The Debriefing of Research Games: A Structured Approach for the Validation of Gaming Simulation Outcomes

Jop van den Hoogen; Julia Lo; Sebastiaan Meijer

Abstract
Debriefing is an overlooked topic in the gaming simulation literature, especially in the context of gaming for applied research and design. Building upon existing frameworks, mostly in the fields of policy and educational games, and linking the design of the debriefing with the intricacies of innovation in complex networked infrastructures, we state that debriefing should touch upon five important topics: qualitative data generation, internal validation, external validation, reliability analysis and a robust planning for action.

Keywords
research gaming simulations; debriefing; validity; reliability; experiment

1 Introduction

Due to co-evolutionary processes of system elements, networked infrastructures are highly hostile towards radical innovations (Geels, 2007; Markard, 2011). However in certain cases, only these innovations allow a system to further adapt to rapidly shifting environmental conditions. We assume that the main function of gaming simulation is to build niches: protective spaces where radical innovations can be envisioned, experimented with and nurtured without the immediate selection pressure working upon the innovation from the incumbent regime.

Our starting assumption is that in accordance with Klabbers (2003a), the design of a game as an artifact, including its debriefing, is intrinsically linked to the process of the design of the real system that is being simulated. As such, debriefing of gaming simulation for these purposes should differ to a large extent from the debriefing of more usual applications such as games for education and training. This paper focuses on the debriefing of gaming simulation for applied research. Using insights from gaming simulation debriefing literature and methodological literature on experimental research, we use a structured approach to

identify how to debrief games for research, focusing on what topics to address and which participants to involve.

2 Simulation and Gaming Simulation

According to Axelrod (2006), simulation can be seen as large scale thought experiments. The power of simulation is that it allows the experiment to incorporate many variables and relationships that a normal person cannot handle and portray the emergent behavior of systems that have multiple interdependent processes operating simultaneously (Harrison & List, 2004). As in the first half of the 20th century the recognition arose that many of the phenomena we see are related to chaotic and emergent properties of non linear dynamic systems, new methods were needed to do justice to these properties (Ackoff, 1974). Thus in the advent of the system sciences, simulation as a means to understand complexity became widespread in fields such as operations research, meteorology, and evolutionary modelling. Bratley, Fox and Schrage (1987, p. 9) define the act of simulation as "driving a model of a system with suitable inputs and observing the corresponding outputs". Hence, simulation involves both modelling, i.e. building an abstract representation of reality, and experimenting, i.e. manipulating the parameters of this model. By studying systems holistically rather than studying them by breaking it down and studying the isolated parts, simulation offers the possibility to capture so-called epiphenomena of collections of interacting elements.

Gaming simulation also offered this advantage but added the possibility to study systems in which technical and social elements both interacted and accordingly followed different rules than purely technical systems. In the beginning this led to its application mainly in the military and logistics domain (Brewer & Shubik, 1979; Mayer, 2009). Later on the recognition arose that the wickedness of the problems involving complex systems was caused by a myriad of incongruent opinions and perceptions around policy issues and led gaming simulation to become more consensus-oriented than scientific-oriented (Geurts & Joldersma, 2001). Rather than testing hypotheses, policy games offer the chance to create consensus between decision makers through the multilogue mode of communication where people with different perspectives engage with each other simultaneously (Duke, 2011). Outcomes of games therefore provide decision makers not with ready to use decisions, rather games help to create a future memory (Wenzler & Chartier, 1999).

2.1 Research games in organizations

Klabbers (2003b) stated that designing effective gaming simulations is an interplay between designing the game itself, i.e. design-in-the-small (DIS) and the intended effects of the game on the design of the referent system, e.g. design-in-the-large (DIL). If we wish to structure a debriefing, and thus make the debriefing

a design consideration as proposed by Crookall (2010), we should inform this process with the pecularities of the context in which simulation is employed. As Klabbers (2006) states, the goal of a gaming simulation (DIS) should serve the meta-goal of DIL-processes. Kriz and Hense (2006) sought to combine these two design processes by linking usual applications of gaming simulation to Greif and Kurtz (1996) model of organizational development (see Figure 1).

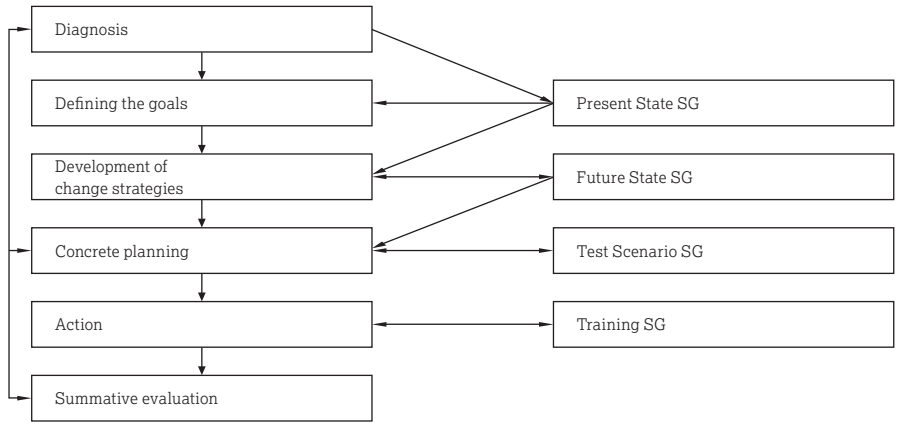

Figure 1 Linking functionalities of gaming simulation to Greif and Kurtz' model (1996) of organizational evelopment (Kriz & Hense, 2006, p. 275).

Organizational development is different from innovation because in the first the organization is intended to change whereas in the latter a product or process the organization governs is intended to change. However, the model shown above helps us in clearly distinguishing different functionalities of gaming simulation, such as diagnosis, design, testing and training. In Table 1, we adapt the four functionalities of gaming simulations to an innovation context.

Table 1 Four functionalities of gaming simulation in an innovation context

Research approaches	Game functionality (Kriz & Hense (2006)	Main activity
Inductive; what theories can be build upon observing the real world? What is happening? What is the system like?	Present State SG	Generation of hypotheses
Constructive; what designs can improve the present state? What should we do to improve the system?	Future State SG	Design of artifacts, policies and strategies
Deductive; does our theory hold in the real world? Is this design really improving the system?	Test Scenario SG	Testing of hypotheses
Instructive; can we instruct people how to go from present state to future state? Can we train people how to handle the new design?	Training SG	Transfer of knowledge and skills

Since the generation and testing of hypotheses is what defines research, we name games 1 and 3 games for research. In game 2, gaming simulation is used for design and we see similarities with policy games. The fourth is well known and comes in the forms of educational games and games for learning. This distinction closely follows Peters and Vissers (1998) categorization of games for research, policy and learning.

3 Debriefing

In the gaming simulation field, debriefing gets less attention than it deserves (Dennehy, Sims & Collins, 1998) and as Crookall (2010) states that learning comes from debriefing, not from the game itself, a stronger focus on debriefing is needed. In addition, studies focusing on the debriefing part of gaming simulation mainly involved gaming simulation for policy making and learning. We see that the debriefing of games for applied research needs more attention. In general, the debriefing literature mostly focused on describing a phased approached using Kolb's (1984) experiential learning cycles as a framework (Van der Meij, Leemkuil & Li, 2013). In debriefing sessions the main focus lies on sharing insights and the transfer of insights to the referent system that had been simulated (Kriz, 2003). Notable examples of elaborations on the debriefing of these kind of games are Sims (2002) and Lederman (1992).

3.1 Phases

Peters and Vissers (2004) touch upon the implications of using gaming simulation for research on the debriefing. They state that it serves three purposes:

1. provide a moment of cooling-down for the participants
2. protect the instrument gaming simulation
3. validate researchers interpretation of simulation outcomes

Whereas the latter is straightforward, at least at first sight, the other two are relevant as well but nonetheless neglected somewhat in the literature on gaming simulation. However for our purposes of debriefing gaming simulations for research the first two are important. Firstly, what we ask from participants in a game and in a debriefing is substantially different. In a game we ask them to be immersed, taking on the game, tasks and responsiblities as-if real. On the other hand, we ask game players in a debriefing to reflect on what has happened in the game and to what extent this was perceived as 'real'. If one is to say that the first involves single-loop learning and the latter double-loop learning, we believe a cooling-down moment is crucial for allowing game players to switch from one mental state to the other. Secondly, in our gaming simulations, game player availability is a big constraint, since an organization can only spare so many employees. Therefore we need to be cautious in treating them and making sure succesive participation is guaranteed. Additionally, participants join the normal organization after playing the game and we need to control to what extent this might have effects. Sometimes, controversial innovations may become subjected to gaming simulation and in the debriefing we need to make sure what information remains inside the realms of the game and what is allowed to enter the organization.

Nevertheless, validation is what we believe to be the most important part of debriefing. Regarding this aspect we see two striking phased approaches from the literature. Lederman and Stewart (1986) and Van Ments (1983) mention three guiding concepts that structure a debriefing process. We adapt both approaches to coalesce them into one (see Table 3).

Table 3 Guiding concepts of a debriefing (Lederman & Stewart, 1986; Van Ments, 1983)

Is the game..?	Lederman and Stewart (1986)	Van Ments (1983)	Our interpretation
Valid	Validity questions	Establishing the facts	What did happen and is this similar to real life?
Reliable	Reliability questions	Analyzing causes	To what extent was the simulation deterministic, chaotic or stochastic?
Actionable	Utility questions	Planning action in real-life	Is the information retrieved from the simulation impetus for action

Kriz (2010) is one of the first to apply a systems perspective to the debriefing process and elaborates on six phases or topics to be addressed during a debriefing following Thiagarajan (1993). Most importantly here is that a systems perspective advocates the holistic studying of complex systems and in the debriefing process, the key idea is to allow for many perspectives on the same phenomena to arise. Firstly, in line with many others, this debriefing process involves a moment of cooling down. Secondly, the analysis of what had happened in the game and to what extent this is externally valid is deemed to take place by mixing many insights from the participants to arrive at a rich picture of the processes and its relation to real world processes. Furthermore in a separate phase, the debriefing should give attention to probable path dependent processes: some early decisions might have long lasting impact on the course of the game. Thus, a debriefing should assess these path dependencies by finding out critical decisions and what-if scenarios. Table 4 provides an overview of the six phases.

Table 4 Phases in a debriefing (Kriz, 2010)

Phase	Topic	Explanation
Phase 1	how did you feel?	Cooling down of participants
Phase 2	what happened?	Data collection
Phase 3	how are the game and reality connected?	External validity
Phase 4	what did you/we learn?	Reaching conclusions
Phase 5	what would happen if..?	Testing replicability / sensitivity
Phase 6	how do we go on now?	Planning for action

If we were to summarize the debriefing literature we see that, overall, a proper debriefing should adhere to a structure that focuses on five topics:

1. Cooling-down
2. Data collection
3. Validity and reliability analysis
4. Planning for action
5. Protecting the instrument

However, since most of the debriefing literature focuses on games for learning or games for policy making, we believe that we are in dire need of some way of properly structuring a debriefing of games for research. Especially games for hypothesis testing follow closely the features of a controlled experiment in which the researcher provides a subject (e.g. a railway system) with a treatment (e.g. an innovation) and wishes to study the effects on some predetermined performance indicators (such as punctuality or capacity). Thus, whereas the emphasis of debriefing in training games lies on a reflection of the lessons learned and its transfer to the real world environment, we assume that reliability and validity issues are key topics in the debriefing of games for research. We therefore build upon the methodological literature on experimental research to further structure our debriefing.

3.2 Format

The debriefing itself can be designed in a number of ways, e.g. through the role of the facilitator, set-up of the seating arrangements, communication structures between participants and use of video-recordings (Crookall, 2010; Kriz, 2010). These debriefing methods haven been rather flexibly applied across and within different phases, as there are a number of factors that can influence the preference of a certain method (Kriz, 2010). As such, group size might influence whether the interactions should take in pairs, small teams or the whole group, and whether the interactions should be structured in accordance to a panel discussion, fishbowl or talk show format.

3.3 Participants

Finally, gaming simulations contain of different types of participants. Next to the presence of game players, there may also a number of nonparticipating person, e.g. colleagues, managers, policy designers, researchers, additional facilitators and observers (Kriz, 2010; Peters & Vissers, 2004). The presence of different types of participants may strongly influence the different phases within a debriefing session, and the applied format within the phases.

4 Experimental Research

The main objective in experiments is to manipulate on one or more independent variables and measure its effects on a dependent variable (Zechmeister, Zechmeister & Shaughnessy, 2001). Experimental research can be divided into two streams: one stream adopting a classical linear perspective on causality and one stream adopting a complexity perspective. Whereas the first sees experimental objects as trivial machines (the same pill given to the same participant will always produce the same results), the latter accounts for non-triviality (dynamic feedback systems show path dependent and chaotic behavior). In the design of experiments this results in treating the subject as either a black box in classical laboratory experiments or treating the subject as a collection of interacting elements in computer simulation.

In general, the quality of any research instrument can be described using two closely related concepts (Pellegrino, 2009): reliability and validity. Whereas the first determines to what extent repeated measured result in similar findings, the latter determine to what extent causal claims are correctly based on the measurements.

4.1 Reliability of experiments

4.1.1 Measurement reliability

Most commonly, the reliability of a research method is defined as the extent to which the measurement methods will measure the same values of a variable if measurements are repeated (Messick, 1975; Hunt, 1983). Quantitatively the reliability of these instruments is portrayed as the error margin of the instrument. For instance, if a temperature measure of an object of 38 degrees half of the times gives a value of 37 and half of the times gives a value of 39, its error margin is 1/38.

4.1.2 Sensitivity

Making sure that measurement instruments reliably measure variables is not the only way we are to make sure that a repeated experiment will show similar results, especially when one experiments with systems. That is because causal relations in an ecology of thousands of bidirectional causal relations are rarely so-cal-

led trivial machines. Stochastic and chaotic properties of dynamic feedback systems might give different results for the same starting conditions or different results for almost the same starting conditions respectively. So in one experiment we might conclude that A leads to B, whereas in a repeat run using the same sample and setup we might conclude that A does not lead to B. Accounting for this is called sensitivity analysis in the realms of computer simulation experiments and is certainly an important aspect of reliability in our context. In computer simulation experiments on stochastic or chaotic systems, multiple runs are executed to see if results are sensitive to initial conditions or to critical game player decisions during game play.

4.2 Validity of experiments

In general, the quality of experimental research can be described using external and internal validity (Zechmeister et al., 2001). Internal validity describes the extent to which one can trust a causal claim to be real inside the scope of the experiment, whereas external validity is "the extent to which findings from an experiment can be generalized to individuals, settings, and conditions beyond the scope of the specific experiment" (Zechmeister et al., 2001, p.161, italics added).

4.2.1 Internal validity

Internal validity is often defined as the extent to which the causal relation was isolated from potential confounders in an experimental setting. These confounders might be different research settings for the treatment group than for the non-treatment group or adverse selection of research participants for the treatment. Experimentalists use random treatment assignment to assure internal validity.

4.2.2 External validity

External validity has been understood in many sometimes conflicting ways (Morton & Williams, 2010) and to better clarify this concept we distinguish between the extent that findings can be translated from sample to population and from isolation to a real-world setting, i.e. the 'fieldness' of the experiment (Harrison & List, 2004). We then arrive at generalizability and ecological validity. Experimentalists ensure generalizability by finding a representative sample from a population the research wishes to study. This representativeness is achieved by finding a subset of a population that shows the most important features that the population also shows. Thus, experimentalists first try to obtain a large enough sample. Secondly, in case the population is very poorly understood, the research may wish to randomly sample from this population. If the population is very clear, than more fine tuned ways of sampling might be done, for instance using stratified or convenient sampling.

Ecological validity on the other hand more resembles Raser's (1969) way of defining validity of gaming simulation using the concepts of psychological realism, structural validity and process validity. Gaming simulations are often highly artificial environments and many contextual cues, processes and structural elements are omitted from the model in order to simulate it. However, these omissions might render the causal claim invalid once the claim is translated from the artificial model to the real world. Experimentalists usually enrich the experimental setting with cues from the real world, in much the same way as game designers try to ensure to properly model the real world by omitting only irrelevant parts. Furthermore the above mentioned sensitivity analysis could profoundly enrich the assessment of external validity. Given that structural and process validity are deemed sufficiently high, do the systemic qualities of the game resemble the qualities of the referent system? Here the assessment would focus on whether parameter sensitivity, tipping points and critical game player decisions hold true in real life. Additionally, the epiphenomena that emerge out of game player interaction, e.g. system level constructs such as punctuality, robustness, group dynamics, social atmosphere, can be assessed on their resemblance to the referent system: do we see the same emergent behavior in the game as we see in real life?

4.2.3 Measurement validity

Measurement or test validity refers to the measurement instrument itself, in which construct, criterion and content validity can be distinguished, see also Table 5 American Psychological Association, American Educational Research Association, and National Council on Measurement in Education, 1966). These concepts have been predominantly applied in the psychology domain, in relation to the use of questionnaires.

We summarize the concepts in Table 5 and show where in the design of experiments these play a role.

Table 5 Validity and reliability concepts in research gaming simulations

	Dimension	Experimental design
Validity		
Internal validity	—	Research designs (pretest-posttest, control group, random treatment assignment, etc.)
External validity	Generalizability	Sample size, sampling procedure
	Ecological validity	Experimental context design or game design
Measurement Validity	Construct validity	Compare results of test with test on similar constructs and opposite constructs (convergent and discriminant validity)
	Content validity	Assess whether items from the test cover all dimensions of the construct
	Criterion validity	Compare results of the test to one or more objective measurements
Reliability		

Sensitivity	-	Multiple runs, sensitivity analysis
Measurement reliability	-	Multiple measurements, triangulation

5 Framework for the debriefing of research gaming simulations

Based on the literature review in previous sections, we identify six phases that need to be addressed in a debriefing session of a research game, in which a large overlap exist with existing literature by Kriz (2010) and Peters and Vissers (2004). However, the current paper recognizes the gap in existing literature regarding the specific topics that need to be addressed within the validity and reliability analysis phase. Table 6 summarizes the findings from the previous sections and integrates the different debriefing phases with the topics and the ideal involved participants per phase.

Table 6 Framework of the phases, addressed topics and involved participants in a research game

Phase	Description	Topics	Involvement of participants
Cooling down	Change mental state of game players from immersion to retrospection	Experience, emotions	Facilitator, game players
Data collection	Additional qualitative data from game players, observers and facilitators	Measurement reliability and validity	All participants
Reliability	Assess whether repeated runs would result in similar outcomes	Sensitivity	Game players, observers
Validity	Assess whether causal claim is internally valid and also holds in real-life (ecological) and for different samples (generalizability)	Internal, external validity	Game players, observers
Planning for action	Determine what follow-up questions need to be answered; determine what concrete actions need to be taken and by whom	Future research questions and actions	All participants
Protect the instrument	Evaluate gaming simulation session; determine what outcomes may be shared; ensure durable relationship with game players	Experience, emotions	Facilitator, game players

6 Discussion and conclusion

Using gaming simulation literature and literature on the methodology of experimental research, this paper proposed a framework for the debriefing of research games that are used in an organizational context. The framework has focused on the identification of the structure/ phases in the debriefing, the topics and the involvement of participants in ideal circumstances. We especially provide an extensive elaboration on how validity and reliability, crucial to any research undertaking, can be assessed and improved by a proper debriefing.

Further research should test the proposed debriefing framework for research games. Furthermore, theoretical implications of the role of the facilitator or observers (e.g. facilitation techniques), the phenomenon of debriefing stress, different organizational cultures, professional codes or ethical considerations should be more in-depth investigated as they might demand the debriefing process to be differently structured.

Acknowledgments

This research was funded through the Railway Gaming Suite program, a joint project by ProRail and Delft University of Technology.

References

Ackoff, R. L. (1974). The systems revolution. Long Range Planning, 7(6), 2—20.
American Psychological Association, American Educational Research Association, and National Council on Measurement in Education (1966). Standards for Educational and Psychological Tests and Manuals. Washington, D.C.: American Psychological Association.
Axelrod, R. (2006). Advancing the art of simulation in the social sciences. In: Rennard, J. P. (Ed.) Handbook of Research on Nature Inspired Computing for Economics and Management. Hershey, PA: Idea Group.
Bratley, P., Fox, B. L., & Schrage, L. E. (1987). A Guide to Simulation. New York: Springer-Verlag.
Brewer, G. D., & Shubik, M. (1979). The War Game. A Critique of Military Problem Solving. Cambridge: Harvard University Press.
Crookall, D. (2010). Serious games, debriefing, and simulation/gaming as a discipline. Simulation & Gaming, 41(6), 898—920.
Dennehy, R. F., Sims, R. R., & Collins, H. E. (1998). Debriefing experiential learning exercises: A theoretical and practical guide for success. Journal of Management Education, 22(1), 9—25.
Duke, R. D. (2011). Origin and evolution of policy simulation: A personal journey. Simulation & Gaming, 42(3), 342—358.
Geels, F. W. (2007). Transformations of Large Technical Systems A Multilevel Analysis of the Dutch Highway System (1950—2000). Science, Technology & Human Values, 32(2), 123—149.
Geurts, J. L., & Joldersma, C. (2001). Methodology for participatory policy analysis. European Journal of Operational Research, 128(2), 300—310.
Greif S., & Kurtz, H.J. (1996). Handbuch Selbstorganisiertes Lernen. Göttingen: Verlag für Angewandte Psychologie.
Harrison, G. W., & List, J. A. (2004). Field experiments. Journal of Economic Literature, 42(4), 1009—1055.
Hunt, E.B. (1983). On the nature of intelligence. Science, 219, 141—146.
Klabbers, J. H. (2003a). Simulation and gaming: Introduction to the art and science of design. Simulation and Gaming, 34(4), 488—494.
Klabbers, J. H. (2003b). Gaming and simulation: Principles of a science of design. Simulation & Gaming, 34(4), 569—591.
Klabbers, J. H. (2006). A framework for artifact assessment and theory testing. Simulation & Gaming, 37(2), 155—173.
Kolb, D. A. (1984). Experiential Learning: Experience as the Source Of Learning and Development. Englewood Cliffs, NJ: Prentice-Hall.
Kriz, W. C. (2003). Creating effective learning environments and learning organizations through gaming simulation design. Simulation & Gaming, 34(4), 495—511.
Kriz, W. C., & Hense, J. U. (2006). Theory-oriented evaluation for the design of and research in gaming and simulation. Simulation & Gaming, 37(2), 268—283.
Kriz, W. C. (2010). A systemic-constructivist approach to the facilitation and debriefing of simulations and games. Simulation & Gaming, 41(5), 663—680.
Lederman, L.C., & Stewart, L.P., (1986). Instruction manual for THE MARBLE COMPANY: a simulation board game. New Brunswick, NJ: Rutgers University.
Lederman, L. C. (1992). Debriefing: Toward a systematic assessment of theory and practice. Simulation & Gaming, 23(2), 145—160.
Markard, J. (2011). Transformation of infrastructures: sector characteristics and implications for fundamental change. Journal of Infrastructure Systems, 17(3), 107—117.
Mayer, I. S. (2009). The gaming of policy and the politics of gaming: A review. Simulation & Gaming, 40(6), 825—862.
Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 30(10), 955.

Morton, R. B., & Williams, K. C. (2010). Experimental Political Science and the Study of Causality: From Nature to the Lab. New York: Cambridge University Press.
Pellegrino, J.W. (2009). Mental models and mental tests. In: Wainer, H., Braun, H.I. (Eds.), Test validity (pp. 49—60). New York: Routledge.
Peters, V. A., & Vissers, G. A. (2004). A simple classification model for debriefing simulation games. Simulation & Gaming, 35(1), 70—84.
Peters, V., Vissers, G., & Heijne, G. (1998). The validity of games. Simulation & Gaming, 29(1), 20—30.
Raser, J. C. (1969). Simulations and Society: An Exploration of Scientific Gaming. Boston: Allyn & Bacon.
Sims, R. R. (2002). Debriefing experiential learning exercises in ethics education. Teaching Business Ethics, 6(2), 179—197.
Thiagarajan, S. (1993). How to maximize transfer from simulation games through systematic debriefing. In Percival, F., Lodge, S. & Saunders, D. (Eds.), The Simulation and Gaming Year- book 1993 (pp. 45—52). London: Kogan Page.
Van Ments, M. (1983). The Effective Use of Role-Play: A Handbook for Teachers and Trainers. London: NP Cogan Page.
Van Der Meij, H., Leemkuil, H., & Li, J. L. (2013). Does individual or collaborative self-debriefing better enhance learning from games? Computers in Human Behavior, 29(6), 2471—2479.
Wenzler, I., & Chartier, D. (1999). Why do we bother with games and simulations: an organizational learning perspective. Simulation & Gaming, 30(3), 375—384.
Zechmeister, J. S., Zechmeister, E. B., & J. J. Shaughnessy (2001).Essentials of Research Methods in Psychology. New York, NY: McGraw-Hill.

Authors / Contact

Jop van den Hoogen
Delft University of Technology
Delft, The Netherlands
j.vandenhoogen@tudelft.nl

Julia Lo
Delft University of Technology
Delft, The Netherlands
j.c.lo@tudelft.nl

Sebastiaan Meijer
KTH Royal Institute of Technology, Delft University of Technology
Stockholm, Sweden
smeijer@kth.se