Proceedings of the 2013 Winter Simulation Conference
R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, eds

USING GAMING SIMULATION EXPERIMENTS TO TEST RAILWAY INNOVATIONS: IMPLICATIONS FOR VALIDITY

Julia Lo
Jop van den Hoogen

Faculty of Technology, Policy and Management
Delft University of Technology
Delft, 2628BX, THE NETHERLANDS

Sebastiaan Meijer

Division of Traffic and Logistics
KTH Royal Institute of Technology
Stockholm, 100-44, SWEDEN

ABSTRACT

Gaming simulation in the railway sector often uses the same conceptual model as in computer simulation, and enables operators to interact with this model during a simulation run. Therefore, gaming simulation validation poses different challenges. This paper aims to answer the question to what extent gaming simulation can be used as an experimental research setting, due to its loosely demarcated experimental features. Focusing on validity issues, we study five cases in which the Dutch railway sector used gaming simulation to test innovations in a controlled environment. The results show that in addition to traditional external validity issues, human game players inherently open up this controlled environment, bringing in many confounding variables. By signaling what the specific validity threats are, this paper strives to improve gaming simulation for testing innovations that tackle social and technical elements of a system.

1 INTRODUCTION

Innovation in the railway sector is increasingly focused on achieving improvements by altering the internal architecture of the system rather than by expanding the system. In the Netherlands, recent costly projects like the Hanzelijn-extension and the High Speed Rail between Amsterdam and the Belgian border have left little financial room for further improvements. Additionally, spatial constraints inhibit ProRail, the Dutch railway infrastructure manager, in focusing exclusively on improving the system through large civil engineering projects. Thus, upcoming innovations tend to focus on altering the way the railway system is built up. Examples are decoupling railway lines and improving traffic control procedures.

These specific kind of innovations put more demands on the extent the involved decision makers understand the system and the internal causal mechanisms. However, in looking for regularities, human beings tend to favor linear processes and neglect feedback loops (Brehmer 1980). Since we assume railway systems to be complex systems, a collection of parts that interact in non-simple ways (Simon 1962), decision makers need research tools that allow the system to be studied holistically doing justice to both micro-level mechanisms and emergent properties.

Within the railway sector we see an abundance of the use of computer simulation, mostly discrete-event simulation, to assess innovations on their effect on punctuality, time table robustness and capacity. However, as innovations are more and more a case of fine-tuning both technical and human elements, a need arose to incorporate these human elements into the simulations. Traditionally, whenever computer simulations incorporated human behavior it was in the form of simplified algorithms. For instance, in the simulation software ProRail currently uses, traffic controllers are deemed to handle traffic around railway stations according to a simple first-come-first-serve principle.

Since 2009, the organization gradually employed gaming simulations to test out innovations in a controlled environment. Under the Railway Gaming Suite program, a joint project of ProRail and the Delft University of Technology, a plethora of gaming simulations have been designed and executed, for in-

978-1-4799-2076-1/13/$31.00 ©2013 IEEE

stance to test out traffic control concepts, handling schemes for larger disruptions, high-frequency time tables and new methods for freight transport slot allocation.

We define gaming simulation as an operating model of reality (Ryan 2000) to which gaming elements are added (Meijer 2012). Researchers can use gaming simulation for two purposes: hypothesis generation and hypothesis testing, although the latter purpose is less prominent (Meijer 2009). In the case of the gaming simulations employed at ProRail, we see that in five instances a clear hypothesis was tested using gaming simulation as an experimental method. When testing a solution in a game before implementing it in the reference system, validity becomes highly important (Peters, Vissers and Heijne 1998).

Our paper wishes to contribute to the body of methodological knowledge on computer simulation and gaming simulation and provide a theoretical and empirical contribution to the validation of gaming simulation. Therefore, the following research question is posed: *"how can we position gaming simulation as a hybrid between laboratory experiments and field experiments and what are the resulting validity threats?"*. Answering this question asks for a structured approach. We start by building a theoretical framework through which we can discover and assess validity threats in the use of gaming simulation. Additionally, we use five specific cases to look how in the design, facilitation and debriefing of gaming simulation experiments, we have tried to encounter these validity threats.

2 GAMING SIMULATIONS FOR INNOVATIONS IN COMPLEX SYSTEMS

As a design object, railway systems are a set of interrelated elements that together function to serve some chosen goal. We can therefore describe railway systems as a purposeful system (Ackoff 1971). Similar to Frenken (2006) we portray the design process of this railway system as a search process over a set of finite combinations of system elements. This set, typically called the design space (Frenken 2006), is the multiplication of all possible states of system elements. For instance, if we could conceptualize a railway system as a system having 10 elements (signaling, switches, trains, operational procedures, etc.) with each element having two states, the design space of this railway system would be $2^{10} = 1024$ combinations. Complexity in these systems involves the epistatic and pleiotropic properties of system elements. Respectively, these properties describe the extent to which the contribution of a system element to system level behavior is dependent on other elements and the extent to which a single element contributes to multiple system level behaviors. Because of these properties the fitness landscape of these systems tends to be rugged: multiple local optima of system configurations exist (Kauffman and Macready 1995).

2.1 The Use of Simulation in Innovation

Based on complexity models of innovation we find three requirements. Firstly, any system can be described using an unlimited amount of elements and states. However, cognitive limitations place restrictions on the amount of elements a decision maker can consider. Some form of abstraction makes the problem at hand more manageable. This abstraction however, still needs to consider the most dominant parameters. By experimenting with these dominant parameters, it is expected to find more effects on system behavior than considering less dominant parameters. Secondly, a designer wishes to know the internal rules an element applies in relating to other elements. Technical elements are relatively easily understood, e.g. wear-and-tear of railway tracks in relation to intensity of use can be described using a mathematical function. However, as human operators are an important part of a complex sociotechnical system, understanding how they relate to other elements is crucial in any process that wishes to optimize a system that is partly technical and partly social. Thirdly, in complex systems the epistatic properties of system elements limit the extent to which designers can work using simple rules-of-thumb about causal links between states of elements and system behavior. Because of these interaction effects, relations between states of elements and system behavior are highly non-linear, and evaluating a proposed design change can only be done by a holistic comparison of the current system and the system with the design change.

2.2 Building a Gaming Simulation

Comparing a system with and without an innovation resembles much an experiment, in which a treatment group is evaluated before and after the treatment. If a gaming simulation is employed for this purpose, it deviates from more common applications such as training and consensus building. Moreover, it is desirable for participants to portray similar cognitive processes and behavior as they do in their real work environment. Although gaming simulations differ in forms and purposes, still a set of fundamental design characteristics can be distinguished. In figure 1, a meta-framework that includes gaming simulations for research, training and policy purposes is specified to analytical science and design science (slightly adapted from Meijer (2009)). Analytical science refers to the research purpose of games, which mainly focus on hypothesis testing (Klabbers 2006). In this approach, gaming simulations are used as a research environment instead of traditional laboratory settings. Games for training and policy purposes reside under the noun of design science, which focuses on a change of participant(s) or an organization, based on experiences in the game session.

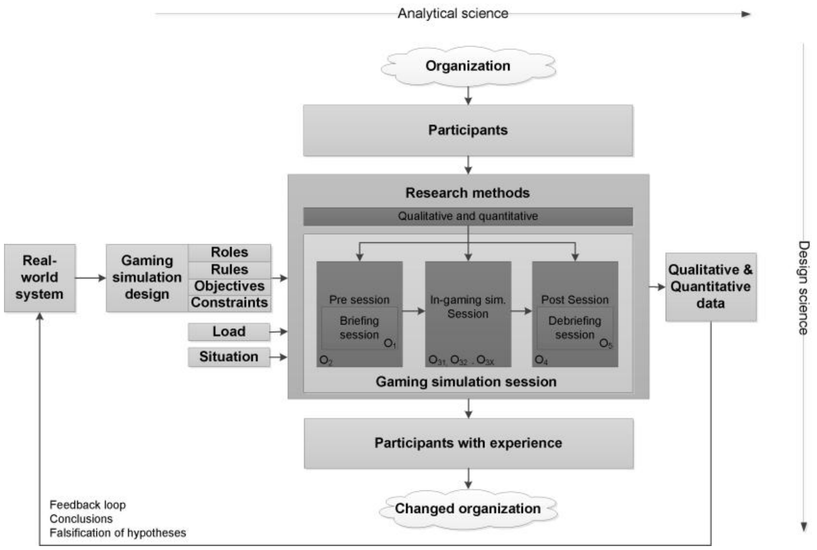

Figure 1: Meta-framework for design and analytical science with research and evaluation procedures included (slightly adapted from Meijer 2009).

Following the framework in figure 1, components within the real world provide input for the gaming simulation design aspects. These design aspects are related to the roles, rules, objectives and constraints of the gaming simulation with parameter settings, such as load and external influencing factors. Roles within gaming simulations can exactly match the roles of participants in the real-world environment or rather abstract representations (i.e. a fantasy role). Rules refer to behavioral limitations in the reference system or artificial constructs in what is allowed or forbidden within the simulated system. The nature of objectives need to be determined, to include individual and/or team goal(s) that are (implicitly) present in the reference system. Through the constraints, the range of actions that participants can take are limited within the gaming simulation. Additionally, the value of the variables in the design of the gaming simulation (load) and external factors that are in present in a gaming simulation, set the parameters of the gaming simulation. The abstraction level with regards to physical elements, which can be operationalized by the level of similarity and accuracy and use of isomorphism with the reference system, is determined by the choice of the scientific approaches.

Additionally, these design choices are also influenced by the emphasis of the validity of the gaming simulations, which differs between design and analytical purpose. Four types of validity have been identified for gaming simulations that are used for research, policy and educational purposes (Peters et al. 1998). Psychological reality refers to the perceived realism of the gaming simulation environment (i.e. simulated system). Structural and process validity refer to the degree of isomorphism in the simulated system with regards to the underlying structure and resulting processes in the referent system. Lastly, predictive validity denotes the degree to which the outcomes of gaming simulation correspond to historical or future outcomes in the reference system. It is expected that gaming simulations that serve the purpose for research require high validity levels on all four validity types, followed by educational gaming simulations which have a lower priority on predictive validity, and policy gaming simulations that only need medium levels of validity.

The next block in the framework describes the gaming simulation session, with a particular focus on the qualitative and quantitative data that is acquired to feedback the participants for training and policy (design science) games or to collect data for hypothesis testing in research (analytical science) games. In case of an analytical science approach, the research design and methods need to be carefully aligned and integrated. The gaming simulation session is usually consisting of a pre-session that can be separated in a briefing session, in which one or more participants are briefed on the session, and a window for measurement before the start of the session. During the session usually more qualitative and observational methods are used, followed by a possible measurement directly after the end of the session, and a final debriefing in which the participant(s) reflect about their experiences in the game session.

Gaming simulations exist in different forms, e.g. from high-tech individual human-in-the-loop simulator alike environments to low-tech multi-actor gaming simulations. The latter uses isomorphic elements, in which the information systems are made more abstract e.g. trains are represented by sponges or pegs. Train traffic operators take part in the gaming simulation in their own professional role. All necessary information is provided for the operators to make similar decisions as in their real work environment. Section 4 provides an elaborate description of the different low-tech gaming simulation.

2.3 Three-leveled Challenges

As mentioned earlier, different methods can be applied to test innovations in complex systems. Computer simulation as well as gaming simulation are both methods of simulating a reference system, each with their own properties and related strengths and weaknesses. Different purposes guide the development of both types of simulations. For computer simulations that are used for research, it is necessary to look into the process of simulation and conducting the research, which include the development of the model, the data analysis and the feedback of the results to others (Axelrod 2003). However, this is also the case for gaming simulations. In essence, gaming simulations experiments (or direct experiments) follow more or less the same research process as computer simulations (or thought experiments) (Axelrod ibid, Sterman 1987). In figure 2, the research process of both types of simulations is presented, which focuses on three levels: 1. to model or create a simulated system that represents the reference system, 2. to select valid simulation strategies or facilitate natural behavior by participants whilst controlling the research environments for confounding factors, and 3. to identify and obtain valid and accurate outcomes of the system that need to be translated to clients or researchers. This is in line with the process where a problem entity is translated into a computerized model through a conceptual model (Sargent 2004).

As the focus in this paper is on the use of gaming simulations in an experimental setting, a more in-depth description follows, to take upon a structured approach to identify characteristics in the research process. In figure 2, the three levels are accompanied by a set of validity challenges that have certain assurance for the following level. In order to have a valid simulated system, the *external validity* (the degree to which the findings can be generalized (Campbell and Stanley 1966)) needs to be assured. To confidentially make causal claims from the collected data (also defined as *internal validity* (Zechmeister, Zechmeister and Shaughnessy 2001)), the session needs to controlled for internal validity threats. Finally to draw conclusions based on the used research methods, these research methods need to be assured of a high *test validity*.

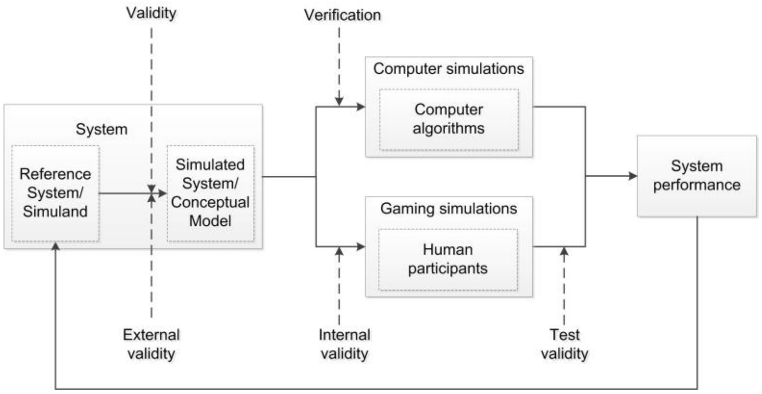

Figure 2: Three-leveled challenges in the research process of computer and gaming simulation environments.

In comparing computer simulation with gaming simulation, especially internal validity and test validity issues become significant. As a closed system, and thus lacking the problems of confounding factors, computer simulation does not have internal validity issues. Even in non-deterministic simulations, Monte Carlo methods help in averaging out the influence of an independent variable and a dependent variable and showing if this influence is statistically significant. However, internal validity-like issues appear during the computer programming of a conceptual model into a computerized model (Sargent 2004). In computer simulation literature the mitigation of this validity threat is done using verification activities. In gaming simulation sessions, the introduction of game players makes the experiment inherently open, allowing all sorts of confounding variables to distort the causal picture of one independent variable and one dependent variable. Furthermore as more soft variables are used to assess system behavior, e.g. work load and resilience, which do not need to be fully operationalized, gaming simulation, more than computer simulation runs the risk of not measuring exactly that what was intended to be measured.

The following section will describe what the characteristics of gaming simulations experiments are based on a comparison with more established types of experimental settings. Further on, external, internal and test validity are more thoroughly discussed.

3 EXPERIMENTAL SETTINGS

Laboratory and field experiments are two mainstream research settings in social science. The current section looks into the position of multi-actor board gaming simulations between these types of research environments, their components and the congruence on inherent and conflicting characteristics.

3.1 The Research Design of Experiments

The main objective in experiments is to manipulate on one or more factors (independent variables) and measure its effects on the manipulated variable (dependent variable) with a strong reliance on quantitative statistical methods (Zechmeister et al. 2001). The difference between experimental designs is related to the approach for which the sample procedure is conducted, whether a control group has been applied, and when and which measures have been used (see also table 1) (Creswell 2003). Experimental designs are also known as a configuration of set of research design characteristics, e.g. a one-shot case study is a form of a pre-experimental design, which includes no random sampling, no control group and solely a posttest.

Traditional experimental research usually takes place in an laboratory setting, which is characterized by low contextual cues. Field experiments on the contrary are a type of experimental setting that pertain high contextual cues, in which often a representative sample of situations and participants are involved. (Harrison and List 2004; Vissers, Heyne, Peters and Geurts, 2001).

Table 1: Research Design Characteristics for Three Types of Experimental Designs.

Research design characteristic	**Types of experimental designs**		
	Pre	**Quasi**	**True**
Sample procedure	Non-random, e.g. convenience sampling	Non-random, e.g. convenience sampling	Random
Conditions	No control group	No control group/control group	Control group (and multiple group conditions)
Measures	Pretest and/or posttest	(Multiple) pretest(s) and post-test(s)	Pre and/or posttest

Harrison and List (ibid, p.1012) describe the difference between laboratory and field experiments by the following characteristics:

- Nature of the subject pool: the degree of a nonstandard, representative sample, e.g. professionals
- Nature of the information that the subject brings to the task: the field knowledge and expertise that the participants bring to the experiment
- Nature of the commodity: the presence of physical field characteristics in the experiment
- Nature of the task: the domain-specific tasks in the experiment
- Nature of the stakes: the urgencies of risks in field settings
- Nature of the environment that the subject operators in: the environment of the experiment

Based on these characteristics two more hybrid lab-field experimental settings can be identified. Artefactual field experiments relate closer to laboratory settings, to the extent that an abstract frame and imposed set of rules is used in combination with a higher degree of a representative sample of the researched population. Framed field experiments build on the characteristics of artefactual field experiments, but additionally entail the field context as well with regards to the commodity, task or information.

Gaming simulation resembles mostly the latter type of field experiment, but distincts itself by the use of game design components, which are the presence of facilitators, the use of game design principles and components, such as immersion and play and the emphasis on the value of the debriefing session.

All in all, laboratory and field experiments make a trade-off between internal and external validity by respectively guaranteeing that the treatment variable is the only variable impacting the experiment and by guaranteeing that the experiment provides enough contextual cues for the experimental results to also hold in real life. Since gaming simulation somewhat hovers between these two ends of a continuum, validity threats come from both sides. In addition, researchers use test methods like observations, surveys and interviews to see how the dependent variable reacts to the treatment. Thus, the external, internal and test validity need to be secured.

3.2 External Validity

External validity is defined as "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). Issues or threats that can occur for external validity are (Campbell and Stanley 1966, p.6):

- Reactive effect: the effect of the pretest on the participants' sensitivity or responsiveness to the experimental variable
- Interaction effects: the interaction effects of biases in the selection of participants and the experimental variable
- Reactive effects of experimental arrangements: effects of the experimental variable upon participants being exposed to it in non-experimental settings. These include behavioral reactions of participants to the knowledge of being observed (e.g. Hawthorne effect) and the interactions between participants (contamination). When one of these validity threats occur in either one of the groups, but not in both, this becomes an issue for internal validity
- Multiple-treatment interference: effects of prior treatments remain present, thus possibly interacting with the new intervention

3.3 Internal Validity

Internal validity is defined as the ability to confidentially "state that the independent variable caused differences between groups on the dependent variable" (Zechmeister et al 2001, p. 149). In order to make a causal inference, the experiment needs to establish a relationship between the independent and dependent variable, the cause must precedes the effect, and finally, plausible alternative explanations should be outruled. To ensure the latter, the following factors (confounding factors or internal validity threats) need to be controlled (Campbell and Stanley 1966, Zechmeister et al. ibid):

- History: specific events that might occur between the first and second measurement next to the experimental variable
- Maturation: natural changes of participants over time, e.g. tiredness
- Testing: the effects of taking a test on subsequent testing
- Instrumentation: changes in the measurement of participants, due to the calibration of a measuring instrument or changes in the observers
- Regression: changes in the performance of participants that are due to the selection of participants on the basis of their extreme scores
- Subject mortality: loss of respondents in the different groups
- Selection: difference in individuals between the groups at the start of the study
- Interaction with selection (or selection-maturation interaction): different response of one group of participants to other internal validity threats, such as history, instrumentation

3.4 Test Validity

Finally, an experiment needs research methods to extract the information about causality from the experimental run. In a computer simulation, the information is mainly about primary qualities, such as speed, travel time or punctuality. In gaming simulation often dependent variables, or constructs, come in the form of more secondary or subjective qualities such as work load, operator reasoning or quality of the handling of disruptions. This adds to the importance of measuring exactly what was intended to be measured. This test validity refers to the validity of measurement instruments, in which the following three types are in line with the American Psychological Association (Van den Brink and Mellenbergh 1998):

- Construct: the extent to which the instrument measures what it is supposed to measure
- Content: the extent to which the test can be reflected to a spectrum of situations or topics
- Criterion: the extent to which the test correlates to one or more external variables, which are a direct measure for the variable.

4 MULTI-ACTOR BOARD GAMING SIMULATION EXPERIMENTS IN THE RAILWAY SECTOR

Since its conception, the Railway Gaming Suite has delivered a range of gaming simulations, from single player high-tech games to low-tech multiplayer board games. Focusing on multiplayer games, we see that most of the gaming simulations focus on traffic control concepts and more specifically on how traffic controllers and the higher echelons can best tackle disruptions. In the ETMET (lit.: a train each ten minutes) and the Bijlmer Junction games, the goal was to find out how different ways of handling a disruption would work out under conditions of higher frequencies. For the NAU game (lit.: new action plan Utrecht) the challenge was to find how robustness of the network was influenced by arranging traffic control along corridors rather than geographical areas and by removing railway switches. To test if it was possible to park already cleaned trains on Amsterdam Central station, the Platform Overnight Parking (POP) game focused on the Watergraafsmeer and Hoofddorp rail yard. Finally, the 1st phase game focused on the influence of a new way of managing a disruption on the speed and quality by which this disruption would be solved in real life. For a more thorough description of the games we refer to Meijer (2012). An overview of the projects and the design of the experiments can be found in table 2 and an overview of the validity threats can be found in table 3.

Table 2: Summary of Five Gaming Simulation Projects.

	ETMET	**NAU**	**Bijlmer Junction**	**POP**	**1st phase**
Goal	Test a different traffic control concept for handling disruptions under a metro-like time table	Test a different traffic control concept to mitigate second-order delays around Utrecht	Test a traffic control concept based on time slots rather than fixed time points under a metro-like timetable	Test the possibility of parking serviced trains on Amsterdam central station	Test a traffic control concept, especially in the first phase of a disruption.
Research design					
Sample procedure	Convenience	Convenience	Convenience	Convenience	Convenience
Conditions	Traffic control concept, Different time table	Traffic control responsibilities, Different infra lay-out	Traffic control concept, Different time table	Different maintenance and cleaning schedule	Traffic control concept
Measures	Pre-test and post-test	Pre-test and post-test	Post-test only	Post-test only	Pre-test and post-test
Control group	No	No	No	No	No

4.1 Research Design

In earlier studies on the design of railway systems within the organization of ProRail we found that decision makers were severely limited in their search space, i.e. the range of elements that they could manipulate and study (Van den Hoogen and Meijer 2012). In this study, we see that in three instances gaming simulation allows designers to increase their search space by incorporating multiple conditions to the design. However, in all cases the amount of treatment variables remained one. For instance, when two conditions were used, one condition always remained unchanged for both the pre-test and the post-test.

Gaming simulation uses real-life operators as behavioral input for a simulation run. This advantage also poses a disadvantage as finding available operators has proven to be cumbersome in multiple instances. Railway traffic control is a 24/7 operation and operational staff is scheduled accordingly. A fully random sampling procedure was impossible since operator availability was the decisive factor determining the sample. Furthermore we have learned through the course of executing gaming simulations that more experienced operators are more suitable than less experienced ones. Firstly, they are better equipped for new and complex problems, such as dealing with disruptions in general and under conditions of new innovations specifically. Secondly, we have noticed that using a certain level of abstraction increases the need for game players to translate this abstraction. More experienced players seem better able to do so.

4.2 External Validity Threats

External validity issues (table 3) appear when a design needs to be tested in a simulated experimental environment. Thus, building this needs to incorporate and preferably tackle these issues. We see some profound issues here that need further explaining. Firstly, models are inherently more abstract than the reference system. There seems to be a negative parabolic relation between abstraction level and ecological validity. The Bijlmer game used little abstraction, but was deemed less realistic by traffic controllers due to slight changes in the interface. Other games were more abstract, e.g. using sponges for trains instead of the standard traffic control interfaces. These models were less confusing to the game players and we saw in the debriefing that psychological reality was still perceived as high. Secondly, experimental arrangements might threaten the external validity. As far as we can see, two factors are most important here. Firstly, the benefit of gaming simulation is that processes that are normally spatially and temporally dispersed are now brought together. For designers, managers and decision-makers this allows them to study this processes in more detail. As they are observers that bring more scrutiny to the behavior of traffic controllers, strong Hawthorne effects might take place. Although inconclusive to this respect we see that each game caused high levels of immersion of the game players and we feel that this somehow decreases potential Hawthorne effects. Furthermore, both a highly observed game and a less observed game have both been validated in real-life and for effects of increased scrutiny on external validity we saw no indication.

Table 3: Validity Threats in Five Gaming Simulation Sessions.

External validity	**ETMET**	**NAU**	**Bijlmer Junction**	**POP**	**1st phase**
Ecological validity	More abstract but still all relevant information presented to players	More abstract but still all relevant information presented to players	High detail, small errors in context cues caused problems for immersion	More abstract but still all relevant information presented to players	High-tech-low-tech-hybrid
Immersion	High	High	Medium	Medium - High	Low - Medium
Reactive effect	N.a.	N.a.	N.a.	N.a.	N.a.
Interaction effects	Unknown	Unknown	Unknown	Unknown	Unknown
Reactive effects of experimental arrangements	Many observers; game players separated	Many observers; game players in one room	Many observers; game players separated	Low amount of observers; game players in one room	Many managerial observers; game players separated
Multiple-treatment interference	N.a.	N.a.	N.a.	N.a.	N.a.
Internal validity					
History	N.a.	N.a.	N.a.	N.a.	Learning effect for facilitators
Maturation	N.a.	During post-test some traffic controllers became tired	N.a.	N.a.	During post-test some traffic controllers became tired
Testing	Medium learning effect for traffic controllers	Medium learning effect for traffic controllers	High learning effect for traffic controllers	High learning effect for cleaning personnel	Intensive discussion about game and scenario between pre- and post-test
Instrumentation	N.a.	N.a.	N.a.	N.a.	Some observers were replaced during the experiment
Regression	N.a.	N.a.	N.a.	N.a.	N.a.
Subject mortality	N.a.	N.a.	N.a.	N.a.	N.a.
Selection	Respondents more experienced than their average real-life counterparts	Respondents more experienced than their average real-life counterparts	Respondents more experienced than their average real-life counterparts	Respondents more experienced than their average real-life counterparts	Respondents more experienced than their average real-life counterparts
Interaction with selection	N.a.	N.a.	N.a.	N.a.	N.a.
Test validity					
Construct	Resilience, Robustness	Resilience, Robustness	Robustness	Throughput capacity	Resilience, work load
Content	Punctuality and capacity as proxy, measured on train level	Punctuality and capacity as proxy, measured on train level	Punctuality and capacity as proxy, measured on train level	Amount of trains	Punctuality and capacity as proxy, Work load measured using self-rating
Criterion	Video, quantitative data, observers, debriefing	Quantitative data, observers, debriefing	Quantitative data, observers, debriefing	Quantitative data, debriefing	Video, questionnaires, observers, debriefing:

4.3 Internal Validity Threats

Internal validity issues appear when other variables within the experiment might explain the change in the dependent variable as well. Since all gaming simulations did not use a control group, it is hard to control for internal validity issues. A critical examination of possible confounding variables in the five cases showed that learning effects of players and facilitators, player fatigue and dynamic instrumentation are the main factors decreasing the internal validity.

When using less experienced operators we see that high learning effects take place during a simulation run, making it difficult to compare a pre-test with a post-test. For instance, during the overnight parking game cleaning personnel had difficulties in dealing with the abstraction and game mechanics during the first parts of the gaming simulation. Additionally, during the 1st phase game we saw that game facilitators, including one of the authors of this paper, had problems in facilitating the game and became more apt only as the game evolved. Although in this case only a minor problem, it points to the importance of training facilitators in the task they are responsible for during the session. If neglected, the learning effect of a facilitator might be mistaken for a treatment effect. Finally, we have noticed how gaming simulation sessions are demanding sessions that drain the energy of game players. To still be able to realistically compare a pre-test and a post-test, experimenters should incorporate fatigue effects.

4.4 Test Validity Threats

Construct validity is hampered by a problem of focus and a lack of a clear definition of often used concepts like resilience and robustness. Furthermore, we saw in the NAU game that disagreements occur on what the focal construct should be. While ProRail was interested in system performance, the Dutch railways was more interested in what the effect of the innovation was on the work load of their train controllers. Related to content validity, we see a very narrow focus on resilience and robustness as the extent to which punctuality and capacity can be maintained throughout a disruption and that these proxies were measured on a train level and not on a traffic level. However, this specific focus is also present in the reference system. A main and unique advantage of gaming simulation is that it easily allows for triangulation of data sources and thus increases criterion validity. In almost all instances we see that logs of punctuality are combined with observer logs, video reports and data from the debriefing to see if these data sources corroborate each other's findings.

4.5 Input for Next Cycle

In three cases the hypothesis was rejected (see table 4), much to the surprise of the involved project managers. However, they saw the gaming simulation session as externally valid enough to trust the outcomes and included the findings in the continued work on their proposed solution. In addition, the gaming simulation gave much valuable and rich information about what measures where needed parallel to their solution. These measures could stem from the simulation sessions itself (endogenous) or could be signaled by game players during the run or after the session (exogenous). For instance, using time slots in controlling high frequency traffic did not work quite as expected, but game players signaled additional directions for improvement, e.g. by changing platform lengths and building a railway track dedicated for overhauling.

Table 4: Results from Gaming Simulation Sessions.

Input for next cycle	**ETMET**	**NAU**	**Bijlmer Junction**	**POP**	**1st phase**
Hypothesis	Rejected	Accepted	Rejected	Accepted	Rejected
Additional data	-	Validation; additional endogenous dominant parameters found: cooperation between traffic control echelons	Additional exogenous and endogenous dominant parameters found: infrastructure and procedural changes	Validation	Additional endogenous dominant parameters found

NAU serves as a prime example of a gaming simulation of which the findings could be to some extent validated in real life. Some months after the session, this new way of handling traffic around the central node of the Dutch network was indeed altered. Different from the game, the switches were kept and their nonuse could only be guaranteed by work arrangements. It appeared that the same behavior was seen in real life as in the game: stability of single corridors, e.g. Amsterdam - Den Bosch, was sacrificed for the robustness of the total network. However, the fact that the railway switches could still be used, mostly in situations where flexibility was demanded by traffic controllers, meant that the system had a natural tendency towards a less robust but more flexible way of controlling traffic. In 2015, the measures tested in the NAU game are to be made more permanent by changing the whole infrastructure around Utrecht station, decreasing the amount of switches five- to tenfold.

5 DISCUSSION AND CONCLUSION

The current paper describes the position of gaming simulation as a direct experiment and as a form of a hybrid lab-field experimental setting, in which experimental research is conducted for railway innovations. Three levels in the research process are accompanied by three validity challenges, i.e. external validity threats in the development of the simulated system in the gaming simulation, internal validity threats to ensure causal relations that are drawn from data in the gaming simulation session, and test validity issues related to the selected research methods. An analysis on five cases is conducted, which showed issues with the level of internal validity, but a relatively high external validity and test validity.

A possible reason for this manifestation of internal validity issues might be due to the occurred learning effects of respondents, learning effects of facilitators or other effects of the experimental arrangements for which the lack of a control group meant that controlling for these issues was infeasible.

Overall, the gaming simulations seem to have less external validity issues, although there seems to be a relation between the level of abstraction and ecological validity. Finally, the test validity in the gaming simulations seems to be very high, because of the use of multiple quantitative and qualitative research methods. Out of the five cases, two innovations have been implemented and positively validated, in which a preconceived hypothesis about the effect of a specific innovation was accepted. The latter conclusion means that gaming simulation is useful for hypothesis testing but that researchers need to tackle a range of validity threats inherent to testing under organizational constraints. Furthermore, in some cases we have seen that additional information is gathered about dominant parameters that were either endogenous or exogenous to the model being studied. Especially in sessions where internal validity threats were present, due to learning effects of respondents, discussion between and with participants led to valuable insights in dominant factors that are present in the innovation process, for instance concerning the improvement of coordination and communication in traffic control and the improvement of infrastructure and station layout. This notion was not touched upon by this paper but serves as a promising avenue for further research. On top of that, this paper only focused on validity and left usability, meaning the link experimental findings back to real world implementation unexplored. Finally, it can be studied what factors influence the extent to which data from a gaming simulation can be used to alter a design.

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. 1971. "Towards a System of Systems Concepts." *Management Science*, 17, 1: 661-671.

Axelrod, R. 2003. "Advancing the Art of Simulation in the Social Sciences." *Japanese Journal for Management Information Systems* 12, 3: 1-19.

Brehmer, B., 1980. "In one Word: Not from Experience." *Acta Psychologica*, 45,1-3: 223-241.

Campbell, D. T., and J. C. Stanley. 1966. *Experimental and Quasi-Experimental Designs For Research*. Skokie, Illinois: Rand McNally & Company.

Creswell, J. W. 2003. *Research Design: Qualitative, Quantitative and Mixed Methods Approaches*. Thousand Oaks, California: Sage Publications, Inc.

Frenken, K. 2006. *Innovation, Evolution and Complexity Theory*. Cheltenham: Edward Elgar.

Harrison, G.W., and J. A. List. 2004. "Field Experiments." *Journal of Economic Literature* 42,4: 1009–1055.

Kauffman S., and W. Macready. 1995. "Technological Evolution and Adaptive Organizations." *Complexity*, 1-2: 26-43.

Klabbers, J. H. G. 2006. "Guest Editorial. Artifact Assessment vs. Theory Testing," *Simulation & Gaming* 37, 2: 148-154.

Meijer, S. A. 2009. *The Organization of Transactions: Studying Supply Networks Using Gaming Simulation*. Wageningen: Academic Publishers.

Meijer, S.A. 2012. "The Power of the Sponges." CESUN 2012 Third International Engineering Systems Symposium, Delft, 18-20 June.

Peters, V., G. Vissers, and G. Heijne. 1998. "The Validity of Games." *Simulation & Gaming* 29,1: 20- 30.

Ryan, T. 2000 "The Role of Simulation Gaming in Policy-making." *Systems Research and Behavioral Science* 17,4: 359-364.

Sargent, R. G. 2004. “Validation and Verification of Simulation Models.” In *Proceedings of the 2004 Winter Simulation Conference,* Edited by R .G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, 17-28. Piscataway, New Jersey: Institute of Electrical and Electronics Engineering, Inc.

Simon, H. A. 1962 “The Architecture of Complexity” Proceedings of the American Philosophical Society, 106,6: 467-482.

Sterman, J. D. 1987. “Testing Behavioral Simulation Models By Direct Experiment.” *Management Science* 33,12: 1572-1592.

Van den Brink, W. P., and G. J. Mellenbergh. 1998. *Testleer en Testconstructie*. Amsterdam: Boom.

Van den Hoogen, J. and S. A. Meijer. 2012. “Deciding on Innovation at a Railway Network Operator: A Grounded Theory Approach.” CESUN 2012 Third International Engineering Systems Symposium. Delft, 18-20 June.

Vissers, G., G. Heyne, V. Peters, and J. Geurts. 2001. “The Validity of Laboratory Research in Social and Behavioral Science.” *Quality & Quantity* 35: 129-145.

Zechmeister, J. S., E. B. Zechmeister, and J. J. Shaughnessy. 2001. *Essentials of Research Methods in Psychology*. New York, NY: McGraw-Hill.

AUTHOR BIOGRAPHIES

JULIA LO is a PhD candidate in the Policy, Organization, Law and Gaming department at Delft University of Technology. Her research focuses on studying the (team) situation awareness of operators in the railway sector through the use of (gaming) simulation methods. Her email address is j.c.lo@tudelft.nl

JOP VAN DEN HOOGEN is a PhD candidate in the Policy, Organization, Law and Gaming department at Delft University of Technology. His research focuses on systemic innovation processes in railroad infrastructures. His email address is j.vandenhoogen@tudelft.nl

SEBASTIAAN MEIJER is associate professor Transport Systems at KTH Royal Institute of Technology, Department of Transport Science, and part-time assistant professor at Delft University of Technology, Faculty of Technology, Policy and Management. His email address is smeijer@kth.se