze limitatie is gerelateerd aan het gebrek van ontwerp methoden voor games en het meestal laag aantal deelnemers aan de game. Het eerste punt is geanalyseerd als onderdeel van het onderzoek naar het ontwerpen van games (zie voorgaande sectie). Het laatste punt, i.e., het aantal deelnemers, speelt een significante rol in de toepassing van game resultaten. Een kleine groep deelnemers is makkelijk te verkrijgen, maar heeft gelimiteerde mogelijkheden voor analytische conclusies en dus beperkte mogelijkheden voor het generaliseren van de observaties van de game. Een groter aantal deelnemers zal het analytische probleem oplossen, en het generaliseren van de resultaten mogelijk maken, maar is vaak moeilijk te verkrijgen vanwege kosten en is veel lastiger te coördineren.

Validatie van het Simulatie Level is de laatste drie tientallen jaren uitgebreid onderzocht waarbij meerdere formele methoden en statistische technieken zijn geïntroduceerd. Bovendien zijn er methoden voor verificatie van kleine steekproefgrootten voorgesteld. Aan de andere kant is de overvloed aan onderzoek een remmende factor vanwege het grootte aantal methoden dat eruit voortkomt. In deze thesis is een framework voorgesteld voor het selecteren van de best passende simulatie validatie methode en statistische techniek gegeven de verschillende karakteristieken van de simulatie. Hiernaast is een web tool verkend die validatie automatiseert en waarin de geleerde lessen van de toepassingen in twee casussen uit de spoorsector zijn opgenomen.

Validatie van het Game Level is niet zo eenvoudig als dat van het Simulatie Level vanwege de onzekerheden die ontstaat door menselijke activiteiten. Aan de ene kant kan formalisatie van het ontwerpen van games meer structuur geven aan game validatie. Aan de andere kant, met betrekking tot de steekproefgrootte, zal het Game Level alleen voordeel ondervinden door het gradueel uitbreiden van de kennis door voort te bouwen op eerder werk. Dit aspect is direct gekoppeld aan het organiseren van kennis in de zin dat, als meer game sessies zijn gehouden, er meer bewijs voor het gedrag van het systeem is ontdekt en de opgetelde steekproefgrootte groot genoeg wordt om de uitkomsten van de game te generaliseren.

De games in engineering organisaties, zoals de belangrijkste casussen in deze thesis, hebben een dominant Simulatie Level, en dit is dus ook het geval voor de games gebruikt in deze thesis. Het resultaat hiervan is dat het onderdeel dat gaat over de validatie van games grotendeels is gefocust op validatie van pure simulaties. Niettemin, de implicaties van de validatie van het simulatiemodel in de game worden duidelijk door de thesis.

Game sessies

Een game sessie bestaat uit drie fasen: voorbespreking, spelen van de game, en nabespreking, waarbij de laatste fase als meest belangrijke onderdeel van de game wordt gezien. Niettemin, de bijna complete synthetische aard van games doet de vraag rijzen: worden game sessies, en in het bijzonder de nabespreking hiervan, geanalyseerd op een wetenschappelijke manier? Met andere woorden, zijn ze consistent gestructureerd gegeven de verschillende karakteristieken van de game, en is het duidelijk wat het betekent voor een game en de nabespreking om succesvol te zijn. Het antwoord op deze vragen is nee. De reden voor deze negatieve uitkomst is dat de expertise met game sessies en de nabespreking bijna volledig onderdeel is van het tacit kennisspectrum.

Het resultaat hiervan is dat de kennis en de beste praktijken over hoe men een vruchtbare game sessie en na bespreking moet uitvoeren zijn verspreid zonder begrip van de reden waarvoor een bepaald besluit voordelig zou zijn, of ze zijn helemaal niet verspreid. Dus, het doel van deze thesis is om aandacht te geven aan de tacit kennis vervuld door experts, en om begrip te krijgen voor waarom het ene voorbeeld meer succesvol is dan de andere, net als het naar voren brengen van voorbeelden die onbekend zijn gebleven. Om dit doel te bereiken zijn twee ronden interviews met experts in het faciliteren van games en andere game professionals uitgevoerd.

Kennismanagement

Een game sessie bestaat uit drie fasen: voorbespreking, spelen van de game, en nabespreking, waarbij de laatste fase als meest belangrijke onderdeel van de game wordt gezien. Niettemin, de bijna complete synthetische aard van games doet de vraag rijzen: worden game sessies, en in het bijzonder de nabespreking hiervan, geanalyseerd op een wetenschappelijke manier? Met andere woorden, zijn ze consistent gestructureerd gegeven de verschillende karakteristieken van de game, en is het duidelijk wat het betekent voor een game en de nabespreking om succesvol te zijn. Het antwoord op deze vragen is nee. De reden voor deze negatieve uitkomst is dat de expertise met game sessies en de nabespreking bijna volledig onderdeel is van het tacit kennisspectrum.

Het resultaat hiervan is dat de kennis en de beste praktijken over hoe men een vruchtbare game sessie en na bespreking moet uitvoeren zijn verspreid zonder begrip van de reden waarvoor een bepaald besluit voordelig zou zijn, of ze zijn helemaal niet verspreid. Dus, het doel van deze thesis is om aandacht te geven aan de tacit kennis vervuld door experts, en om begrip te krijgen voor waarom het ene voorbeeld meer succesvol is dan de andere, net als het naar voren brengen van voorbeelden die onbekend zijn gebleven. Om dit doel te bereiken zijn twee ronden interviews met experts in het faciliteren van games en andere game professionals uitgevoerd.

Samenvatting van de resultaten

De vier hiervoor genoemde gebieden van games moet men niet zien als vier compleet verschillende en niet gerelateerde onderdelen. Omdat ze allemaal en onderdeel zijn van games zijn ze vaak met elkaar vervlochten en beïnvloeden ze elkaar. Keuzes voor het ontwerp van de game kan de validatie positief beïnvloeden, maar ook verhinderen afhankelijk van hoe het systeem is vertaald naar de game. Deze verbinding tussen ontwerp en validatie wordt duidelijk door te veranderen van een meer synthetisch naar een meer formele ontwerp benadering waarbij de formalisatie het gebruik van formele validatie methoden in staat stelt. Validatie verhoogt de geloofwaardigheid van games in de ogen van professionals, wat ertoe leidt dan een game sessie meer effectief is. Op een zelfde manier, wanneer een game sessie gezien wordt als “succesvol” verhoogt dat de game zijn geloofwaardigheid, bruikbaarheid, en toepasbaarheid, wat de game weer meer valide maakt. Als zodanig, de methoden die geassocieerd zij met kennismanagement moeten aan de ene kant geschikt zijn voor het ontwerp, de validatie en de game sessie, maar het moet ook passen bij de kenmerken van de gebruiker en cultuur van de organisatie.

Toevoegend aan de invloed die de game gebieden op elkaar hebben in de gehele thesis, kunnen ze ook naast elkaar geplaatst worden in drie onderzoeksthema’s: i. de bijdrage van de analytische en ontwerp wetenschap voor elk gebied, ii. de verschillende soorten complexiteit die voorkomen in elk gebied, en iii. op welke manier deze gebieden bijdragen aan de kwaliteit van het besluitvormingsproces. Het niet alleen bekijken van de verschillende game gebieden heeft toegevoegde waarde. Aan de ene kant kunnen games gezien worden als het complete artefact om te begrijpen hoe ze als geheel bijdragen aan de oplossing van het probleem. Aan de andere kant, wanneer we de games meer holistisch bekijken, dan is de relatieve, en absolute, bijdrage van elk gebied in het grotere geheel van de game makkelijker te bepalen en verder te onderzoeken.

CONCLUSIE

Deze studie voegt vier onderdelen samen waarvan elk een aspect van games bevat. Voor elk van deze aspecten draagt de thesis bij aan nieuwe methoden of conceptuele frameworks, en de thesis promoot een nieuwe manier van nadenken over games. Voor het ontwerp van games is een framework gebaseerd op speltheorie voorgesteld. Het framework is in staat om elementen van spel theoretische concepten direct te vertalen naar game ontwerp keuzes (bijv. Actoren → Spelers, Pay-offs → Beloningen, Uitkomst → Doel). Terwijl voor meer kwalitatieve elementen (bijv. Acties, Strategieën) geeft het framework inzicht aan game ontwerpers. Bovendien stelt het gebruik van game concepten het framework in staat om een situatie, die gekarakteriseerd is door een bepaald concept, zich kan ontwikkelen in de toekomst naar een ander concept. Uiteindelijk, in een bepaalde casus, heeft het framework een worst-case scenario geindentificeerd en wijst daarmee het gebied aan dat het meest voordeel heeft van het gebruik van een game.

Op het gebied van validatie draagt de thesis bij aan twee gebieden van simulaties die beide gekoppeld zijn met games voor engineering systemen. Ten eerste, een framework voor validatie en verificatie methoden en statistische technieken is voorgesteld. Het framework bevat verschillende eigenschappen en karakteristieken van simulaties en systemen. Daarbij heeft het gedemonstreerd dat een aantal van deze de selectie van methoden en technieken beïnvloeden en daarmee dus het uiteindelijke simulatie onderzoek. Ten tweede, de lessen die we hebben geleerd van de ontwikkeling en toepassing van een tool voor het automatiseren van een validatie onderzoek zijn gepresenteerd en de mogelijke voordelen in games zijn geïllustreerd. Een aantal van de kritische succesfactoren van validatie onderzoek bevatten de noodzaak voor organisaties om kennismanagement protocollen te ontwikkelen naast het aanpassen van verschillen in educatie en professionele achtergrond van betrokken actoren. Daarnaast is er de noodzaak voor de analyse en visualisatie van data die alle mogelijke informatie bevat in eerste instantie om vervolgens te worden gereduceerd, geabstraheerd of gesimplificeerd door analisten om te voldoen aan het gewenste doel.

Met betrekking tot game sessies zijn een aantal beste praktijken geïdentificeerd door middel van een aantal interviews met nabespreking experts en andere game belanghebbenden. Er werden in totaal drie interviewronden gehouden. De eerste ronde bevatte 8 game facilitator experts wat resulteerde in verschillende valkuilen gedurende de nabespreking van games. Voorbeelden zijn onvoldoende voorbereiding door facilitators en onvoldoende emotionele veiligheid voor de deelnemers. De tweede ronde interviews is gehouden met 19 experts, 7 game ontwerpers, 6 projectleiders, 4 deelnemers, en 2 afdelingsmanagers. Q-methodologie is gebruikt om factoren te bepalen die het succes van een game verhindert, zoals strenge regels en tijdsdruk. De derde ronde interviews in gehouden met 21 game facilitator experts, deze ronde werd gekarakteriseerd door tegenstrijdige antwoorden op bijna alle vragen. Tegelijkertijd kunnen heeft de studie een aantal factoren kunnen aanwijzen die bepalend zijn voor een succesvolle game sessie zoals vrijheid en gevoel van veiligheid bij de deelnemers om hun ervaringen te delen en of de kennis verkregen door de game was gebruik in de praktijk.

Met betrekking tot kennismanagement is een framework voorgesteld voor een kennismanagementsysteem dat games ondersteunt. Het frameworks is vervolgens toegepast en gevalideerd in drie casussen, i.e., games, van de Nederlandse spoorsector. De games hebben inzicht gegeven in hoe zo een framework kan bijdragen aan zowel het validatieproces als het groeien van de organisatie. In zijn totaliteit geeft het framework algemene richtlijnen voor de componenten om mee te nemen bij de ontwikkeling van een module die gaat over bijzonderheden van games om zo de kennis verkregen door games te integreren in een kennismanagementsysteem.

Appendix A: Validation and Verification Methods

1. *Acceptance Testing:* Acceptance Testing is testing the model using the actual hardware and data to determine whether all the specified requirements are satisfied (Perry, 2007).

2. *Alpha Testing:* Alpha Testing is the operational testing of the alpha version of the model in a department within the company, yet not the one involved with the model development (Beizer, 1990).

3. *Assertion Checking:* Assertion Checking checks what is happening as opposed to what the modeller assumes is happening thus detecting potential errors in the model (Balci, 1998).

4. *Audit:* An Audit is undertaken to assess how adequately the simulation study is conducted with respect to established plans, policies, procedures, standards and guidelines. The audit also seeks to establish traceability within the simulation study (Balci, 1998).

5. *Beta Testing:* Beta Testing is the operational testing of the beta version of the model under realistic field conditions (Miller et al., 1993).

6. *Bottom-Up Testing:* Bottom-Up Testing is testing each submodel, when the model is developed with a bottom-up development strategy, and once every submodel belonging to the same parent is finished and tested, then these submodels are integrated and tested again (Balci, 1998).

7. *Cause-Effect Graphing:* Cause-Effect Graphing aids in selecting, in a systematic way, a high-yield set of test cases and it is effective in pointing out incompleteness and ambiguities in the specification (Myers et al., 2011).

8. *Comparison Testing:* Comparison Testing is testing the different versions of the same simulation model (Pressman, 2015).

9. *Compliance Testing:* Compliance Testing tests how accurately different levels of access authorisation are provided, how closely and accurately dictated performance requirements are satisfied, how well the security requirements are met, and how properly the standards are followed (DoD-5000.59-M, 1997). It consists of the following techniques:

(a) Authorization Testing, which tests how accurately and properly different levels of access authorisation are implemented in the model and how properly they comply with the rules and regulations (Perry, 2007).

(b) Performance Testing, which tests whether (a) all performance characteristics are measured and evaluated with sufficient accuracy, and (b) all established performance requirements are satisfied (Perry, 2007).

(c) Security Testing, which tests whether all security procedures are correctly and properly implemented in conducting a simulation study (Perry, 2007).

(d) Standards Testing, which substantiates that the simulation model is developed with respect to the required standards, procedures, and guidelines (Balci, 1998).

10. *Control Analysis:* Control Analysis analyses the control characteristics of the model. It consists of the following techniques:

    (a) Calling Structure Analysis, which is used to assess model accuracy by identifying who calls whom and who is called by whom (Miller et al., 1993).

    (b) Concurrent Process Analysis, in which model accuracy is assessed by analysing the overlap or concurrency of model components executed in parallel or as distributed (Rattray, 1990).

    (c) Control Flow Analysis, which requires the development of a graph of the model where conditional branches and model junctions are represented by nodes and the model segments between such nodes are represented by links (Beizer, 1990).

    (d) State Transition Analysis, which requires the identification of a finite number of states the model execution goes through and shows how the model transitions from one state to another (Balci, 1988).

11. *Data Analysis:* Data Analysis ensures that (1) proper operations are applied to data objects (e.g., data structures, event lists, linked lists), (2) the data used by the model are properly defined, and (3) the defined data are properly used (Perry, 2007). It consists of the following techniques:

    (a) Data Dependency Analysis, which involves the determination of what variables depend on what other variables (Dunn, 1987).

    (b) Data Flow Analysis, which is used to assess model accuracy with respect to the use of model variables (Adrion et al., 1982).

12. *Debugging:* Debugging identifies errors causing the model to fail and changes the model accordingly in order to correct these errors (Dunn, 1987).

13. *Desk Checking:* Desk Checking is when a person other than the modeller thoroughly examines the model to ensure correctness, completeness, consistency and unambiguity (Beizer, 1990).

14. *Documentation Checking:* Documentation Checking ensures accuracy and up-to-date description of the model logic and its results (Balci, 1998).

15. *Execution Testing:* Execution Testing collects and analyses execution behaviour data in order to reveal model representation errors. It consists of the following techniques:

    (a) Execution Monitoring, which examines low-level information about activities and events taking place during model execution (Balci, 1998).

    (b) Execution Profiling, which examines high-level information about activities and events taking place during model execution (Balci, 1998).

    (c) Execution Tracing, which tracks line-by-line the execution of a model (Balci, 1998).

16. *Face Validation:* In Face Validation, people knowledgeable about the system under study subjectively compare model and system behaviours and judge whether the model and its results are reasonable (Hermann, 1967).

17. *Fault/Failure Analysis:* Fault/Failure Analysis determines if any faults or failures can logically occur and in what context and under what conditions (Miller et al., 1993).

18. *Fault/Failure Insertion Testing:* Fault/Failure Insertion testing inserts an fault or failure into the model and observes whether the model will behave in the expected invalid manner (Balci, 1998).

19. *Field Testing:* Field Testing executes the model in an operational situation for the purpose of collecting information regarding the model validation (Shannon and Johannes, 1976).

20. *Functional (Black-Box) Testing:* Functional Testing ignores the internal mechanism(s) of the model and focuses on the generated outputs based on specific input and execution conditions (IEEE, 2011).

21. *Graphical Comparisons:* In Graphical Comparison, graphs produced from the model are compared to graphs produced by the real-world system under study, in order to detect similarities and differences between the two (Miller et al., 1993).

22. *Induction:* Induction asserts that if every step a model follows is valid and the model terminates, then the model is valid (Balci, 1998). Induction as a term can be

found in many fields, like mathematics in which case it is a tool for directly proving theorems. In simulation model validation, where absolute validity does not exist (Martis, 2006), induction should more correctly be referred to as inductive reasoning, which is based on one or more inductive arguments, and the conclusions are not considered as the absolute truth but rather a strong evidence (Copi et al., 2016).

23. *Inference:* Inference is similar to Induction; it is a mental process by which one proposition is arrived at and affirmed on the basis on one or more other propositions assumed as the starting point of the process (Copi et al., 2016).

24. *Inspections:* Inspection is a five phase procedure conducted by four to six people. The phases include not only a validation phase but also suggestions for improvements and a follow-up (Schach, 2011).

25. *Interface Analysis:* Interface Analysis consists of the following techniques:

    (a) Model Interface Analysis, which is conducted to examine the (sub)model-to-(sub)model interface and determine if the interface structure and behaviour are sufficiently accurate (Balci, 1998).

    (b) User Interface Analysis, which is conducted to examine the user-model interface and determine if it is human engineered so as to prevent occurrences of errors during the user's interactions with the model(Balci, 1998).

26. *Interface Testing:* Interface Testing consists of the following techniques:

    (a) Data Interface Testing, which assesses the accuracy of data inputted into the model or outputted from the model during its execution (Miller et al., 1993).

    (b) Model Interface Testing, which detects model representation errors caused due to interface errors (Balci, 1998).

    (c) User Interface Testing, which deals with the assessment of the interactions between the user and the model, and detects errors associated with those (Schach, 2011).

27. *Lambda (λ) Calculus:* λ-calculus is a mathematical tool for formally defining systems (Barendregt, 1984). λ-calculus offers function that can be translated into validation rules.

28. *Logical Deduction:* Logical Deduction, also known as Deductive Reasoning, is similar to Induction but the conclusions are considered as logically true, or valid, if every step of the model is valid and the model terminates (Copi et al., 2016).

29. *Object-Flow Testing:* Object-Flow Testing assesses model accuracy by exploring the life cycle of an object during the model execution (Balci, 1998).

30. *Partition Testing:* Partition Testing, also known as equivalent class partitioning, partitions the model into functional representatives (partitions), assuming that all elements within each partition bear the same properties, and then, by selecting a representative element from each partition, each partition and subsequently the model is validated, thus eliminating the need for exhaustive validation (Burnstein, 2006).

31. *Predicate Calculus:* Predicate Calculus quantifies simple relationships (predicates) using boolean variables. Since, the model can be defined based on predicates, then its validation can be performed by manipulating these predicates (Balci, 1998). Similarly to Deduction, Predicate Calculus' conclusions are logically true or valid.

32. *Predicate Transformations:* Predicate Transformations, or more formally known as Predicate Transformer Semantics, show that systems (in this case a simulation model) can achieve their goals, i.e., they are valid. Predicate Transformations associate a pre-condition to any post-condition, or in other words transform model output states into all model input states, thus providing the basis for proving model correctness (Dijkstra, 1976).

33. *Predictive Validation:* In Predictive Validation, the model executes with past input data and the results are then compared with data from the real system (Emshoff and Sisson, 1970).
    Product Testing: Product Testing is a preparatory step for the Acceptance Testing, in which all requirements specification are tested in the same way as in the Acceptance Testing, with the only difference being that the test takes place within the development team whereas Acceptance Testing takes place at the client's premises (Schach, 2011).

34. *Proof of Correctness:* A Proof of Correctness expresses the model in a precise notation and then proves that the model terminates and thus satisfies the requirements specification with sufficient accuracy (Backhouse, 1986).

35. *Regression Testing:* Regression Testing ensures that correcting errors in the model during the validation process do not create new errors or adverse side-effects (Balci, 1998).

36. *Reviews:* Reviews are similar to an inspection but the review team also involves managers. Reviews are intended to give management and study sponsors evi-

dence that the model development is being conducted according to the study objectives (Perry, 2007).

37. *Semantic Analysis:* Semantic Analysis attempts to determine the modeller's intent in writing the code (Whitner and Balci, 1989).

38. *Sensitivity Analysis:* In Sensitivity Analysis, selected variables in the model are given different values (within a predetermined range) in order to observe the behaviour of the model with regards to these changes (Shannon and Johannes, 1976).

39. *Special Input Testing:* Special Input Testing assesses model accuracy by subjecting the model in a variety of inputs and consists of the following techniques:

    (a) Boundary Value Testing, which tests the boundary values of the input and output equivalence classes (a set of values that bear similar characteristics and one value can act as a representative for the whole set) (Myers et al., 2011).

    (b) Equivalence Partitioning Testing, which tests the model by partitioning input data into equivalence classes (Perry, 2007).

    (c) Extreme Input Testing, which tests the model based on extreme input values (minimum, maximum, or a mixture of those) (Balci, 1998).

    (d) Invalid Input Testing, which tests the model using incorrect input data (Balci, 1998).

    (e) Real-Time Input Testing, which tests the model using real-time input data from the real system (Balci, 1998).

    (f) Self-Driven Input Testing, which test the model by executing it under input data randomly sampled from probabilistic models representing random phenomena of a real system (Balci, 1998).

    (g) Stress Testing, which tests the model by subjecting it into heavy loads, like large volumes of data, intense activity over a short time span etc (Myers et al., 2011).

    (h) Trace-Driven Input Testing, which tests the model by executing it under input trace data collected from a real system (Balci, 1998).

40. *Structural (White-box) Testing:* Structural Testing is used to evaluate the internal structure of the model and consists of the following techniques:

    (a) Branch Testing, which tests the model under test data in order to execute as many branch alternatives as possible (Beizer, 1990).

    (b) Condition Testing, which tests the model under test data in order to execute as many logical conditions as possible (Balci, 1998).

(c) Data Flow Testing, which tests the model by using the control flow graph as to explore sequences of events related to the status of data structures and to examine data-flow anomalies (Beizer, 1990).

(d) Loop Testing, which tests the model under test data in order to execute as many loop structures as possible (Pressman, 2015).

(e) Path Testing, which tests the model under test data in order to execute as many control flow paths as possible (Beizer, 1990).

(f) Statement Testing, which which tests the model under test data in order to execute as many statements as possible (Beizer, 1990).

41. *Structural Analysis:* Structural Analysis is used to examine the model structure and to determine if it adheres to structured principles (Balci, 1998).

42. *Submodel/Module Testing:* Submodel/Module Testing is a top-down form of testing in which is submodel is tested against its corresponding subsystem (Balci, 1998).

43. *Symbolic Debugging:* Symbolic Debugging is a verification method in which the use of "breakpoints" allows for a direct manipulation of the model execution while viewing the model at the source code level (Balci, 1998).

44. *Symbolic Evaluation:* Symbolic Evaluation assesses model accuracy by executing the model using as an input symbolic values and not the actual data values (King, 1976).

45. *Syntax Analysis:* Syntax Analysis assures that the mechanics of the programming language are applied correctly (Beizer, 1990).

46. *Top-Down Testing:* In Top-Down Testing, the model testing starts from the submodels at the highest level and moves downwards into the base submodels (Sommerville, 2004).

47. *Traceability Assessment:* Traceability Assessment matches, one to one, the elements of one form of the model to another (Miller et al., 1993).

48. *Turing Test:* In a Turing Test, experts are presented with two sets of output data, i.e., the model and reality, and without knowing which one is which, they are asked to differentiate the two (Schruben, 1980).

49. *Visualisation/Animation:* In Visualisation/Animation, the model is tested by observing different graphs of the internal or external behavior of the model (Balci, 1998).

50. *Walkthroughs:* Walkthroughs are used to detect and document faults. They are similar to an Inspection but less time consuming, they have fewer phases (Myers et al., 2011).

Appendix B: Statistical Techniques

In the statistical formulas shown in this appendix, wherever M and R are used as subscripts, they denote that the particular variable refers to the model or reality respectively. Moreover, unless explicitly stated, n with the appropriate subscript denotes the respective sample size.

1. *t-Test:* The t-Test, also known as Student's t-test, is a statistical hypothesis test, which determines whether the mean of a variable is significantly different from a constant value (one-sample test) or whether the mean of two variables is significantly different (two-sample test) (Hahs-Vaughn and Lomax, 2013). The most common usage of t-test in simulation model V&V is the two-sample test (Model and Reality) with unequal sample sizes and variances. The latter is also known as Welch t-test (Welch, 1947) and its formula is:

$$t = \frac{\overline{X}_M - \overline{X}_R}{\sqrt{\frac{s_M^2}{n_M} + \frac{s_R^2}{n_R}}} \tag{16.1}$$

   where $\overline{X}$ and s are the mean and variance respectively. The t-test is one of the most commonly used tests for the equality of means between model and reality.

2. *Hotelling's T^2 Test:* Hotelling's T^2 test is a generalisation of the t-test for multivariate hypothesis testing (Hotelling, 1992). As it is the case with t-test, Hotelling's T^2 test can also be used for one- or two-sample testing. Its formula for the two-sample test is:

$$T^2 = (\overline{X}_M - \overline{X}_R)' \left\{ S_p \left(\frac{1}{n_M} + \frac{1}{n_R} \right) \right\}^{-1} (\overline{X}_M - \overline{X}_R) \tag{16.2}$$

   where

$$X_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}, i = \{M, R\} \tag{16.3}$$

$$S_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \overline{x}_i)(X_{ij} - \overline{x}_i)' \tag{16.4}$$

$$S_p = \frac{(n_M - 1)S_M + (n_R - 1)S_R}{n_M + n_R - 2} \tag{16.5}$$

3. *Analysis of Variance (ANOVA):* ANOVA is a collection of statistical techniques for testing mean equality between three or more datasets (Naylor and Finger, 1967). It is similar to multiple two-sample t-tests but less prone to a Type I error. The most popular ANOVA test is the F-Test. In a nutshell, the F-Test is the ratio of the variability between the datasets to the variability within each dataset (Lomax and Hahs-Vaughn, 2013). The formula is:

$$F = \frac{\sum_{i=1}^{K} n_i(\overline{Y}_i - \overline{Y})^2/(K-1)}{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2/(N-K)} \tag{16.6}$$

where Y_i is the average of the i^{th} dataset, $\overline{Y}$ the overall average of the data, K the number of datasets, Y_{ij} the j^{th} observation of the i^{th} dataset, and N the total sample size.

4. *Multivariate Analysis of Variance (MANOVA):* MANOVA is similar to ANOVA but for cases where the dependent variables are more than one (Warne, 2014). One of the most popular MANOVA tests is the Samuel Stanley Wilks' statistic, which is a summary based on the eigenvalues λ_p of the A matrix ($A = \sum_M * \sum_{res}^{-1}$), where $\sum_M$ is the model variance matrix and $\sum_{res}$ the error variance matrix. Wilks' formula is:

$$\Lambda_{Wilks} = \prod_{1\ldots p}(1/(1+\lambda_p)) = det\Big(\sum_{res}\Big)/det\Big(\sum_{res}+\sum_{M}\Big) \tag{16.7}$$

and is distributed as Λ.

5. *Simultaneous Confidence Intervals:* Balci and Sargent (1984) proposed the validation method of simultaneous confidence intervals (sci) for simulation models with multiple outputs. The sci are formed by the confidence intervals of each model output. They described three approaches for calculating the sci and choosing one approach over the others depends on whether the model is self- or trace-driven. In other words, the choice of the approach depends on whether the model's input data are coming from the same population as the system's input data but they are different or whether the model's input data are exactly the same as the system's.

6. *Factor Analysis:* Using factor analysis, p observed random variables can be expressed as linear functions of m ($m < p$) random variables, also called common factors, along with an error (Jolliffe, 1986). If $x = \{x_1, x_2, \ldots, x_p\}$ are the observed variables, $f = \{f_1, f_2, \ldots, f_m\}$ the common factors, and $e = \{e_1, e_2, \ldots, e_p\}$ the error, then there exists a

$$K = \begin{bmatrix} \kappa_{11} & \kappa_{12} & \cdots & \kappa_{1m} \\ \kappa_{21} & \kappa_{22} & \cdots & \kappa_{2m} \\ & \cdots\cdots\cdots & & \\ \kappa_{p1} & \kappa_{p2} & \cdots & \kappa_{pm} \end{bmatrix} \tag{16.8}$$

so $x = Kf + e$.

7. *Principal Component Analysis (PCA):* The idea behind PCA is that if there is a large number (p) of random correlated variables, orthogonal transformation can

be used to convert these variables into a significantly smaller number (m) of uncorrelated variables, called principal components (Jolliffe, 1986). PCA is similar to factor analysis, and is often considered to be a method of factor analysis. Despite their similarities, PCA and factor analysis are different in the sense that PCA concentrates on the diagonal elements of the covariance matrix, i.e., the variances, whereas the factor analysis focuses on the non-diagonal elements. In mathematical terms, PCA can be defined as follows:

$$f_1 = a_1' x = a_{11} x_1 + a_{12} x_2 + \ldots + a_{1p} x_p = \sum_{j=1}^{p} a_{1j} x_j \tag{16.9}$$

$$f_2 = a_2' x = a_{21} x_1 + a_{22} x_2 + \ldots + a_{2p} x_p = \sum_{j=1}^{p} a_{2j} x_j \tag{16.10}$$

$$\vdots$$

$$f_m = a_m' x = a_{m1} x_1 + a_{m2} x_2 + \ldots + a_{mp} x_p = \sum_{j=1}^{p} a_{mj} x_j \tag{16.11}$$

where f is the m principal components, a' is a transposed vector of constants, and x is the p independent variables. It should be noted that PCA is particularly useful when $m \ll p$.

8. *Kolmogorov-Smirnov Test:* The Kolmogorov-Smirnov test (K-S test) is a non-parametric goodness-of-fit test that it can be one-sample, i.e., test whether a sample is distributed according to a known theoretical distribution (e.g., normal, binomial etc.), or two-sample, i.e., test whether two different samples are drawn from the same empirical distribution (Chakravarty et al., 1967). In simulation model V&V, the two-sample K-S test is the most common, i.e., comparing whether the data from the model and from reality are derived from the same distribution. The two-sample K-S test is calculated as follows:

$$D_{n_M,n_R} = sup_x |F_{M,n_M}(x) - F_{R,n_R}(x)| \tag{16.12}$$

where F denotes the empirical distribution of each dataset, which is calculated as follows:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{[-\infty,x]}(X_i) \tag{16.13}$$

where

$$I_{[-\infty,x]}(X_i) = \begin{cases} 1, & \text{if } X_i \leq x \\ 0, & \text{otherwise} \end{cases} \tag{16.14}$$

Finally, the null hypothesis is rejected for a given α level if:

$$D_{n_M,n_R} > C(\alpha)\sqrt{\frac{n_M + n_R}{n_M * n_R}} \tag{16.15}$$

where $c(\alpha)$ is given in the Kolmogorov-Smirnov table.

9. *Chi-square Test:* The chi-square (χ^2) test is also a goodness-of-fit test which, similarly to the K-S test, it can also be a one- or two-sample test. The idea behind a two-sample chi-square test, which is more commonly used in model V&V, is that the simulation and operational data are partitioned in i bins, and then the number of points in each bin is observed on whether it is similar on both datasets (Fisher, 1924). Accepting the null hypothesis (H_0) means that the samples are drawn from the same distribution. The chi-square test can be calculated as follows:

$$\chi^2 = \sum_{i=1}^{k} \frac{(K_M x_{Mi} - K_R x_{Ri})}{x_{Mi} + x_{Ri}} \tag{16.16}$$

which follows the chi-squared distribution, and where i is the number of bins, x_{Mi} and x_{Ri} the observed values from the model and reality respectively, and K_M and K_R constants adjusting the inequality of the observations of the two datasets, which are calculated as follows:

$$K_M = \sqrt{\frac{\sum_{i=1}^{k} x_{Ri}}{\sum_{i=1}^{k} x_{Mi}}} \tag{16.17}$$

$$K_R = \sqrt{\frac{\sum_{i=1}^{k} x_{Mi}}{\sum_{i=1}^{k} x_{Ri}}} \tag{16.18}$$

10. *Anderson-Darling Test:* The Anderson-Darling test belongs to the class of quadratic empirical distribution function (EDF) statistics, which determine whether a sample is drawn from a specific distribution (one-sample) or whether two samples are drawn from the same distribution (two-sample) (Anderson and Darling, 1952). The two-sample formula of the test is calculated as follows (Darling, 1957):

$$AD = \frac{1}{n_M n_R} \sum_{i=1}^{n_M+n_R} (N_i Z_{(n_M+n_R-n_m i)})^2 \frac{1}{i Z_{n_M+n_R-i}} \tag{16.19}$$

where $Z_{n_M+n_R}$ is the combined and ordered samples of the model and reality and N_i the number of observations in the model that are equal to or smaller than the i^{th} observation in $Z_{n_M+n_R}$.

11. *CramÃľr-von Mises Criterion:* The CramÃľr-von Mises criterion also belongs to the class of quadratic EDF statistics and is quite similar to the Anderson-Darling test (Cramér, 1928). Compared to the CramÃľr-von Mises criterion, the Anderson-Darling test places more weight on observations in the tails of the distribution. The two-sample CramÃľr-von Mises criterion is calculated as follows:

$$T = \frac{U}{n_M n_R (n_M + n_R)} - \frac{4 n_M n_R - 1}{6(n_M + n_R)} \tag{16.20}$$

where

$$U = n_M \sum_{i=1}^{n_M} (r_i - i)^2 + n_R \sum_{j=1}^{n_R} (s_j - j)^2 \tag{16.21}$$

and $(r_1, r_2, ..., r_{n_M})$ and $(s_1, s_2, ..., s_{n_R})$ the ranks of the sorted samples of the model and reality respectively.

12. *Kuiper's Test:* Kuiper's test is a goodness-of-fit test similar to the Kolmogorov-Smirnov test (K-S test) in the sense that it compares two cumulative distribution functions. Compared to the K-S test, Kuiper's test is sensitive not only to the median but also to the tail. Compared to the The Anderson-Darling test, which also provides equal sensitivity at the tails and at the median, Kuiper's test is invariant under cyclic transformations of the independent variable (Kuiper, 1960). Kuiper's test is calculated as follows:

$$V = D_+ + D_- \tag{16.22}$$

where

$$D_+ = max_{-\infty<x<\infty}[S_M(x) - S_R(x)] \tag{16.23}$$

$$D_- = max_{-\infty<x<\infty}[S_R(x) - S_M(x)] \tag{16.24}$$

$$S_M(x_i) = \frac{i - n_M}{n_M} \tag{16.25}$$

$$S_R(x_i) = \frac{i - n_R}{n_R} \tag{16.26}$$

13. *Coefficient of Determination (R^2):* R^2 is yet another goodness-of-fit test that indicates the proportion of the variance of the dependent variable that is predicted from the independent variable or variables. The most commonly used extension of R^2 is the adjusted R^2 ($\overline{R}^2$), which adjusts for the number of explanatory terms in a model relative to the number of data points (Theil, 1958). $\overline{R}^2$ is calculated as follows:

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n_M - 1}{n_M - k - 1} \tag{16.27}$$

where

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}} \tag{16.28}$$

$$SS_{residual} = \sum_{i=1}^{n_M} e_i^2 \tag{16.29}$$

$$SS_{total} = \sum_{i=1}^{n_M} (y_i - \overline{y})^2 \tag{16.30}$$

and k is the number of independent variables. The closer $\overline{R}^2$ is to one, the better the model is considered, since the results are explained in a large degree from the variation of the dependent variables and not from the residuals.

14. *Mann-Whitney-Wilcoxon Test:* The Mann-Whitney-Wilcoxon (MWW) test, also known as Mann-Whitney U test, is a non-parametric test that tests whether two samples derive from populations having the same distribution (Mann and Whitney, 1947). The MWW test can be calculated by first sorting all values from both datasets in an ascending order and assigning numeric ranks starting with 1 from

the end of this sorted list. Then, the MWW values for both datasets are computed as follows:

$$U_M = R_M - \frac{n_M(n_M+1)}{2} \tag{16.31}$$

$$U_R = R_R - \frac{n_R(n_R+1)}{2} \tag{16.32}$$

where R indicates the sum of the ranks for each dataset. Finally, in order to determine whether the two samples derive from the same population, the minimum value between U_M and U_R is compared with the value from the tables.

15. *White Test:* The White test is a test for determining whether the variance of a model is constant, i.e., whether the model is homoscedastic (H_0) (White, 1980). The White test is calculated as follows:

$$\hat{e}_i^2 = \delta_0 + \delta_1 \hat{Y}_i + \delta_2 \hat{Y}_i^2 \tag{16.33}$$

where Y_i are the predicted dependent variables of the model. Upon calculating δ_0, δ_1, and δ_2, the $R^2_{\hat{e}^2}$ can be computed and then the $\chi^2 = n_M R^2_{\hat{e}^2}$, which can then be tested with 2 degrees of freedom against the null hypothesis.

16. *Glejser Test:* The Glejser test also tests for Heteroscedasticity but instead of using the square of the residuals, it uses their absolute values (Glejser, 1969). The Glejser test is calculated as follows:

$$|e_i| = \gamma_0 + \gamma_1 f(x_i) + u_i \tag{16.34}$$

in which case the most common values for the $f(x_i)$ are: $f(x_i) = x_i$, $f(x_i) = \sqrt{x_i}$, and $f(x_i) = \frac{1}{x_i}$. The γ_1 of the equation with the highest R^2 is then tested and if it is found statistically significant, the null hypothesis of homoscedasticity is rejected.

17. *Spectral Analysis:* Spectral analysis tests whether two time series are equivalent (Fitzsimmons, 1974). Spectral analysis is a relatively complex statistical test, especially compared to the tests presented so far, and it is calculated as follows:

$$g_i(f) = \frac{1}{\pi}\left[2\sum_{p=1}^{L} k_L(p)C_i(p)cos(f_i(p)) + C_i(0)\right] \tag{16.35}$$

where $i = \{M, R\}$. $C_i(p)$ is the autocovariance function

$$C_i(p) = \frac{1}{T-p}\sum_{t=1}^{T-p}(x_t - m)(x_{t+p} - m) \tag{16.36}$$

$k_L(p)$ is a Bartlett weighting function for which several possibilities exists (Jenkins and Watts, 1969), and

m = mean of $X(t)$

T = total time period

X_t = observation at time t

f = frequency in cycles per unit of time

L = number of lags

p = number of time periods separating correlated observations (1,2,...,L-1)

Finally, in order to determine whether the two time series are equivalent, i.e., not rejecting the null hypothesis, the ratio $g_M(f)/g_R(f)$ should satisfy the inequality:

$$e^{-\phi} \leq \frac{g_M(f)}{g_R(f)} \leq e^{\phi} \tag{16.37}$$

where

$$\phi = Z_{\alpha/2}(4L/3T)^{1/2} \tag{16.38}$$

and $Z_{\alpha/2}$ = the two tail critical value for the standard normal distribution at a significance level of α.

18. *Durbin-Watson Statistic:* The Durbin-Watson statistic tests for the existence of autocorrelation in the residuals from a regression analysis (Durbin and Watson, 1950). The statistic is calculated as follows:

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=2}^{T} e_t^2} \tag{16.39}$$

where T is the number of observations. The value d is compared to the lower and upper critical values ($d_{L,a}$ and $d_{U,a}$) to test for positive or negative autocorrelation.

The statistical techniques described above as just a sample of the available techniques for simulation model V&V. Nevertheless, it is a representative sample that can be used in the majority of the cases. The aim of this section is to illustrate the various statistical techniques, which facilitates the categorisation of these techniques and thus the selection of the most suitable ones given the problem at hand.

Appendix C: The R Code

```php
<?php
  //Fetch and open the CSV file with the simulation data
  $csvFriso = fopen('friso.csv', 'r');

  //Parse data and insert into database
  while($results = fgetcsv($csvFriso)){
    $database->query("INSERT INTO punctuality (train, station, planned_arrival, planned_departure, actual_arrival, actual_departure) VALUES ('".$results[0]."','".$results[1]."','".$results[2]."','".$results[3]."','".$results[4]."','".$results[5]."')");
  }
?>
```

Listing 16.1. Importing a CSV file using purely PHP

```php
<?php
  //Fetch the path of the SQL file with the simulation data
  $sqlFriso = 'friso.sql';

  //Fetch the contents of the SQL file
  if($sql = file_get_contents($sqlFriso)) {
    //Separate each query based on the semi-column
    $results = explode(';',$sql);
    //Loop through each SQL query
    foreach($results as $result) {
      //Run each SQL query
      mysqli_query($result . ';');
    }
  }
?>
```

Listing 16.2. Importing an SQL file using purely PHP

```
<?php
  //Fetch and open the CSV file with the simulation data
  $csvFriso = 'friso.csv';

  //Execute the shell script and pass the path of the CSV file as an argument
  exec(db.sh $csvFriso);
?>
```

LISTING 16.3. IMPORTING A CSV FILE THROUGH A SHELL SCRIPT

```
#!/bin/bash

#Connect to the database 'friso' and import the CSV file
mysqlimport -h localhost -u root -p friso --local --fields-terminated-by=',' $1
```

LISTING 16.4. THE SHELL SCRIPT

```
<?php
  //execute the R script with the appropriate arguments
  exec("/usr/bin/Rscript ../RScript/R/tt2017.R $trainSerie $departureStationFull
     $arrivalStationFull $tableAbbr $smoothingFactor");
?>
```

LISTING 16.5. EXECUTING AN R SCRIPT FROM WITHIN PHP

```r
#Include library for connecting to MySQL Database
library(RMySQL)

#Enable script to read arguments from PHP
arguments <- commandArgs(TRUE)

tableFriso <- args[1] #The Friso database table
station <- args[2] #Station: Ut, Ht etc.
direction <- args[3] #North or South
activity <- args[4] #Arrival, departures, driving
delayThres <- args[5] #Maximum amount of accepted delay

#The WHERE clause for the Friso data
whereFriso <- paste("WHERE station='", station, "' AND direction='", direction, "' AND activity='", activity, "' AND  delay<", delayThres, sep = "")

#Concatenate the 1st part of the query with the Friso database table and the WHERE clause
Fquery <- paste("SELECT * FROM", tableFriso, whereFriso, sep = " ")

#Connect to database
mydb <- dbConnect(MySQL(), user='root', password='root', dbname='friso', host='127.0.0.1')
#Query the Friso database table
queryFriso <- dbSendQuery(mydb, Fquery)
#Fetch results from query
dataFriso <- fetch(queryFriso, n=-1)
#Disconnect from the database
dbDisconnect(mydb)

#Fetch the delay of Friso data
delayFriso <- dataFriso$delay

#Name of the file containing the histogram
histogramName <- paste("Friso-", station, "-", direction, "-", activity, "-Histogram", sep = "")
#Full path of the file containing the histogram
fullHistogramPath <- paste("/var/www/html/Friso/", histogramName, ".png", sep = "")

#Create the .png file
png(filename=fullHistogramPath, width=500, height=500)

#Create the histogram and define its title and x-axis
h <- hist(delayFriso, main = "Histogram of Friso Delays", xlab = "Friso delays (seconds)", col = "blue")
#Define the histogram's density
h$density = h$counts/sum(h$counts)
#Plot the histogram
plot(h,freq=FALSE,col=c("blue"))

#Finalize the histogram in the .png file
graphics.off()
```

LISTING 16.6. THE R SCRIPT FOR BUILDING THE HISTOGRAMS OF DELAYS IN FRISO'S MODEL

```r
library(ggplot2) #Include library for plots
library(scales) #Include library for breaks and labels
library(directlabels) #Include library for putting labels

#Create the plot
plot <- ggplot() +
#Customize the x axis
scale_x_continuous(breaks = xbreaks, labels = xlabels) +
#Customize the y axis
scale_y_continuous(breaks = ybreaks, labels = ylabels, limits = c(comparison[["ymin"]], comparison[["ymax"]])) +
#Draw vertical line at minimum floored arrival time
geom_vline(xintercept = xbreaks, size = 0.3) +
#Draw horizontal line at throughpoints
geom_hline(yintercept = ybreaks, size = 0.2, color="#D3D3D3") +
#x- and y-axis titles and footnote with production date
labs(x = "Time (in mins)", y = "Train Stations", fill="Legend", caption = production) +
#Centering graph title
theme(plot.title = element_text(hjust = 0.5),
#Make axis line black
axis.line = element_line(colour = "black"),
#Legend background color
legend.key = element_rect(fill = "white"),
#Remove background color
panel.background = element_blank(),
#Change font size of the legend
legend.text=element_text(size=15)) +
#Plot the operational data
geom_line(data = dataOper, aes(x = VertrekVertraging, y = DistAbsolute, group = AnTreinID, color = operationalName),size = 0.1) +
#Plot the simulation data
geom_line(data = dataSim, aes(x = relTime, y = position, color = simulationName),size = 0.4) +
#Plot simulation data with a minute added in time
geom_line(data = dataSimOneMinute, aes(x = relTime, y = position, color = simulation1minName), size = 0.4, linetype="dotted") +
#Plot the fastest 10th percentile smoothed for spikes
stat_smooth(data = fast, aes(x = base, y = adjDistance, color=percentileName), size = 0.2, method=lm, formula = y ~ poly(x,5), se=FALSE) +
#Label for the fastest 10th percentile
geom_text(data = fast, aes(x = pbreaks[[1]], y = pbreaks[[2]], label = plabels[[1]]), hjust = 0.7, vjust = 1, size = 2, color = "blue") +
#Plot the median percentile (50th) smoothed for spikes
stat_smooth(data = median, aes(x = base, y = adjDistance, color=percentileName), size = 0.2, method=lm, formula = y ~ poly(x,5), se=FALSE) +
#Label for the median percentile
geom_text(data = median, aes(x = pbreaks[[3]], y = pbreaks[[4]], label = plabels[[2]]), hjust = 0.7, vjust = 1, size = 2, color = "blue") +
#Plot the slowest 10th percentile (90%) smoothed for spikes
stat_smooth(data = slow, aes(x = base, y = adjDistance, color=percentileName), size = 0.2, method=lm, formula = y ~ poly(x,5), se=FALSE) +
#Label for the slowest 10th percentile
geom_text(data = slow, aes(x = pbreaks[[5]], y = pbreaks[[6]], label = plabels[[3]]), hjust = 0.7, vjust = 1, size = 2, color = "blue") +
#Create legend
```

```
scale_colour_manual("", values=c("#4c4cff", "#ff6666", "black", rgb(0, 0, 0, 20,
     maxColorValue = 255))) +
#Make the 1-minute line dotted and the rest solid
guides(color = guide_legend(override.aes = list(linetype = c("solid", "solid", "dotted",
      "solid"))))
```

LISTING 16.7. THE R SCRIPT FOR BUILDING THE GRAPH OF DRIVING BEHAVIOR IN OPENTRACK'S MODEL

Appendix D: Generalisation Results

The detailed results from the questionnaire for the generalisation of the findings can be seen in Table 16.3. The results were quantified with 1 representing the *Strongly Disagree* and 5 the *Strongly Agree*. The shortened *Exp.* followed by a number indicates one of the 10 experts, who participated and filled the questionnaire. The green cells show the best score for each proposition, whereas the red cell the worst; both of these scores are excluded from the final average, i.e. the column *Corrected Average*, in order to remove possible outliers. Finally, the abbreviation for each proposition is explained below:

P1 Game theory has the ability to support game designers by first providing for a formal methodology for modelling the real system under study and then enabling the translation of certain game theoretical elements, like strategies and pay-offs, in game design decisions.

P2 Validation methods from pure simulations can provide rigour in game validation for improving games' usability and credibility. However, the use of analytical methods in game validation should be done with caution and a balance should be struck between empirical and analytical methods.

P3.P1 The presence of a game manager has a positive impact in the game session.

P3.P2 The managerial guidance and involvement has a positive impact in the game session.

P3.P3 The structured and concrete results have a positive impact in the game session.

P3.P4 Strict rules have a positive impact in the game session.

P3.P5 High variety of roles involved in the game design have a positive impact in the game session.

P3.P6 The game being validated beforehand has a positive impact in the game session.

P3.P7 Structured debriefing has a positive impact in the game session.

P3.N1 A potential high complexity of the game's scope has a negative impact in the game session.

P3.N2 The unexpressed and/or conflicting stakeholders' interests have a negative impact in the game session.

P3.N3 Time pressure has a negative impact in the game session.

P3.N4 Pressure from external factors (for obtaining a solution suitable to their interests) has a negative impact in the game session.

P4 Knowledge management can contribute towards extended usability and credibility of games by providing for a platform in which our knowledge about certain systems and games can gradually be built, thus overcoming important limitations with regards to analytically approaching game validity, which is the usually low number of participants, and partial coverage of the system of study.

TABLE 16.3. THE RESULTS ON THE QUESTIONNAIRE FOR THE GENERALISATION OF THE FINDINGS.

	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5	Exp. 6	Exp. 7	Exp. 8	Exp. 9	Exp. 10	Average	Corrected Average
P1	2	5	5	1	3	4	5	5	3	2	3.5	3.7
P2	3	5	5	4	4	5	5	5	5	5	4.6	4.8
P3.P1	2	3	5	5	3	4	5	4	5	5	4.1	4.3
P3.P2	3	-	5	3	3	3	4	3	3	4	3.4	3.5
P3.P3	4	5	4	3	4	3	3	3	5	5	3.8	3.8
P3.P4	3	3	3	3	3	3	3	3	2	5	3.0	3.0
P3.P5	3	4	2	4	3	3	3	3	4	3	3.2	3.2
P3.P6	5	2	4	4	5	3	3	4	5	4	3.9	3.8
P3.P7	4	4	5	5	4	3	4	4	5	5	4.3	4.3
P3.N1	4	3	4	4	4	3	3	2	2	4	3.4	3.3
P3.N2	3	3	1	4	3	3	3	4	2	4	3.0	3.0
P3.N3	3	4	2	2	3	3	3	5	4	5	3.3	3.3
P3.N4	3	4	3	4	3	3	3	5	2	4	3.5	3.6
P4	3	4	4	2	3	4	4	3	5	3	3.5	3.5

Curriculum Vitæ

Vasileios Roungas

23-03-1984	Born in Athens, Greece.

Education

2001–2006	B.Sc. in Economic and Regional Development Panteion University, Greece
2007–2009	M.Sc. in Business Mathematics Athens University of Economic and Business, Greece
2012–2015	M.Sc. in Game & Media Technology Utrecht University, The Netherlands
2015–2019	Ph.D. in Gaming Simulations Delft University of Technology, The Netherlands

List of Publications

Roungas, B., Bekius, F., Meijer, S., & Verbraeck, A. (2019). Improving the decision-making qualities of gaming simulations. Journal of Simulation. Under Review.

Roungas, B., Lo, J. C., Meijer, S., & Verbraeck, A. (2019). Guidelines for the management and dissemination of knowledge from gaming simulations. Simulation & Gaming. Under Review.

Roungas, B., Meijer, S., & Verbraeck, A. (2019). The tacit knowledge in games: From validation to debriefing. In 50th International Conference of International Simulation and Gaming Association. Warsaw, Poland.

Roungas, B., Bekius, F. A., & Meijer, S. (2019). The game between game theory and gaming simulations: Design choices. Simulation & Gaming, 50(2).

Roungas, B., Meijer, S. A., & Verbraeck, A. (2018). The future of contextual knowledge in gaming simulations: A research agenda. In M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, & B. Johansson (Eds.), Proceedings of the 2018 Winter Simulation Conference. Gothenburg, Sweden: IEEE.

Roungas, B., Meijer, S. A., & Verbraeck, A. (2018). Harnessing Web 3.0 and R to mitigate simulation validation restrictions. In International Conference on Simulation and Modeling Methodologies, Technologies and Applications. Porto, Portugal.

Roungas, B., Lo, J. C., Angeletti, R., Meijer, S. A., & Verbraeck, A. (2018). Eliciting requirements of a knowledge management system for gaming in an organization: The role of tacit knowledge. In 49th International Conference of International Simulation and Gaming Association. Bangkok, Thailand.

Roungas, B., Meijer, S. A., & Verbraeck, A. (2018). A framework for optimizing simulation model validation & verification. International Journal On Advances in Systems and Measurements, 11(1 & 2).

Roungas, B., Meijer, S., & Verbraeck, A. (2018). Validity of railway microscopic simulations under the microscope: Two case studies. International Journal of System of Systems Engineering, 8(4), 346-364.

Roungas, B., Meijer, S., & Verbraeck, A. (2018). Knowledge management of games for decision making. In H. Lukosch, G. Bekebrede, & R. Kortmann (Eds.), Simulation Gaming. Applications for Sustainable Cities and Smart Infrastructures. ISAGA 2017. Lecture

Notes in Computer Science, vol 10825. Cham, Switzerland: Springer.

Roungas, B., De Wijse, M., Meijer, S., & Verbraeck, A. (2018). Pitfalls for debriefing games and simulation: Theory and practice. In A. Naweed, M. Wardaszko, E. Leigh, & S. Meijer (Eds.), Intersections in Simulation and Gaming (pp. 101-115). Cham, Switzerland: Springer.

Roungas, B., Meijer, S., & Verbraeck, A. (2017). A framework for simulation validation & verification method selection. In SIMUL 2017: The Ninth International Conference on Advances in System Simulation (pp. 35-40). Athens, Greece.