Proceedings of the 2014 Winter Simulation Conference
A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, eds.

DEBRIEFING IN GAMING SIMULATION FOR RESEARCH: OPENING THE BLACK BOX OF THE NON-TRIVIAL MACHINE TO ASSESS VALIDITY AND RELIABILITY

Jop van den Hoogen
Julia Lo

Faculty of Technology, Policy and Management
Delft University of Technology
Jaffalaan 5
Delft, 2628 BX, THE NETHERLANDS

Sebastiaan Meijer

Department of Transport Science
KTH Royal Institute of Technology
Teknikringen 10A
Stockholm, 10044, SWEDEN

ABSTRACT

Gaming simulation allows for experiments with sociotechnical systems and has as such been employed in the railway sector to study the effects of innovations on robustness and punctuality. Systems work as non-trivial machines and the effect of an innovation on a dependent variable is potentially context, time and history dependent. However, several constraints inhibit the use of validity increasing measures such as repeated runs and increasing sample size. Based on a debriefing framework, insights from qualitative process research and six games with Dutch and UK railway traffic operators, we provide a guide on how to assess and increase reliability and validity. The key is for game players, observers and facilitators to open up the black box and thereby assessing how the innovation brought about any changes, if these changes are insensitive to changes in parameters and if the conclusions hold outside the game.

1 INTRODUCTION

Gaming simulation can be defined as an operating model of reality (Ryan 2000) using gaming methods to allow human participants to interact with this operating model during a simulation run (Meijer 2009). The debriefing of participants is seen as crucial for gaming simulation sessions that serve the purpose of training and education, during which participants learn (Crookall 2010) and reflect on the experiences (Lederman 1992). Contrary to gaming simulation for experiential learning, the literature on the debriefing of gaming simulation for research is rather sparse. The structure and components of a debriefing of experimental gaming simulations (hypothesis testing) or exploratory gaming simulation (hypothesis generation) have not been specifically identified and addressed by many studies. However, the importance of a debriefing for games for research cannot be overstated. Usual techniques for assessing the validity and veracity of the model and the simulation outcomes, such as repeated runs, elaborate factorial designs and sensitivity analysis (Kleijnen 1998, Sargent 2005) are hard to employ because of constraints on resources such as game players. Therefore game players and observers, under the guidance of a facilitator, should discuss and potentially resolve these issues in a debriefing session. Following this observation, the current authors defined a framework to identify the phases of a debriefing in a gaming simulation used for research purposes that is applied in an organization (Van den Hoogen, Lo, and Meijer forthcoming). The framework includes and addresses different validity and reliability dimensions of the gaming simulation session itself, in which the debriefing is recognized as an opportunity to assess and improve these dimensions.

This framework serves however as best as a topic guide, providing the interested reader a structured approach to cover all relevant aspects of a debriefing. The current paper therefore wishes to operationalize this framework by exploring ways with which all these topics can be practically addressed. Using a

combination of our experiences as game designers and borrowing insights from methodological literature on qualitative process research methods, we build a more fine grained structure for the debriefing.

2 GAMING SIMULATION IN THE RAILWAY SECTOR

In 2009 ProRail, the Dutch railway infrastructure manager, and Delft University of Technology started with the Railway Gaming Suite (RGS), which intended to introduce gaming simulation as a decision support tool to the organization. Most of the games so far focused on testing a preconceived innovation in an experimental setup. Similar to medical research we thus applied a treatment (e.g. removing switches, adding new communication protocols or changing job roles) to a test subject (i.e. the railway system comprised of all relevant technical and social elements) and if possible, the effects on a dependent variable (e.g. capacity or punctuality) was measured without and with the treatment, similar to a pretest-posttest research design. In Table 1 we show six research games we employed for ProRail (five) and Network Rail in the United Kingdom (one). Meijer (2012) provides a more in-depth overview.

Table 1: Railway research gaming simulations.

Game	Project	Treatment	Data collection
BIJLMER	Starting a metro-like timetable	New traffic control concept	Logs of punctuality, debriefing with game players
ETMET	Starting a metro-like timetable	New disruption management concept	Logs of punctuality, debriefing with game players and stakeholders
NAU	Complete overhaul of infrastructure of Utrecht Station	Removal of switches, new traffic control concept	Logs of punctuality, debriefing of game players and observers
LEEDS-BRADFORD	Introduction of traffic management	New traffic control concept, new roles	Debriefing of game players
1st PHASE	New ways of handling disruptions by operators	New disruption management principle	Debriefing of game players and observers
OV-SAAL	$ 1bn. upgrade of Amsterdam Airport – Lelystad corridor	New infrastructural expansions	Debriefing with game players

2.1 The Impetus for Adding Gaming to Simulation

The organization wished to employ gaming simulation because they increasingly became aware that social elements and rules determined largely the behavior of the system, especially when the system is in a disrupted state. For normal train traffic conditions, the organization makes extensive use of computer simulation software to experiment with different time tables and track layouts. This is feasible because in the Netherlands traffic control is, within certain bounds, fully automated. The rules that are applied in this automation are easily transferred into algorithms for computer simulation. However, in case of large disruptions which the automatic track assignment program cannot solve, human traffic controllers step in. Their job is then for instance to reroute, cancel and combine trains; activities that in most computer simulation software cannot be sufficiently simulated.

In general our gaming simulations can be described as low-tech multi-player table top games, with some exceptions that employed elaborate computer simulations, with which game players, such as traffic controllers, interacted. Although mostly low tech (sponges for trains, paper table tops for infrastructure, etc.), our gaming simulation experiments always involve making a model that contain many real world elements such as realistic time tables, realistic infrastructure and real life operators. Thereby, operators

interact with the railway system similar to reality: they are able to see trains moving correctly over the tracks and switches according to a known timetable and they are able to make the same decisions, communicate and coordinate the same as in real life.

2.2 Gaming as an Experimental Research Tool

Given the purpose of gaming simulations for research, experimental and exploratory gaming simulation requires a high validity and reliability (Lo, Van den Hoogen, and Meijer 2013). Combining literature from experimental studies in computer simulation and psychological research, we identify and address two reliability concepts: sensitivity and measurement reliability (Van den Hoogen, Lo, and Meijer 2014). As validity and reliability are crucial concepts underlying the research question and the purpose of the research gaming simulation, the interpretation of the simulation outcomes should be addressed in the debriefing (Kriz 2010, Peters and Vissers 2004). However, the specific structure of how to conduct this phase of the debriefing remains open. Table 2 lists different dimensions of validity and reliability. In addition to the previous literature, we further distinguish external validity in generalizability (the extent to which the results are generalizable based on the representativeness of the sample in relation to the required population), and ecological validity (the extent to which the design of the simulated system sufficiently reflects the reference system).

Table 2: Validity and reliability concepts and dimensions, slightly adapted from Lo, Van den Hoogen, and Meijer (2013) and Van den Hoogen, Lo, and Meijer (2014).

Dimension	**Definition**	**Experimental design**	**Possible threats**
Measurement reliability	Similarity of the measurement outcomes over a number of runs	Multiple measurements, triangulation	Qualitative observations used, high observer dependence
Sensitivity	Variation of the session outcomes over a number of similar runs	Multiple runs, sensitivity analysis	Number of runs usually one, hard to assess sensitivity in the game
Internal validity	Causal claims are true inside the scope of the simulated environment	Research designs (pretest-posttest, control group, random treatment assignment)	Resources constrain use of pretests, possible confounding variables due to social processes and errors in game facilitation
Generali-zability	Extent to which findings hold beyond the sample and for the overall population of possible system configurations	Sample size, sampling procedure	Resources constrain use of large samples, other sample dimensions such as time tables are picked according to representativeness of population of system configurations
Ecological validity	Extent to which findings hold in real life, in an ecology of omitted elements and processes	Experimental context design or game design	Use of stylized contextual cues, other properties of reference system are omitted such as neighboring control centers and passenger flows

Designing, facilitating and debriefing the gaming simulations from 2009 onwards has provided us with many insights on the pros and cons of using gaming simulation as an experimental research tool for testing innovations. In Table 2 we have added insights from our work on potential validity threats (Lo, Van den Hoogen, and Meijer, 2013) that arise when employing gaming simulation in an organization.

3 FRAMEWORK FOR DEBRIEFING RESEARCH GAMING SIMULATIONS

Based on a literature review related to general guidelines on debriefing (e.g. Kriz 2010, Lederman 1992, Peters and Vissers 2004), we developed a debriefing framework for research gaming simulations which is showed in table 3 (Van den Hoogen, Lo, and Meijer 2014). This framework is based on several unique notions to gaming simulation as an applied research tool: firstly, we feel that for a proper debriefing a different mental state is required from the game players than during gameplay. Whereas immersion is crucial for externally valid player behavior, we need them to go to a more retrospective mental state if they are asked to reconstruct and evaluate the dynamics of the gameplay. Sharing experience and emotions should enable more ease in the reflection on the gaming simulation session in the first phase of the debriefing. Secondly, our simulations run real time and only a few runs can be finished within one day due to limited availability of organizational resources such as time, money and personnel. Since many simulation validation techniques rely on running a model many times with different test data, parameters and different factorial designs in the design of experiments (DOE) - see for instance Balci (2013), Sargent (2005) and Kleijnen (1998) for an overview of these techniques - real time gaming simulation methods are limited in the extent to which validity and sensitivity can be assessed in-game. Therefore in the second, third and fourth phase of the debriefing, these issues should be discussed. Thirdly, we apply gaming simulation in an organization. This invokes a number of challenges that influence the design of the game and the debriefing, such as the (amount of) employees that can be arranged to participate in the gaming simulation and the possibility to use subject matter experts (SMEs) as observers. Another challenge is that employees return to their real job after the gaming simulation is finished, in which they take lessons learned from the gaming simulation back into the organization. These insights may want to be recognized and reflected upon in the debriefing stage. As the gaming simulation takes place in an organization, phase 5 'planning for action' is aimed at converging the results into actions and identifying next steps. The final phase of the debriefing 'protect the instrument' focuses on an overall conclusion about the experience of the session, before participants return to different parts of the organization. Given the flexibility of the gaming simulation, the framework provides guidelines in ideal circumstances. That is, depending on the presence of certain participants, the format of the debriefing session might change, e.g. observers may not be available or support facilitators may not be needed.

Table 3: Framework of the phases, addressed topics and involved participants in a research game.

Phase	**Description**	**Topics**	**Involvement of participants**
1. Cooling down	Change mental state of game players from immersion to retrospection	Experience, emotions	Facilitator, game players
2. Data collection	Additional qualitative data from game players, observers and facilitators	Measurement reliability and validity	All participants
3. Reliability	Do repeated (or slightly different) runs result in (slightly) similar outcomes	Sensitivity	Game players, observers
4. 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
5. 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 and actions	All participants
6. Protect the instrument	Evaluate gaming simulation session; determine what outcomes may be shared; ensure durable relationship with players	Experience, emotions	Facilitator, game players

4 OPENING THE BLACK BOX

Since we use gaming simulation as an experimental tool, we intend to measure causal links between some predetermined set of independent variables, constituting the innovation, and a set of dependent variables, constituting the performance indicators assumed to change as a consequence of the innovation. However, different from usual experiments in for instance medical and psychological research, we do not apply our innovation to a assumed single atomistic entity, but rather to a complex system comprising many interdependent, adaptable, and dynamically interacting elements such as acting and reacting traffic controllers. The difference between trivial machines (TM) and non-trivial machines (NTM) by Von Foerster (1984) helps in explaining how this impacts the way we claim any causality.

4.1 Trivial and Non-trivial Machines

In so called 'text-book' style experiments, researchers adhere to the notion of the TM (Klabbers 2006), assuming that some conceptual device transforms an input x into an output y, irrespective of time, history and context. The device brings about a mechanistic and linear causal link between x and y and the world is thought to consist of a web of these mechanistic links. Because the transformation is understood as a simple one, opening the black box is not needed: for prediction we simply need to establish *that* x leads to y, not *how* x leads to y. Complexity perspectives have resulted in a different notion of these conceptual devices. NTM's are devices that transform x into y in a far more complex manner, caused by some internal processing scheme that brings about causality in a manner highly dependent on context, time, place and history. Therefore the internal structure and processes within this black box do matter (Von Foerster 1984) and cannot be neglected as is done for TM's. A perfect example of a NTM is a social system comprising of adaptive, acting and reacting human actors (Klabbers 2006). Since we simulate a system that is partly social we must acknowledge the possibility that our games are non trivial machines and that any causality found is potentially dependent on many uncontrollable factors, path dependence and emergence and distorted by possible chaotic properties of the dynamics of the system. This prevents us from just comparing a pretest (without the innovation) and a posttest (with the innovation) on some predetermined performance indicators, as is done in classical medical and psychological experiments.

4.2 Research on Processes and Narratives

The fact that causality is often brought about by a complex interplay of nonlinear feedback loops, path and context dependent chaotic processes on multiple levels of analysis, has since long been recognized in the historical and sociological sciences (Griffin 1993, Hedström and Bearman 2009) with a strong reliance on narrative explanation using sequences of events and timing and conjunctures of event-chains rather than a variable-based explanation using independent and dependent variables (Abbott 2001, Geels 2011). According to Weber (1949) most events are too complex to allow for causal generalizations. 'Narratives', rather than causal models, can provide insight into the dynamic interplay of agency and social structure, in and through time (Giddens 1979, Sewell 1992, Griffin 1993). Thus, phenomena in these fields of study are deemed too complex and too context dependent that isolation, as is common in other sciences, renders any scholarly undertaking invalid. In the realms of management and policy sciences, these processed-based or narrative approaches have influenced the works of Van De Ven and his Minnesota Innovation Research Program in the 80s (Van de Ven, Angle, and Poole 2000), scholars such as Pettigrew (1992), Langley (2007) and Tsoukas and Hatch (2001) and research on transitions of complex sociotechnical systems (Geels 2011).

4.3 Methodologies

Whereas the variable based sciences have an extensively developed repertoire of methodologies, process based sciences still lack a well developed methodology (Geels 2011). According to Abell (2001) this is related to the ontological primacy of narrative studies compared to the epistemological primacy of

variable centered studies. However, since what really happens in a gaming simulation is a sequence of events and not a link of causally related variables, we believe that we can seriously improve our debriefings if we borrow some of the methodological insights from the historical, sociological and qualitative management sciences. We see event-structure analysis (Heise 1988) as a much used methodology in the process literature. This methodology focuses on the sequencing of events, how accumulations of past actions constrain future actions and the introduction of contingent and unpredictable events that are able to capture novelty (Griffin 1993). For an overview of similar narrative analysis methodologies we refer to Manzo (2010).

The analysis starts by building a narrative in which events have a specific temporal ordering, much like a timeline of a story. To avoid just portraying a sequence of events, researchers need to know to what extent one event causally triggered a following event or, through temporal side branches, indirectly triggered an event later on. Here, much importance is placed on counterfactuals. The core idea is for every event to be analyzed as if it were just an instantiation of another possible event that is the negation or modification of that specific event, basically demanding from the research to ask for every event a 'what if-question' (Griffin 1993). The condition is however that the counterfactual world is a possible world and conceptually close to the real past, hence these counterfactuals are also called 'objective possibilities' (Weber 1949). For example, a game player can decide to continue the service of a delayed train and the counterfactual might be to cancel the service. If however cancellation would leave thousands of passengers stranded in a meadow, this counterfactual might be deemed impossible, i.e. not likely to happen in real life. If the hypothetical absence or modification of an event triggers a totally different unfolding of events, than this event can said to be essential and causally triggering all following events (Griffin 1993). Usually, researchers determine the effect of counterfactuals based on either what theoretically should be consequences or what generally, i.e. in other cases, are consequences. For instance, if we portray the decision of one operator to reroute a train from track A to track B as an event and we see other operators reacting to this, we wish to know whether an alternative decision could have also been made and to what extent this different decision would have led to a completely different game process. This can be determined either based on previous experiences of the consequences of this other decision or based on theoretically predicting the consequences. Notable for this approach is that it assumes that operators have the ability to rationally assess and reflect on their decisions contrary to theories on intuitive or unconscious cognitive processes. Thus, we assume them to be able to explicitly recall and reason on their decisions. In Table 4 we briefly summarize the steps of most narrative analyses that focus on causality.

Table 4: Event structure analysis (loosely based on Griffin 1993).

Step	**Description**
1. Determine events	Map all game player decisions, changes in parameters of the game and the context
2. Determine counterfactuals	Map for every event the potential counterfactual events
3. Assess realism of counterfactual	Determine whether the counterfactual is close to the real past and is realistic in real life.
4. Determine counterfactual world	Assess to what extent the different event would trigger different following events.

5 DEBRIEFING RAILWAY GAMING SIMULATIONS

At the debriefing different participants will be present, e.g. game players, observers, facilitators, managers and other stakeholders, often varying in numbers. We identify different roles for each participant during the debriefing, in which managers and other stakeholders should only – if they should be included at all - in a round-up of the debriefing, to ensure a safe environment where participants can discuss and share

their thoughts. Using our framework and combining this with insights from historical, sociological and process research, we show how each topic can be addressed in a debriefing and what each involved participant should do. In the end this provides a means to open up the black box of gaming simulation.

5.1 Cooling Down

There are several methods on how the cooling down in a debriefing session should look like (Kriz 2010, Peters and Vissers 2004). Although it might appear trivial, the first step in cooling down participant is allowing for a break between the end of the game and the start of the debriefing. Game players, by being occupied with something else then the game, are then more able to get out of the immersed mode that was demanded from them during the game. In all games, we see how fully immersed players are and how sometimes heated debates arise, as in real life, during the resolution of disruptions. Furthermore, we start the debriefing with a general question, such as 'how did you experience the game?' Additionally, it is important to note that research games that take place in an organization with participants from the organization itself may be affected by political or organizational sensitivities. Both organization culture and maturity of the organization toward the use of gaming simulation should be taken into account by the lead facilitator: he or she should start by looking what controversies were touched upon, such as new job roles, and should make clear to the participants of the debriefing how and to what extent these controversies may and can be discussed during the debriefing.

5.2 Data Collection

In the five years we have been conducting gaming simulation experiments we slowly realized the enormous potential of gaming simulation as a tool to observe a dynamic system holistically. Otherwise dispersed system elements, think of different operation centers located in different parts of the country, are now brought together and their interaction patterns immediately become visible. Therefore, in addition to more traditional data that is logged, such as capacity and punctuality, we started to use more and more observational data that allowed us to open up the black box. For this purpose we rely on retrospective accounts of game players as well a special observers. For instance, in the 1st PHASE game we invited several SMEs which we provided with a topic guide that focused their observations on certain processes during the game. However, as quantitative data is relatively reliable and self-explanatory, observations are highly observer dependent. The debriefing thus first needs to focus on building a coherent picture of what happened inside the black box. For this purpose we use the event-structure analysis methodology, albeit somewhat loosely, and as we will show later on, this analysis serves as a perfect tool to more accurately assess the validity and reliability of the outcomes as well.

We start by mapping all events that occurred during the game in a timeline. A better metaphor would be a stave, where each line represents an element within the system and each music notation represents an event instigated by that system element. The story is then the temporal progression of individual events, for instance traffic controller 1 decided A (event 1), upon which trains moved from B to C instead of from B to D (event 2), triggering a reaction by traffic controller 2 (event 3) and train controller 1 (event 4), and so on. A graphical representation in Figure 1 better explains this.

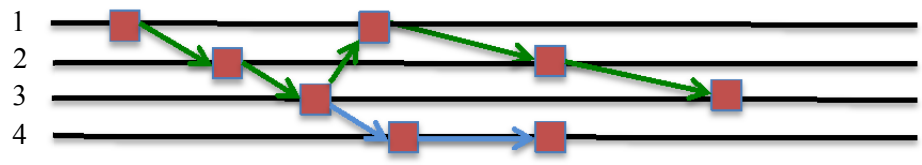

Figure 1: Event-chain of a system with four elements with actual sequence of events (green) and an objective possibility (blue) had actor 3 made a different decision.

We note here that the extreme detail we use here is for didactical purposes. In the games we debriefed so far the level of detail is much lower, focusing on around 10 to 15 events that describe the processes in the game. Basic debriefing questions that guide this process are: what did happen? What decisions did you make? Based on what events was this decision triggered? What processes did you observe? For these questions we always combine the insights of game players and observers since game players better know the relevance of events whereas observers are better to recall all events since they have been appointed to do so. In bringing together all these insights, the lead facilitator should then assess to what extent peoples' observations are reliable (would they observe the same in a rerun?) and valid (did they use the correct constructs in their qualitative measurements?).

5.3 Sensitivity

System behavior in games might be chaotic, or partly random and therefore sensitive to slight changes in initial parameters or highly contingent on specific decisions made by players during the game. However, usual ways of assessing these properties such as multiple runs are infeasible. In the debriefing we thus need to find out how sensitive the course of the game was to initial settings and game player decisions. We apply our event-chain to determine which decisions most significantly had an impact on the further course of the game. Usually game players are too much immersed to recall all critical decisions and we often add the insights of observers for this purpose. Using these critical decisions, a debriefing should focus on the extent to which these decisions might be tipping-points. So we ask during the debriefing whether these critical decisions were stemming from a range of other decisions and to what extent these decisions are likely (objective possibilities, see the blue arrow in figure 1). The likelihood of these other decisions is often a matter of practical experience and a knowledge of what is actually desired of the operational controllers, therefore again this part of the debriefing needs both game players and SMEs. Consequently we assess for every likely other decision the severity of its impact on the further course of the game. The same as in historical research, we ask game players what theoretically would follow from this other decision (where we rely on their own 'mental' simulations of the changed game) or what in general would follow. Next to that, a big advantage of gaming simulation is that the materials are still available and setting up a new simulation run would take a few minutes. In OV-SAAL, we discussed one traffic controllers' decision and not being sure what the consequences would be of an alternative decision, we started the game and let it run for 15 minutes to see how the game progressed under these new conditions.

5.4 Internal Validity

With the gaming simulation experiments we intend to determine whether an innovation and a performance indicator are causally linked. However, the specifics of our experimental design serious limit the internal validity if we were to leave the black box closed. In that case, we at best can establish a correlation between the two variables. Agreeing with George and Bennett (2005), if we establish *how* A leads to B, we are better able to establish *that* A leads to B. Focusing again on the events, we ask the game players how the innovation specifically influenced the instantiation of this event and not one other. By doing so we are to translate a narrative of event sequences to a narrative of causal variables, providing the causal story between variables A and B. Furthermore, by making specific events as the topic of discussion rather than the general game, game players and observers seem to better determine to what extent other internal validity threats played a role.

5.5 External Validity (Generalizability)

Since our unit of analysis is on the system level, the population for which the outcomes should hold as well as the sample that is part of the study must be defined as configurations of system elements. Usual elements we apply in defining the sample are infrastructure, operators, timetable and a scenario. For

instance, for LEEDS-BRADFORD we used a set of operators, an afternoon timetable and the whole infrastructure around this conurbation and a realistic disruption as a scenario. In the game we studied the effects of changing the job roles, giving operators new responsibilities on their effectiveness in coping with disruptions. Generalizing the results would involve asking to what extent the results would hold for the Leeds-Bradford area when a different timetable is applied, when different operators are working or when different disruptions occur.

As a start, in this part of the debriefing the facilitator should explore on what facets the sample (infrastructure, timetable, scenario, game players, etc.) differed from the population on which the experiment wished to shed light on. Usually, we have made some assumptions in the design process of the game, e.g. using representative disruptions in our scenarios, using peak hour timetables, and using local operators as game players. However, certain facets are always overlooked and the local knowledge of game players and observers might add to the comparison between sample and population. Using the dimension by which the sample and population can be compared, we again use the chain of events and ask the question: would this event occur when a different timetable was used? Would you make the same decision if the disruption was less severe? Would another traffic controller in the same situation make the same decision as you did in the game?

5.6 External Validity (Ecological Validity)

Raser (1969) conceptualized ecological validity according to three dimensions: structural and process validity and psychological realism. We have seen during the debriefing sessions that game players are able to determine realism through general questions and without focussing on specific events and decisions. Therefore we often tackle this point with a simple question such as: did it feel real? As soon as players state that certain parts of the game felt unrealistic, we assess what events or decisions were impacted by this unrealism and how this affected the game outcomes. To assess structural and process validity (the extent to which structure and process in the game resemble the structure and processes of the referent system) we rely firstly on the facilitator who has knowledge on the rationale of the game design and which structural and process elements were omitted. Often game players are able to signal as well elements they felt belonged to the game to make it more ecologically valid. Differences between the game and the referent system are then collectively assessed on their impact on key events.

5.7 Planning for Action

What we have seen in the games is that although in the first place the objective was to test hypotheses, many additional hypotheses were generated during the debriefing. For instance, during the debriefing of NAU, we collectively found that removing switches and changing traffic control procedures was indeed beneficial to robustness but that additionally better cooperation protocols were needed for higher echelons of traffic control. Often, the facilitator is appointed with setting out these additional topics for further research. Even more important than additional research is the question who is responsible for what action. In the debriefing we therefore often have the client of the game and other stakeholders as well, and in a round table setting we discuss what actions will be taken based on the game and by whom.

5.8 Protecting the Instrument

Finally, as participants return to their organization after the end of the gaming simulation, it is necessary to come to a fruitful conclusion, in which important and bothering issues have been addressed by participants. This phase of the debriefing is related to ‘protecting the instrument’ (Kriz 2010, Peters and Vissers 2004), which is not only necessary for a constructive wrap-up, but also to ensure the validity of game players’ behaviour in future gaming simulation sessions, e.g. negative attitudes that influence the immersion, resulting in a low psychological reality with its related cognitive and behavioural differences.

5.9 Overview

In table 5 we present a brief overview on how additional data can be collected in the debriefing and how the sensitivity and validity of the outcomes can be assessed. We have omitted 'cooling down', 'planning for action' and 'protecting the instrument' from this table since we assume them rather self-explanatory.

Table 5: Debriefing through the reflection of validity and reliability issues.

Dimension	**Role of participants in the debriefing**		
	Player (Operator)	**Observer (SME)**	**Facilitators**
Data Collection	Establishment of event-chains	Establishment of event-chains	Juxtaposing statements; assessing measurement validity and reliability
Sensitivity	Determine counterfactuals and their effects on subsequent events (based on experience)	Determine counterfactuals and their effects on subsequent events (based on theory, rules, etc.)	Ask players and observers about crucial events and objective possibilities
Internal Validity	Determine how treatment impacted the events; determine effect of confounding variables	Determine how treatment impacted the event-chain; determine effect of confounding variables	Identification of potential confounding variables due to experimental context
Generalizability	Comparison own decisions with probable decisions made by peers	Identification of differences between sample and the population	Linking differences found by observers with players' comparisons
Ecological Validity	Determine perceived realism and effect of omissions of elements and processes of referent system on event-chains	Determine effect of omissions of processes and structural properties of referent system on event-chains in game	Discuss what omissions were applied during game design

6 DISCUSSION AND CONCLUSION

In the five years we have been designing research games for the railway sector we found that the debriefing session serves as an opportunity to collect data, validate findings and come to general and actionable conclusions. We learned that for these purposes it was of vital importance to open up the black box. Rather than simply assuming *that* some innovation influenced a performance indicator, we used the debriefing to assess *how* this influence took shape. Rather than having game players and observers evaluating the game process holistically, we have learned that focusing on specific chains of events is better suited for our debriefing purposes. Firstly, our use of game players and observers as empirical input for analysis potentially raises measurement reliability and validity issues: do we measure the same constructs the same if we were to rerun the game? By explicating the game process through describing chains of events, we are better able to uncover these issues. Furthermore, through asking to what extent the in-game dynamics and outcomes are highly contingent on events (sensitivity), in what ways these events were affected by the innovation (internal validity), and if the chains of events would take place for all facets of the population (generalizability) and in real life (ecological validity), many of the validity threats can be assessed and alleviated. Thus, by collectively reconstructing the chains of events that happened in the game and envisioning other possible worlds by continuously asking 'what if'-questions, we are better able to draw up conclusions on causality and assess the validity of these conclusions.

One of the main assumptions we rely on in applying this focus on studying events and asking 'what if'-questions is that game players are able to think through how objective possibilities would play out over the course of the game. Thus we ask from them to mentally simulate the effects of a modification or

negation of an event. To what extent they are able to validly do so and the reach of this mental simulation however remains uncertain and deserves more scientific attention. Besides that, the framework deserves a more rigorous study on the possible challenges in its application over a broader range of topics, innovations and in different organizational and national settings.

Furthermore, participants may identify tipping points during the debriefing session, in which they also open up to (a part of) their reasoning process. We believe that studying these reasoning processes in-game and in-action might be more valid than usual studies of operator reasoning through surveys and interviews. By doing so, we are able to improve the design of agents as substitutes for algorithms in discrete event simulation. Difficulties in obtaining alternative decision options might indicate the involvement of unconscious, intuitive or implicit knowledge processes, which might show the need for specific modelling methods. Gaming simulation has shown to have its shortcomings, as does any other tool, and this way of coalescing the two might be a case of combining the best of both worlds: the realism of gaming and the rigor of simulation.

ACKNOWLEDGMENTS

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

REFERENCES

Abbott, A. 2001. *Time Matters: On Theory and Method*. Chicago, Illinois: University of Chicago Press.

Abell, P. 2004. "Narrative Explanation: an Alternative to Variable-centered Explanation?" *Annual Review of Sociology* 30: 287-310.

Balci, O. 2013. "Verification, Validation, and Testing of Models." In *Encyclopedia of Operations Research and Management Science*, edited by S. I. Gass and M. Fu, 1618-1627, Boston, Massachusetts: Springer.

Crookall, D. 2010. "Serious Games, Debriefing, and Simulation/Gaming as a Discipline." *Simulation & Gaming* 41(6): 898-920.

Geels, F. W. 2011. "The Multi-level Perspective on Sustainability Transitions: Responses to Seven Criticisms." *Environmental Innovation and Societal Transitions* 1(1): 24-40.

George, A. L., and A. Bennet. 2005. *Case Studies and Theory Development in the Social Sciences*. Cambridge, Massachusetts: MIT Press.

Giddens, A. 1979. *Central Problems in Social Theory: Action, Structure, and Contradiction in Social Analysis*. Berkeley, California: University of California Press.

Griffin, L. J. 1993. "Narrative, Event-structure Analysis, and Causal Interpretation in Historical Sociology." *American Journal of Sociology* 98(5): 1094-1133.

Hedström, P., and P. Bearman. 2009. *The Oxford Handbook of Analytical Sociology*. Oxford: Oxford University Press.

Heise, D. R. 1989. "Modeling Event Structures." *Journal of Mathematical Sociology* 14(2-3): 139-169.

Klabbers, J. H. 2006 "A framework for Artifact Assessment and Theory Testing." *Simulation & Gaming* 37(2): 155-173.

Kleijnen, J. P. C. 1998. "Experimental Design for Sensitivity Analysis, Optimization, and Validation of Simulation Models." In *Handbook of Simulation – Principles, Methodologies, Advances, Applications and Practice*, edited by J. Banks, 173-224. Hoboken, New Jersey: John Wiley & Sons.

Kriz, W. C. 2010. "A Systemic-constructivist Approach to the Facilitation and Debriefing of Simulations and Games." *Simulation & Gaming* 41(5): 663-680.

Langley, A. 2007. "Process Thinking in Strategic Organization." *Strategic Organization* 5(3): 271-282.

Lederman, L.C. 1992. "Debriefing: Toward a Systematic Assessment of Theory and Practice." *Simulation & Gaming* 23(2): 145-160.

Lo, J., J. Van den Hoogen, and S. A. Meijer. 2013. "Using Gaming Simulation Experiments to Test Railway Innovations: Implications for Validity." In *Proceedings of the 2013 Winter Simulation Conference*, edited by R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, 1766-1777. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Manzo, G. 2010. "Analytical Sociology and its Critics." *European Journal of Sociology* 51,1: 129-170.

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

Meijer, S. A. 2012. "Gaming Simulations For Railways: Lessons Learned From Modeling Six Games For The Dutch Infrastructure Management." In *Infrastructure Design, Signaling and Security in Railway*, edited by X. Perpinya, 275-294. Rijeka, Croatia: InTech.

Peters, V. A., and G. A. Vissers. 2004. "A Simple Classification Model for Debriefing Simulation Games." *Simulation & Gaming* 35(1): 70-84.

Pettigrew, A.M. 1992. "The Character and Significance of Strategy Process Research." *Strategic Management Journal* 13(2):5-16.

Raser, J.R. 1969. *Simulations and Society: an Exploration of Scientific Gaming.* Boston, Massachusetts: Allyn & Bacon.

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

Sargent, R. G. 2005. "Verification and Validation of Simulation Models." In *Proceedings of the 2005 Winter Simulation Conference,* edited by M. E. Kuhl, N. M. Steiger, F. B. Armstrong and J. A. Jones, 130-143. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Sewell, W. H. 1992. "A Theory of Structure: Duality, Agency, and Transformation." *American Journal of Sociology* 98(1): 1-29.

Tsoukas, H., and M. J. Hatch. 2001. "Complex Thinking, Complex Practice: the Case for a Narrative Approach to Organizational Complexity." *Human Relations* 54(8): 979-1013.

Van den Hoogen, J., J. Lo, and S. A. Meijer. 2014. The Debriefing of Research Games: A Structured Approach for The Validation of Gaming Simulation Outcomes. In *The Shift from Teaching to Learning,* edited by W. Kriz, 88-99, Bielefeld: W. Bertelsmann Verlag.

Van de Ven, A. H., H. L. Angle, and M. S. Poole. 2000. *Research on the Management of Innovation: The Minnesota Studies*. New York, New York: Oxford University Press.

Von Foerster, H. 1984. "Principles of Self-organization—in a Socio-managerial Context." In: *Self-Organization and Management in Social Systems*, edited by H. Ulrich and G. J. B. Probst, 2-24. Berlin: Springer.

Weber, M. 1949. *The Methodology of the Social Sciences*. New York, New York: Free Press.

AUTHOR BIOGRAPHIES

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 large networked infrastructures and the role of gaming. His email address is j.vandenhoogen@tudelft.nl.

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

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