

Assessing gaming simulation validity for training traffic controllers.

van Lankveld, G.; Sehic, E.; Lo, J.C.; Meijer, S.A.

DOI

[10.1177/1046878116683578](https://doi.org/10.1177/1046878116683578)

Publication date

2016

Document Version

Final published version

Published in

Simulation & Gaming: an international journal of theory, design and research

Citation (APA)

van Lankveld, G., Sehic, E., Lo, J. C., & Meijer, S. A. (2016). Assessing gaming simulation validity for training traffic controllers. *Simulation & Gaming: an international journal of theory, design and research*, 48(2), 219-235 . <https://doi.org/10.1177/1046878116683578>

Important note

To cite this publication, please use the final published version (if applicable). Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights. We will remove access to the work immediately and investigate your claim.

Assessing Gaming Simulation Validity for Training Traffic Controllers

Simulation & Gaming
2017, Vol. 48(2) 219–235
© The Author(s) 2016



Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1046878116683578
journals.sagepub.com/home/sag



G. van Lankveld^{1,2}, E. Sehic³, J. C. Lo¹,
and S. A. Meijer^{1,4}

Abstract

Background. The Dutch **railway** company ProRail is performing large-scale capacity upgrades to their infrastructure network. As part of these upgrades, ProRail uses **gaming simulations** to help prepare **train traffic controllers** for new infrastructure situations. Researching the **validity** of these gaming simulations is essential, since the conclusions drawn from gaming simulation use may result in decisions with large financial and social impact for ProRail and Dutch train passengers.

Aim. In this article, we aim to investigate the validity of the gaming simulations for **training** traffic controllers for new situations in rail infrastructure. We also aim to contribute to the discussion on the minimum level of **fidelity** required to develop and conduct gaming simulations in a valid way.

Method. We investigate the validity by using training sessions in conjunction with **questionnaires**. We based the approach and questionnaires on the earlier work of Raser.

Results. Our results show that the validity of the gaming simulation ranges from medium to good. They also show that while the fidelity of the gaming simulation is not like the real-world operating conditions, this does not reduce validity to low levels.

¹Delft University of Technology, The Netherlands

²Open University in the Netherlands, Heerlen, The Netherlands

³ProRail, Utrecht, The Netherlands

⁴KTH Royal Institute of Technology, Stockholm, Sweden

Corresponding Author:

G. van Lankveld, Open University of the Netherlands Postbus 2960 6401 DL Heerlen, Netherlands.

Email: giel.vanlankveld@ou.nl

Conclusions. We conclude that the gaming simulation used in this study was of medium to good validity. We also conclude that maximum fidelity is not required in order to run a valid gaming simulation session.

Keywords

fidelity, gaming simulation, infrastructure, questionnaire, railway, teaching, traffic controller, train, training, validity

ProRail, the public organization in charge of maintenance of railway infrastructure and train routing in the Netherlands, is substantially increasing its rail traffic capacity (Meijer, 2012b). One of the consequences is the restructuring of infrastructure around train stations, causing significant changes in the work of train traffic controllers. In the new situation, train traffic controllers have far fewer track switches available to reroute train traffic. This means that they have to anticipate routing problems further in advance. Potentially, the restructured situations could lead to problems in routing and consequently to large delays. Examples of potential problems are: (1) an increased number of conflicts between the routing for passenger transport operators and freight transport operators, (2) an accelerated failure rate of remaining track switches, which may lead to traffic jams for the remaining track switches, and (3) a decreased amount of time to route trains to shunting yards. Because of the potential problems with large scale restructuring, ProRail intends to use gaming simulations to prepare controllers before major infrastructure changes take place (Middelkoop, Meijer, Steneker, Sehic, & Mazzarello, 2012). Preparations using gaming simulations may help prevent routing related problems such as disruptions of traffic, and calamities, which have the potential to cause large delays, thereby preventing financial damage as well as preventing delays and discomfort for passengers.

A gaming simulation is a model of a reference system, such as the Dutch railway network (for a more extensive definition of the term gaming simulation we refer to the work of Meijer, 2012a). This simulation allows the user to experience the simulated situation without the risks involved in experiencing these situations in the reference system itself. The gaming simulation in this article mimics a train traffic controller's workstation and shows adapted infrastructure in order to explore future infrastructure situations.

ProRail cooperated with Delft University of Technology in order to model future infrastructure and timetabling situation for railway stations in a flexible gaming simulation environment called PRLGAME (Meijer, 2015). Figure 1 shows an example of an interface of this gaming simulation. ProRail's train traffic controllers play gaming simulations in order to (1) become acquainted with future traffic situations and (2) to facilitate an early recognition of difficult or risky situations (such as the busiest track switches and the available timeframe for using shunting yards). This approach follows developments in gaming simulation theory and educational theory (Amory & Seagram,

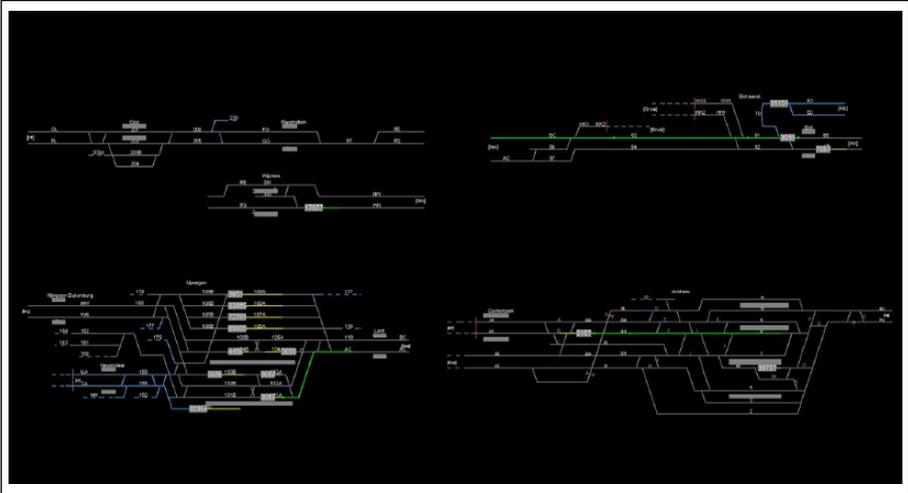


Figure 1. The image shows an example of the modeled future infrastructure in the Utrecht area as presented in the gaming simulation.

2003). Although the same technical gaming simulation framework allows for other purposes as well, in this article we focus on the training capabilities only.

The question we pose in this article is whether the applied gaming simulation is valid for training railway traffic experts. An invalid game may not lead to an increase in traffic controller's efficiency. In a worst-case scenario, invalid gaming simulation may lead to a drop in efficiency and increase routing related problems. Empirical research of gaming simulation validity is necessary because lack of validity may result in lack of a gaming simulation's desired effects. For example, in the work of Whiteley, Leduc, and Dawson (2004) an empirical investigation showed that their gaming simulation does not improve player's knowledge. Whitely et al. stated that this lack of effectiveness could be mitigated by checking (and ensuring) internal validity. We base our approach to investigating the validity of ProRail's gaming simulation on three validity aspects proposed by Raser (1971) and Peters, Vissers, & Heijne (1998). We have not found any pre-existing questionnaires of Raser and Peter's validity; because of this, we have created our own questionnaire.

A second goal of this article is to contribute to the ongoing discussion on the necessity of high simulation fidelity. High fidelity may not contribute positively to the desired learning effects of our gaming simulation (Alessi, 1988; De Winter, Dodou, & Mulder, 2012). In fact, high fidelity may hurt learning performance (Martin & Waag, 1978). The gaming simulation investigated in this research has a high fidelity when observed in the light of Feinstein and Cannon's (2002) framework. E.g., the simulation visually matches a train traffic controller's workstation and behaves in a nearly identical fashion. We investigate whether our gaming simulation does indeed have a high fidelity and we make a connection between the level of fidelity and the validity of the gaming simulation.

Section two (background) discusses previous work on validity, as well as background on the ongoing fidelity in gaming simulation discussion. Section three (methods) presents our research setup, the gaming simulation and scenario, the participants, and the questionnaires that were used. Section four (results) presents and explains a systematic overview of our results, starting with the background information questionnaire and finishing with the validity questionnaire. In section 5 (discussion) we discuss the results in the order of (1) psychological reality, (2) structural validity, (3) process validity, (4) subjects, and (5) fidelity. Finally, Section 6 (conclusions) presents a brief overview of the results followed by conclusions and future work.

Background

Historically, developers have applied gaming simulations with various goals. Peters et al. (1998) identified three categories of tasks for which gaming simulations have been applied: (1) training, (2) research, and (3) policy. In this research, we focus on gaming simulations for training. Validation is required in order to make sure the gaming simulation meets its goals. Because of the variety in possible approaches for reaching the goal, validation becomes a challenging process.

Raser (1971) proposed an approach to validating gaming simulations. In his approach, gaming simulation designers should take into account four relevant aspects of validity. These relevant aspects are: (1) psychological realism, (2) structural validity, (3) process validity, and (4) predictive validity. If a gaming simulation is valid in these four aspects, designers can be confident that the gaming simulation meets its design goals and can be trusted as long as it falls in the category of research gaming simulations. Peters et al. (1998) stated that for the category of training gaming simulations, only three of the four validity aspects are relevant. These relevant aspects are psychological realism, structural validity, and process validity.

In relation to the second goal of this article, fidelity appears to have a partial overlap with the validity aspects proposed by Raser (1971). We come to this conclusion because a decrease in complexity of the simulator would result in both reductions in fidelity as well as a decrease in either structural or process validity.

The Fidelity Discussion

Hamstra, Brydges, Hatala, Zendejas, and Cook (2014) defined fidelity as the degree to which a simulation is perceived as physically similar to the part of the real world that it is meant to simulate. E.g., the degree to which the simulator looks, feels, and acts like the real world.

Although some researchers have expressed doubts about the need for using simulations with high fidelity (as opposed to low fidelity) (Alessi, 1988; De Winter et al., 2012), much effort has been put into developing high fidelity simulations. Since this is of critical importance to health care training success, already a literature search in Google Scholar using the terms “high fidelity” and “nursing” for the years 2014 and 2015 reveals 43 articles with specific emphasis on the level of fidelity in the field of

nurse education alone. As the discussion on this topic suggests, simulators may not need high fidelity. Instead, low fidelity simulations may already be sufficient. If this position is true, efforts in simulation development may be devoted to different areas instead of on fidelity improvement.

Proponents of high fidelity stipulated that high fidelity helps in reaching simulation goals (Klipfel et al., 2011; Weaver, 2011). Some researchers went even further and claimed that some forms of simulation require a sufficiently high fidelity in order to produce useful results (Kadir, Zuhra, & Xu, 2011).

Opponents claimed that the level of fidelity has been shown to be independent of simulation goals in terms of teaching effectiveness (Hamstra et al., 2014). Empirical results support this theory by showing outcomes of simulation-based training that are independent of simulation fidelity (Conlon, Rodgers, Shofer, & Lipschik, 2014). In this research, fidelity may influence some intermediate effects. However, we do not find that fidelity has a significant effect on the overall simulation goals of education effectiveness. If this is true in general, investing effort in raising the fidelity of simulations might be ineffective.

Some opponents advocated a new definition of fidelity. They felt that fidelity should focus less on physical realism and more on functional realism (Hamstra et al., 2014). To the authors of this article, this line of reasoning seems to be in line with Raser's (1971) reasoning. However, Raser used the conceptualization of validity instead of fidelity. Feinstein and Cannon (2001) had also observed this high level of overlap of terms. We may look at Raser's validity as an alternative definition of fidelity. Alternatively, validity and fidelity might be seen as (partially) overlapping because the concepts seem to complement each other. As Feinstein and Cannon (2001) noted, the concepts of physical realism in fidelity and structural validity are very similar. This literature review shows that the fidelity discussion, and specifically the ideal level of fidelity, has not yet been decided. Although we do not claim to solve the issue in this article, we will attempt to provide our own perspective on this matter during the course of this article and contribute to the discussion.

Methods

In this section, we present the setup, the gaming simulation and scenario, and the two questionnaires that were used for this research. We will present our analysis techniques together with our results in the results section.

Research Setup and Preliminary Questionnaire

The testing location for our research was at the ProRail control center for the Utrecht region (called 'post Utrecht'). In this investigation, 22 train traffic controllers participated. The participants had varying levels of experience at post Utrecht and as a train traffic controller in general. Every controller had experience working at the topic of our simulation: the workstation controlling the routing for the Utrecht Central station area.

The investigations consisted of 11 sessions. In every session, we tested two controllers simultaneously. The structure of our sessions was always the same. Two traffic controllers arrived at a previously agreed upon time in the gaming simulation room. Three investigators as well as one or two senior traffic controllers were present in the room. One workstation was available for each train traffic controller. A more extensive explanation of the workstations will be provided in the subsection 'the gaming simulation and scenario'.

We gave instructions before starting the gaming simulation session. At the start, the investigators introduced themselves and the principal investigator explained the goal and the structure of the session. Following the introduction, the senior traffic controllers provided an explanation on the infrastructure changes that would occur in the future and on the changes that were already present in the gaming simulation. Participants filled out a preliminary questionnaire before the start of the experiment. This questionnaire contained five questions on the controller's experience levels and on their interest in participating in the investigation. We instructed the controllers that all the data for this investigation would be processed anonymously and that their answers would not be retraceable to them personally.

After the introduction and the preliminary questionnaire, the gaming simulation scenario started. The scenario lasted for about 60 minutes. After the scenario, the participants completed the validity questionnaire.

Finally, we debriefed the controllers on the new infrastructure situation and on the experiences that the controllers could take away from the gaming simulation. In this debriefing the controllers, the senior controllers, and the investigators participated and all participants had the opportunity to share their experiences of the simulation.

The Gaming Simulation and Scenario

In order to present the contents of the gaming simulation we should first clarify the job description of a train traffic controller. Controllers have the task of routing trains through a control area. The trains run based on a basic rush hour timetabling. This pattern repeats each hour during the course of rush hour with no or minor variations.

In normal traffic circumstances, usually some changes occur to the routing due to delays and problems in the control area. Some of the problems arise due to the limitations of the infrastructure in a given control area, other problems arise due to problems with the trains, with the passengers, or with complications from outside the train system (such as unexpected damage to infrastructure, vandalism, maintenance works, etc.). It is the controller's responsibility to solve as many of these problems as possible by routing trains around problem areas to the best of his abilities.

The contents of the gaming simulation were as follows. Traffic controllers had to work 60 minutes of the basic rush hour timetabling for Utrecht central station. Unlike the regular infrastructure that was present in Utrecht central station in the real world, the simulation contained the infrastructure and basic rush hour timetabling that would be present in the future, after a major infrastructure upgrade. The scenario contained light delays similar to those present in normal traffic, as train traffic controllers simply



Figure 2. The image on the left shows the reference system, the image on the right shows the gaming simulation (we made the train traffic controller anonymous).

needed to become acquainted with the new infrastructure and to the new basic rush hour timetabling.

The major change in the future infrastructure is the removal of 66 out of the 250 track switches. This change allows for far less routing options for traffic controllers. This is the first time in ProRail's history that such a large-scale reduction of switches will be implemented. Therefore, the infrastructure changes constitute a large change in the working routine of the traffic controllers. The severity of the change may influence traffic controller responses. However, since this is also the first time this type of validity investigation takes place we are unable to compare responses to other investigations.

The gaming simulation session used a replication of the real workstation (computer) setup with four screens and a replication of the software the controllers normally use, called PRLGAME. In the simulation software, we implemented the infrastructure as it would be in the future, after the infrastructure update had taken place. Part of the software was an extensive simulation algorithm that mimics the movements of the trains through the simulated traffic area. The software is capable of simulating delays and other problems but for the purpose of this investigation, we used an undisturbed situation. We based the routing tables of the trains on the routing tables that will be used after the infrastructure update had taken place. Figure 2 provides an impression on what the reference system (the traffic controller's screens) looks like and on what the gaming simulation looks like. Although a telephone is present in the gaming simulation image, it was non-functional and significantly different from the phone system that is normally used.

Two features in our gaming simulation differ from the real traffic situation and workstation. Firstly, the used simulation has no implementation of the safety actions that traffic controllers should take in case of specific delays or disturbances. We considered the absence of this feature as irrelevant for our investigation because the intended purpose of this gaming simulation was to train the traffic controllers in the new infrastructure setting without disturbances. The second difference is the fact that our game had no communication between controllers and train drivers or other parties

involved in normal traffic control. This was also no relevant part of the investigation because the scenarios did not require any communications given the small delays.

The Background Information Questionnaire

The preliminary questionnaire contained five questions. The first three questions concerned the experience level of the participating traffic controller, the fourth question asked about the level of interest the controller had in participating in the investigation, and the final question was about whether the controller also had experience as planner for the simulated workstation.

The Validity Questionnaire

The validity questionnaire consisted of 18 questions about the gaming simulation's validity. In the questionnaire, we posed three questions per validity aspect. The questionnaire was built up of an earlier questionnaire used by Lo, Sehic, and Meijer (2014) with an addition of nine new validity items. The questionnaire focused on validity measurements of the simulated environment in relation to the task, e.g. 'the representation of the time tables is sufficient for the task I perform in the simulator' for structural validity. We refer to these original items as 'VT' (validity – task) items. The new items focused on validity with regards to the similarity of the simulated environment in relation to the work environment, e.g. 'the simulated workplace looks the same as normal' for structural validity. We refer to these items as 'VW' (validity - workplace) items.

Results

In this section, we provide the results of this investigation and the analysis methods that we used to obtain them. We present results per individual questionnaire item and per validity aspect.

The Background Questionnaire

We used this questionnaire to control potentially confounding variables. The results for the background questionnaire questions can be seen in Table 1. We translated the questions in all the results tables from Dutch to English wherever translation was required. The level of interest item gauged the willingness of traffic controllers to participate in this research.

We found no correlation between the outcomes shown in Table 1 and the scores on the validity questionnaire.

Means and Standard Deviations for the Validity Questionnaire

Table 2 shows the means of the 18 validity items in the validity questionnaire. We sorted Table 2 by mean values in descending order. The first column contains a code identifying the question number and section of the questionnaire in which it was

Table 1. Means and Standard Deviations (SD) of the First Questionnaire.

	N	Mean	SD
How many years of experience do you have as controller in your current workstation?	22	11.52	9.67
How many years of experience do you have working for ProRail?	22	13.88	12.95
How many years of experience do you have working at the Utrecht workstation?	21	8.86	8.47
What is your level of interest in participating in this gaming simulation? (1-5)	22	3.95	1.05

contained. The original questions by Lo et al. (2014) are identified by the letter ‘VT’ (validity task) while new questions are identified by ‘VW’ (validity workplace). A salient detail is that item VT4 had been inverted (this inversion is already accounted for in the mean result value). Since all participants were in the same group between groups analyses were not necessary. Table 2 shows the trend that simulation VT questions are rated higher than VW simulation questions.

The questionnaire results can be found in Table 2. All questions are Likert type questions with possible answers ranging from one to five. Likert type items are items containing a statement for which the subject has to indicate his amount of agreement. For this type of item, choosing one (1) indicates that the subject completely disagrees with the item statement and choosing five (5) indicates that the subject completely agrees with the item. The item scores follow normal distributions and may therefore be analyzed as interval variables.

Overall Results for the Validity Aspects

We constructed the validity questionnaire to measure the three aspects of Raser’s (1971) method (psychological realism, structural validity, process validity). We constructed six items per aspect in the second questionnaire. We divided these items equally between the VT and the VW part of the questionnaire. The averages for each aspect can be found in Table 2 by referring to the category column.

Results Per Validity Aspect

We have calculated group averages for categories containing the three validity aspects for VT items and for VW items. These calculations resulted in six means, which can be found in Table 3. The results show that all validity means fall around 3.00 or higher except for one (mean = 2.80 for process validity in the VW items).

Discussion

In this section, we discuss and interpret the obtained results. We start with psychological reality, followed by structural validity, then process validity, then the number subjects in relation to validity, and finally we conclude with fidelity related points.

Table 2. Means, Standard Deviations (SD), and Item Categories.

		N	Mean	SD	Item category
VT7	The simulated scenario is similar to a situation that appears in real life	21	4.19	.68	Process validity
VT1	The representation of the time tables is sufficient for the task I perform in the simulator	22	3.91	.81	Psychological reality
VT3	The infrastructure model is sufficiently realistic for the task in the simulator	22	3.82	.73	Psychological reality
VT5	The train movements in the simulator work with a similar process to those in reality	21	3.67	.91	Process validity
VT8	Information from sources in the simulator can be used in the same way as the information in reality	21	3.61	1.02	Structural validity
VT2	The simulation environment felt like my normal working environment	22	3.45	1.14	Psychological reality
VT6	The simulator contains the necessary functionalities to perform the task set in the simulation	22	3.32	1.13	Structural validity
VW3	The train movements appear realistic	22	3.32	1.04	Psychological reality
VW6	All normal infrastructure and systems are present in the simulation	19	3.26	.81	Structural validity
VW4	All normal hardware is present in the simulation	20	3.25	1.12	Structural validity
VW1	The simulated workplace looks the same as normal	22	3.09	1.19	Psychological reality
VT4	(Inverted) I do not have all the necessary information needed to perform my task in the simulator	22	2.95	.95	Structural validity
VW2	The software looks the same as normal	22	2.91	1.06	Psychological reality
VT9	The processes (interactions, communication) in the simulator are the same as those found in a similar situation at my workplace	22	2.86	1.08	Process validity
VW7	The computer equipment works the same as at a real workplace	21	2.86	1.06	Process validity
VW9	The trains react normally	21	2.76	.83	Process validity
VW8	The software work the same as normal	20	2.65	.93	Process validity
VW5	All the normal usage options are available in the software	21	2.57	.87	Structural validity

Note. Answers ranged from 1 to 5.

Table 3. Mean and Standard Deviation (SD) Per Validity Aspect.

		N	Mean	SD
Validity task	Psychological reality	22	3.73	.71
	Process validity	20	3.62	.66
	Structural validity	21	3.33	.74
Validity workplace	Psychological reality	22	3.11	.92
	Process validity	20	2.80	.80
	Structural validity	17	3.16	.80

Psychological Reality

As reported earlier in this article, we have measured psychological reality using six variables. We can see that psychological reality has a high average score, especially in the VT section of the questionnaire (mean = 3.7 out of 5). This response average shows that traffic controllers consider the simulated workplace to feel very similar to the real world workplace. Furthermore, this score indicates that traffic controller’s perceptions are is that the representation of the simulated timetables and the simulated infrastructure appear to be fairly realistic.

We notice that the items with lower scores (Mean = 3.00 or lower) are related to the appearance of the software and the appearance of the simulated workplace. We may explain this result by noting that the original software has a slightly different layout compared to the simulated software. This difference in layout is caused by the difference in programming languages used to produce the software, this results in slightly different implementations. However, most of the simulation software’s appearance is the same as the original software, which most likely is the cause for the other reasonable evaluation scores (3.00 or higher for the other software items).

The appearance and organization of the workplace also receives a lower score than expected (Score = 3.09, a score in the region of 4.00 was expected). One possible explanation may be the lack of telephone equipment in the setup. We left this equipment out because, as explained earlier, our simulation lacked the disturbances that required communication. We may be able to remedy this issue by simply placing the phone equipment and not using it. Furthermore, the physical setup and layout of the testing room different from the real traffic controller’s room because the actual traffic control room was in use in the daily routing activities. This may also account for lower perceived realism in appearance.

Structural Validity

Our research shows a slightly above average structural validity score (Mean = 3.33 and Mean = 3.16). We can interpret this as structural validity being slightly above sufficient. We can see this result in both individual questionnaire items and in averages, for both the VT and the VW part of the questionnaire. The only exception to this result

is the item 'VW5': 'All the normal usage options are available in the software'. This item is slightly below average (Score = 2.57).

Raser (1971) explained structural validity as the isomorphism between a simulator's structure and the structure of the reference system. Raser's structure refers to all physical qualities as well as actors and use options.

The goal of our simulator was to enable train traffic controllers to explore the future infrastructure of Utrecht Central station as a form of training. As explained earlier, we focused on implementing the required software and providing a correct representation of future infrastructure. However, in our design process we have not implemented the safety and communication tasks that are normally also present in the work of controllers. However, routing related communication was possible between the two simultaneously testing traffic controllers. It is evident that the absence of safety and communication has led to some simplification of the simulator's structure. We feel that this difference between the simulator and reality may be responsible for the lower scores on the structure items.

Conversation with the controllers (both during the debriefing and informal) has revealed that controllers missed the safety aspect in the simulation. Controllers made several remarks about this difference. Curiously, the scores on the psychological realism items indicate that the absence of the safety and communication did not seem to influence the opinions that the simulator appeared to be realistic.

The apparent opposition between the mean psychological reality and the mean structural validity leads us to wonder whether we need to strive for a high level of structural validity at all for the purpose of training (such as in our simulator). Possibly, a simulation needs only a minimal level of structural validity. We advise that further research is needed to show whether low structural validity leads to reduced learning effect, especially if psychological reality is high. In these findings, we recognize the long on-going discussions on realism and complexity versus learning effectiveness (Dittrich, 1976). We also recognize the discussion on the amount of focus games should have on realism and verisimilitude as opposed to on holistic and other features (Myers, 1999). Finally, we can relate this to the discussion of realism versus symbolic representation (Dormans, 2011).

Process Validity

In process validity, we see results varying widely from below to above average (Scores from 3.61 to 2.57). In the VT part of the questionnaire, we see that two out of the three items have an above average score (Scores 3.61 and 3.32). Contrastingly, in the VW items we see a below average score in all process validity items.

We note that the below average item in the VT part may be explained by the presence of the word 'communication' in the item. Since the entire communication process was absent from the simulation (and no telephone equipment was present) it is feasible that this prompted controllers to score this item lower.

In the below average process validity in VW items we note that a common attribute of all these items is the word 'normal'. We also find this term in the lower scoring items of psychological reality. It is possible that this term puts controllers in a more

critical frame of mind. It is of course also possible that items in the VW part of our questionnaire were not representative enough for process validity. Alternatively, including subjective terms such as “normal”, “similar to”, and “sufficient” may invite additional variability because of the multi-interpretable nature of the terms.

As a concluding remark for the validity aspects, we note that whether something feels real (psychological reality) may be related to the way something behaves (process validity). It is also true that if a process has all the expected parts (which indicate structural validity), abnormalities in behavior (which would indicate a lack of process validity) may not be recognized, especially in a system that has a lot of variation and complexity in its behavior in real life. We suspect that the validity aspects proposed by Raser (1971) may not be independent of one another. We can firmly state that this topic merits further investigation.

Subjects

We performed this research using ProRail personnel as research subjects. This approach brings along the benefit that the validity of the simulation is being investigated with the future user group of this specific simulation. The downside of this group of subjects is that we are working with a small number of subjects (looking from the perspective of statistical methods). Aside from the small number of subjects, a difference between the subject sample and the population did not present itself. In other words, all possible individuals that could in the future be using this particular simulation and scenario combination were actually using it. However, it is true that with the hiring of new traffic controllers in the future the population will change.

The consequences of having a small number of subjects are that smaller statistical effects may be missed as being significant or effects may be missed because of low statistical power. In order to prevent this type of problem we need larger sample sizes.

Options to increase the sample size do exist. We could use additional students in pilot studies or we could use traffic controllers from posts other than post Utrecht central. Disadvantages to using these alternative sources of subjects are that these subjects have far less experience. In the case of the students, they would even have to be (partially) educated as traffic controllers before they would be capable of using the workstation for post Utrecht at all. The required education for students makes adding students as a sample nearly untenable. However, testing controllers from different posts should be considered as a serious option.

Fidelity

In relation to the long-running realism versus symbolism discussion, our results support the position that high physical realism is not necessarily a prerequisite of a high-perceived realism. Our results show that psychological reality items can be high even though the structural validity is rated lower. We posit that leaving details out of the simulator makes it less complex and more abstract and symbolic. In our perspective, a game may be seen as a system with multiple parts in which a reduction of fidelity for one part may not influence the fidelity on other parts of the system. An example in this

study is leaving a whole process like communication and a physical phone out of the simulator does not reduce the perception of realism for the simulator overall. The players seem to be able to compartmentalize their gaming simulation perception. Further research is required to test this hypothesis.

Some researchers have claimed that striving for maximum realism can make flaws in a simulation's realism more obvious and may focus attention of players on those gaps instead of on the intended contents and lessons of the simulation (Alessi, 1988). In our gaming simulation, gaps in realism were obvious to the participants (they were able to spot flaws in structural validity) and may have been the cause of lowered validity. However, the gaps had no functional bearing on the simulation goals or simulation success for the players (psychological reality remained high). We feel that the need for high levels of fidelity is related to the exact definition of fidelity and to the requirements of the gaming simulation that is being considered.

Generalizability

In this research, we have investigated the validity of our gaming simulation. Due to the specific nature of our system: computer assisted simulation combined with the topic of train traffic controller training, and furthermore due to the low number of participants, we assume that issues will arise with replicability. Therefore, we feel that using a close replication of our methodology is not guaranteed to yield good results in a different ProRail traffic control post. Specifically, with regards to the physical setup and the ability of the computer simulation to accurately simulate the issues of each given ProRail post. However, we feel we have demonstrated that through the careful consideration of validation items the questionnaire technique can be re-used in the ProRail railway systems on different locations. This approach to validation may even be used in different countries for different railway management companies if the proper care is taken. Finally, the validation of gaming simulation systems is an important issue. Gaming simulation developers should always consider whether their systems validly portray the reference system that they aim to simulate.

Conclusions

In this investigation, we have used questionnaires to investigate three facets of validity proposed by Raser (1971) and Peters et al. (1998). We have done this investigation during a gaming simulation session for the ProRail organization using a mockup workstation for train traffic controllers. ProRail used the simulation as a training method to prepare train traffic controllers for future railway infrastructure configurations. Specifically, we have investigated the validity of gaming simulation training for the future infrastructure of Utrecht central station.

We may conclude that the subjective ratings for our simulation's validity were good (overall above average). Our findings show that validity scores for this simulation ranged from slightly below average (around 3 out of 5) to good (scores of 4 out of a

maximum of 5). These results are applicable to individual items and to averaged scores over types of validity. Both in the questionnaires and in the debriefing, participants indicated approval of the representation of the railway systems and the systems used in day-to-day operations. As for the relevance of high fidelity, we conclude that our findings support the notion that higher fidelity does not necessarily mean higher perceived simulation effectiveness. Our results show that psychological reality may remain sufficient even with lapses in fidelity (stimulation missing the entire communication aspect) that are obvious to the gaming simulation players.

In future work investigations could focus on research into the learning outcomes for different levels of validity. We should also focus on improving sample size in order to increase statistical power and improve significance levels. The current questionnaire items are fully representative of the concepts they are intended to measure in a theoretical sense. However, to continue using the current validity questionnaires in the future it would be advisable to perform a methodological analysis of the items to see if all relevant sides of the validity constructs are covered and to see if the item quality is high enough. One technique to consider for evaluating and potentially improving the quality would be confirmatory factor analysis.

Furthermore, a potentially interesting research topic would be the compartmentalization of fidelity. One option would be to create gaming simulation systems in which it is possible to manipulate the fidelity of its parts in order to tease apart the types of manipulations that influence the overall perception of the system and the teaching effectiveness of the system. Finally, we consider investigating overlap between Raser's structural and process validity to be a very promising research topic.

Acknowledgment

We would like to thank our colleague Gert-Jan Stolk for his efforts and support in developing our gaming simulation.

Declaration of Conflicting Interests

The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: The authors are the designers and developers of the gaming simulation described in this research. They are not responsible for the education of train traffic controllers. This research was partly (50%) funded by ProRail, the final user of the gaming simulations in this research. Therefore, there are potential conflicts of interest with regards to the success of gaming simulations for ProRail with regards to research success and publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded through the Railway Gaming Suite program, a joint project by ProRail and Delft University of Technology NWO EXPLORAIL RAILROAD (no nr) Dutch National Science Foundation grant, 50% financed by the Dutch Rail Administration (ProRail).

References

- Alessi, S. M. (1988). Fidelity in the design of instructional simulations. *Journal of Computer-Based Instruction*, 15(2), 40-47.
- Amory, A., & Seagram, R. (2003). Educational game models: Conceptualization and evaluation. *South African Journal of Higher Education*, 17(2), 206-217.
- Conlon, L. W., Rodgers, D. L., Shofer, F. S., & Lipschik, G. Y. (2014). Impact of levels of simulation fidelity on training of interns in ACLS. *Hospital Practice*, 42(4), 135-141.
- De Winter, J. C. F., Dodou, D., & Mulder, M. (2012). Training effectiveness of whole body flight simulator motion: A comprehensive meta-analysis. *The International Journal of Aviation Psychology*, 22(2), 164-183.
- Dittrich, J. E. (1976). Realism in business games: A three game comparison. *Developments in Business Simulation and Experiential Learning*, 3, 273-280.
- Dormans, J. (2011). Beyond iconic simulation. *Simulation & Gaming*, 42(5), 610-631.
- Feinstein, A. H., & Cannon, H. M. (2001). Fidelity, verifiability, and validity of simulation: Constructs for evaluation. *Developments in Business Simulation and Experiential Learning*, 28, 57-67.
- Feinstein, A. H., & Cannon, H. M. (2002). Constructs of simulation evaluation. *Simulation & Gaming*, 33(4), 425-440.
- Hamstra, S. J., Brydges, R., Hatala, R., Zendejas, B., & Cook, D. A. (2014). Reconsidering fidelity in simulation-based training. *Academic Medicine*, 89(3), 387-392.
- Kadir, A., Zuhra, A., & Xu, X. (2011). Towards high-fidelity machining simulation. *Journal of Manufacturing Systems*, 30(3), 175-186.
- Klipfel, J. M., Gettman, M. T., Johnson, K. M., Olson, M. E., Derscheid, D. J., Maxson, P. M., . . . Vierstraete, H. T. (2011). Using high-fidelity simulation to develop nurse-physician teams. *Journal of Continuing Education in Nursing*, 42(8), 347-357.
- Lo, J. C., Sehic, E., & Meijer, S. A. Explicit or implicit situation awareness? Situation awareness measurements of train traffic controllers in a monitoring mode. In D. Harris (Ed.), *Engineering psychology and cognitive ergonomics* (pp. 511-521). Switzerland, Europe: Springer International Publishing.
- Martin, E. L., & Waag, W. L. (1978). *Contributions of platform motion to simulator training effectiveness: Study I—Basic contact*. Brooks Air Force Base, TX: Air Force Human Resources Laboratory.
- Meijer, S. A. (2012a). Gaming simulations for railways: Lessons learned from modeling six games for the Dutch infrastructure management. In X. Perpinya (Ed.), *Infrastructure design, signaling and security in railway* (pp. 275-294). Rijeka, Croatia: IntechOpen.
- Meijer, S. A. (2012b). Introducing gaming simulation in the Dutch railways. *Procedia - Social and Behavioral Sciences*, 48, 41-51.
- Meijer, S. A. (2015). The power of sponges: Comparing high-tech and low-tech gaming for innovation. *Simulation & Gaming*, 46(5), 512-535.
- Middelkoop, D., Meijer, S., Steneker, J., Sehic, E., & Mazzarello, M. (2012). Simulation backbone for gaming simulation in railways: A case study. In C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, & A.M. Uhrmacher, (Eds.), *Proceedings of the Winter Simulation Conference (WSC'12)* (pp. 3262-3274). Berlin, German: IEEE.
- Myers, D. (1999). Simulation as play: A semiotic analysis. *Simulation & Gaming*, 30(2), 147-162.
- Peters, V., Vissers, G., & Heijne, G. (1998). The validity of games. *Simulation & Gaming*, 29(1), 20-30.

- Raser, J. (1971). *Simulation and society: An exploration of scientific gaming*. Boston, MA: Allyn & Bacon.
- Weaver, A. (2011). High-fidelity patient simulation in nursing education: An integrative review. *Nursing Education Perspectives*, 32(1), 37-40.
- Whiteley, T. R., Leduc, R., & Dawson, B. (2004). A cognitive investigation of the internal validity of a management strategy simulation game. *Developments in Business Simulation and Experiential Learning*, 31, 290-298.

Author Biographies

G. van Lankveld is a post-doctoral researcher at Welten Institute at the Open University of the Netherlands. His research focuses on player modeling, psychological profiling, and validation of assessment methods and models in serious games and simulations.

Contact: giel.vanlankveld@ou.nl.

E. Sehic is project manager at the innovation and development department at ProRail, the Dutch rail infrastructure manager. He is responsible for the subproject of the Railway Gaming Suite, linking of Prorail simulation tools for gaming simulation. He is also PhD student at Delft University of Technology.

Contact: emdzad.sehic@prorail.nl.

J. C. 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 using (gaming) simulation methods.

Contact: j.c.lo@tudelft.nl.

S. A. Meijer is a professor in Health Care Logistics at the KTH Royal Institute of Technology, Stockholm, Sweden. He is also part-time associate professor at Delft University of Technology in The Netherlands. He is specialised in gaming simulation and other interactive methods to involve the operational level of organisations in innovation processes.

Contact: sebastiaan.meijer@sth.kth.se.s