The Goal of Scoring: Exploring the Role of Game Performance in Educational Games

Casper Harteveld
Northeastern University
140 Meserve Hall
360 Huntington Ave
Boston, MA 02115
c.harteveld@neu.edu

Steven C. Sutherland
Northeastern University
567 Holmes Hall
360 Huntington Ave
Boston, MA 02115
st.sutherland@neu.edu

ABSTRACT
In this paper the role of game performance as an assessment tool is explored and an approach is presented for designing and assessing learning-centered game scores. In recent years, attention has shifted from focusing on games for learning to games for assessment. The research question this paper tries to address is how valid games are as an assessment tool, and more specifically, how valid the use of game scores are as a measure of assessment. To explore this use, we looked at the role of game performance in a game where the goals were designed based on its learning objectives. We hypothesized that because of this design the scores could be used as a measure of learning. The results of our mixed-methods study confirmed this hypothesis. However, the scores are influenced by factors such as computer skills and age. Further analysis revealed that the design of the game and the game-based training were also of influence. These insights will help in designing better predictive game scores in the future.

Author Keywords
game-based assessment; education; game design; game analytics; game-based training.

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous; I.6.8. Simulation and Modeling: Types of Simulation—*Gaming*; K.4.3. Computers and Society: Organizational Impacts

INTRODUCTION
In recent years, the idea of using games for assessment has sparked a widespread interest. In fact, according to Shaffer and Gee [21] "We have been designing *games for learning* when we should have been designing *games for testing*" (p. 3). However, its use so far has only been ascribed as a great promise, but because of its potential it has increasingly been getting more attention in academia and industry, see e.g., [14, 23]. This increased interest in this potential use of games has likely been due to the emerging field of *game analytics* [20]. While game analytics (analogous to learning analytics) is primarily used by the gaming industry for business rather than educational purposes, its use for education holds great promise. By instrumenting games, behavioral telemetry data can be retrieved to analyze player behavior. In the context of education and this paper, this data allows instructors to get the needed insight into student learning.

To date a few commercial examples exist (e.g., *Balloon Brigade*) and some initial, promising academic groundwork [22, 24]. One of the recent projects using game-based assessment is *SimCityEDU* [14]. This adapted educational version of the *SimCity* franchise, which the developers describe on their website as a "game-based learning and assessment tool for middle school students covering the Common Core and Next Generation Science Standards," provides tools to formatively assess students' problem-solving abilities, explanations of relationships in complex systems, and abilities to read informational texts and diagrams. However, until now little research exists to document its success and few provide detailed, validated guidelines on how to accomplish game-based assessment [23]. To that end, this paper seeks to answer how valid games are as an assessment tool. In this research the focus will be on the role of *game performance* (also known as game score) as a measure of assessment for learning.

To explore the role of game performance, and thereby retrieve evidence for the validity of games as assessment tool, we investigated the design and evaluation of the game *Levee Patroller*. This unique game was developed in 2006 by the first author and its name refers to the game's target group. Levee patrollers inspect levees, the artificial and natural barriers that protect a region from flooding, and report any risks they encounter. Similar to the actual practice it simulates, in the game players have to find all virtual failures in a region and report them. If they do not find the failures in time or report them incorrectly, it could result in a levee breach that floods the whole virtual region. In this game, the goals were purposefully designed based on its learning objectives. We hypothesized that because of this design the scores could be used as a measure of learning. We will discuss the details of the design of *Levee Patroller* in a later section, after presenting the related work that informed this study.

CHI 2015, April 18–23, 2015, Seoul, Republic of Korea.

http://dx.doi.org/10.1145/2702123.2702606

Our work contributes to the use of games and gamification for training and education, which are increasingly applied in and beyond HCI. A need exists for guidelines and evidence for these applications for learning and assessment. Our main contributions to HCI are the design approach for creating *learning-centered game scores* around learning objectives and the validation approach through a unique mixed-methods study and by dissecting player behavior using game data. By providing the necessary details and discussing the limitations in a careful manner, this paper can serve as an exemplary study for the design and evaluation of gameful artifacts for education and training.

BACKGROUND

Before arguing for using game performance as an assessment of learning, it is necessary to highlight that performance and learning are not necessarily similar [27]. A person can learn much and not perform, whereas another can perform excellent and not learn much. However, if performance improves, then we can speak of learning. The traditional role of scores is to indicate how well the player performs in the game. Therefore, scores essentially measure performance. If that performance is related to learning content, scores are arguably predictive of learning/performance regarding the content. Despite this logic, the use of game scores for assessment is not well understood and certainly not in widespread use. In this section we provide the theoretical background for when and why scores can be predictive of performance. This discussion is of immediate relevance to the related work of gamification, and specifically gamification of education [11], and this will help to see the broader relevance of our work. We further elaborate on how game data such as scores can be investigated to clarify what approach we have taken in this paper.

Motivation to Learn

As a measure of performance, scores are generally viewed as an extrinsic motivator used to reinforce competition and engage players, while the desire to master a subject (learning) is an intrinsic motivator. By tying scores to learning outcomes and using them as an assessment of learning, it provides the score as a cue to achieving mastery rather than for obtaining the score itself. Such learning-centered game scores provide useful formative assessment (learning psychology) or informational feedback (motivational psychology) on competence [24]. Existing literature suggests that *intrinsic motivation* (IM) is highly correlated with *self-efficacy* (SE) [18, 28], which is one's belief in one's own ability to perform. If scores constitute informational feedback on competence, then it will support IM through SE. Scores that are used as extrinsic motivators have the reverse effect, as *extrinsic motivation* (EM) is known to undermine IM. Therefore, we argue that scores are more likely to support assessment, learning, and motivation if they are clearly connected to learning objectives.

Gamification for Education

The distinction between intrinsic versus extrinsic motivation is key to understanding successful adaptation for using learning-centered game scores in gamification applications for education [11]. The main criticism against the majority of gamification applications is that the game elements are used as extrinsic motivators [16]. Scores, points, badges, and achievements, which are elements that are frequently associated as part of implementing gamification, are all used as external rewards. Grades are too. Several scholars are arguing that more meaningful gamification needs to happen, which can happen by using game design elements to provide informational feedback to the user, which can then foster IM and less so EM. In 2013 a symposium was put together to go beyond badges and points and create for more meaningful assessment systems through gamification by putting together "an elaborated research and design agenda" [6]. Our work is meant to further this design agenda by suggesting carefully designed in-game scores for performance assessment.

Analytics for Assessment

There are two popular approaches for game data: (1) low-level behavioral data collection (unlabeled data); and (2) interpreted behavioral data collection (labeled data) [20]. Although we agree that there are benefits to the first approach, this paper involves the second approach. The first approach is largely structured around interpreting behavioral data to make sense of player characteristics or, in educational games, to infer necessary competencies. The second approach assumes knowledge already exists about what the data mean. For establishing assessment tools, a common method under the first approach is evidence-centered design (ECD) of measures [14, 24, 23]. In this paper we suggest a framework under the second approach that relies upon subject-matter experts to identify key competencies, to provide subject-specific information for labelling data, and ensuring the correct data and in-game behaviors are represented in scores. Similar to the ECD framework, we believe that it is necessary to validate the behavioral assessments of the competencies they are intended to assess. We provide one example, as an exemplar of our framework and show how scores were validated as an assessment for the key competencies identified by experts.

DESIGN

To understand why the game *Levee Patroller* is a good case to investigate the meaning of game performance, it requires providing more detail about the design of the game [7], in terms of the learning objectives and gameplay. Based on the design, we will explain what specific role we hypothesized that the game performance would have.

Learning Objectives

Levee Patroller was developed in the Netherlands, a country where levee inspection has deep historical roots. Levee Patrollers are considered the "eyes and ears" of the Dutch water authorities, which are organizations that are responsible for the water quality, quantity, and safety in the Netherlands. The reason for the game's development is that the patrollers have to deal with rare but disastrous failures. Levee failures hardly occur and it is difficult to get practical experience. In fact, despite its "virtuality" the game provides the only means to get experience in finding and reporting levee failures. The consequences of a potential failure are furthermore too large, and technology can only assist with inspecting levees to some extent. For example, there are simply too many levees to equip them all with sensors. Therefore, personnel is required that

knows what the risks are and how to deal with them. The game *Levee Patroller* was developed to fulfill this gap.

When designing a game as an assessment tool it is important to involve subject-matter experts in the design process to ensure that the goals of and measures in the game are aligned with the desired skills relevant to the task being learned. For that reason, *Levee Patroller* was (co-)developed in close collaboration with the water authorities. From the interactions it became clear that the exact roles and responsibilities differ per water authority for patrollers, and also how they are implemented. The responsibilities are generally to: 1) find failures; 2) report signals that make up a failure; 3) communicate reports to others; 4) diagnose the situation; and 5) take measures when necessary. Patrollers report their findings to a coordinating field office. This office collects all of the information and provides further instructions to individuals in the field. In developing the game, five learning objectives were identified based on these responsibilities: observing, reporting, assessing, diagnosing, and taking measures.

An important design decision was to use the learning objectives to design the activities in the game and as criteria for a game score. Table 1 provides an overview of the criteria in the game, what they mean, and what learning objective is associated with each. There are two criteria for observing, one for observing the failure itself and one for recognizing the signals. A failure could consist of one or more signals, so it is important to make this distinction. For reporting we included two criteria as well based on consultation with the water authorities. They mentioned that reporting the location is an important and often overlooked aspect in the inspection process and they wanted to stress this importance in the game. By weighting the criteria and summing the weighted scores for each failure in a scenario, a *total score* is calculated. This total score is presented as a percentage of the maximum score that a player can achieve in a scenario. The weights for each criterion are mentioned in Table 1. By working with subject-matter experts, game developers can build scores and weighting schemes to better highlight the importance of key objectives and to ensure scores are tied to and assess the most important learning outcomes.

Gameplay

In this single player game, players can walk freely around a 3D environmentthe purpose of the movement being to find all potential failures and appropriately deal with them using the tools of the game (e.g., handbook, map, etc.). Dealing with failures requires players to *recognize* them. Subsequently, players have to *report* the location of the problem and the signals that constitute the potential failures and their characteristics. Next the player must communicate the findings to a field office and *assess* the severity of the situation. Failures can change over time so players need to return to the problem locations to determine if the failure potential worsened. If it did, players make another report to a field office. If the failure becomes critical, i.e., it is likely a levee breach will occur, players have to indicate this to the field office. After *diagnosing* the situation, which involves determining what mechanism is causing the impending failure, players can *take action* and implement a stabilizing measure. If they do not implement this in time, a catastrophic failure and flooding results. Throughout and at the end of an exercise, players receive feedback on how they recognized, reported, assessed, diagnosed, and took action against failures.

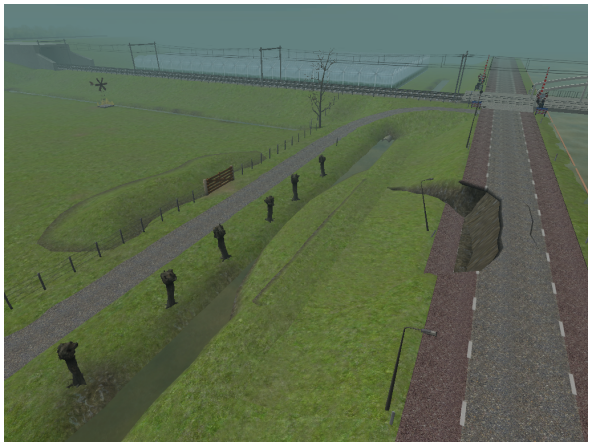

Figure 1. Screen shot of a virtual failure. ©Deltares

The game has different regions each with specific characteristics and includes a variety of failure types. Figure 1 illustrates one of the failures. A training level was created to teach players how to navigate the 3D environment and teach them the various tools they can use to deal with failures.

Hypotheses

We had several expectations with regards to the role of game performance based on the design of the game. First, since the learning objectives are directly coupled to the game scores, it is arguable that if someone is better in playing the game, that person has more knowledge and skills about levee inspection.

HYPOTHESIS 1. *Game scores will be positively related to the results on the main learning outcomes; participants with higher game scores will have higher results on the main learning outcomes.*

Other than that care is needed in equating performance to learning [27], what people learn may be modified by the kind of skills they have. Considering that *Levee Patroller* is a digital game, players will need to have considerable computer skills. Participants who have the necessary skills can pick up the game faster and are better able to focus on the content. For participants with significant experience in playing games this will arguably have an even stronger effect.

HYPOTHESIS 2. *Computer skills (Hypothesis 2.1) and game skills (Hypothesis 2.2) will be positively related to the game scores; participants with higher computer and game skills will have higher game scores, respectively.*

A frequent re-appearing argument is the division between people who have grown up playing digital games and those who have not [17]. People growing up with games have different expectations and skills, and it can be expected that they are better in playing games.

Criterion	Weight	Description	Learning Objective
Observed failures	10	Finding a failure	Observing
Location accuracy	1	Specifying the failure location	Reporting
Observed signals	5	Reporting what signals are part of a failure	Observing
Reporting accuracy	5	Filling out a report for each signal	Reporting
Assessment accuracy	2	Estimating how severe the situation is	Assessing
Diagnose accuracy	5	Determining the failure pattern	Diagnosing
Measure effectiveness	5	Taking an action to prevent flooding	Taking measures

Table 1. Overview of Levee Patroller's scoring system.

HYPOTHESIS 3. *Age will be negatively related to games scores; younger participants will achieve higher game scores compared to older participants.*

METHOD

A game-based training with *Levee Patroller* was designed and implemented to study the game's effectiveness and retrieve a substantiated understanding on the use of games game for training [8]. This training was designed by keeping the needs for a successful training in mind, such as the ability for extensive practice, with some common sense ideas about the willingness and commitment to participate, and with practical considerations on how it could be implemented. To have people participate over an extended period of time, it was deemed important to allow the participants to *play at home*. In addition, studies suggest that players feel more involved in the game and feel a higher level of competence compared to playing the same game in a laboratory setting [26], making playing a game that tests performance a better experience when at home. The training length was set to three weeks. We were concerned that a shorter duration would not give enough time for practice and a longer duration would be too strong of a commitment for a voluntary activity and lead to attrition.

Having people play at home also has its disadvantages, and the many online learning programs confirm that the majority of the learners do not complete them [13]. The following was done to increase commitment of the participants. First, participants would physically attend a start- and end-meeting. The end-meeting established a clear deadline to finish all the assigned exercises, whereas the start-meeting made sure that the participants went home knowing what to do. In between the two meetings, they would play at home. The meetings seemed the perfect compromise in giving the participants the flexibility to play at home. Including the meetings had the additional advantage of making sure that everyone completed the research material. Also, considering the target audience, a meeting to play under guidance was considered necessary.

Second, *weekly assignments* were included. Every week participants received a code to unlock the next two exercises of the game. The assignments enforced a structured manner for playing the game and spread the game play across the training. It also provided an unobtrusive reminder to participants to play the game and for the facilitator to stay in touch.

Participants

In total 145 participants were recruited with the help of three participating water authorities in the Netherlands. The majority were volunteers who help out when needed; the remaining participants were employees from the water authorities. The participants were relatively old; were practically all male; came from varying educational backgrounds and diverse occupations, with the main sectors among volunteers being the construction and agriculture industry; and had little failure and game experience and computer skills. Some did not even own a computer. These participants received a laptop for the duration of the training. Although some participants (23%) were required by their organization to participate, participation was otherwise voluntary. Remuneration for participation as a thank you included a small gift certificate. Participants also received reimbursement for their travel expenses. About 5% dropped out of the training and for various reasons–predominantly a dislike regarding the game.

The purpose of the present study was to look at the overall relationship of game scores and performance. Although we acknowledge that individual and inter-group differences existed, this level of analysis will be further examined in future work. Table 2 provides descriptive statistics of key demographic information and of predictor and outcome measures.

Material

Various methods were used to investigate the impact of the training and measure the player experience and behavior [8]. The relevant material is discussed here in more detail.

Pre- and post-questionnaire

Before and after the training, participants made a self-assessment of their knowledge and attitudes toward levee inspection. The pre-questionnaire was further used to gather contextual variables, such as age and game experience, and the post-questionnaire to determine how participants judged the training. The questionnaires made use of 5-point and 7-point Likert items. Using principal components analysis questionnaire items were reduced. This resulted in the composite items pre-knowledge perception (α = .928), game attitude (α = .662), post-knowledge perception (α = .909), and judgment (α = .914). Pre-knowledge perception and post-knowledge perception reflect how knowledgeable participants perceive themselves to be about levee inspection; game attitude is about the participants disposition regarding playing games and the use of games for serious purposes; and judgment is the evaluation by participants of the game. Measures of computer and game skills are individual items.

Pre- and post-test

In addition to self-assessment, participants' knowledge was more objectively tested by letting them make sense of pictures with virtual and real failures before and after the train-

Variable	Statistic
Age, *M* (*SD*)	47.6 (12.1)
Gender, % male	97
Water Authority Employees, %	27
Failure Experience, % yes	52
Game Experience, *Mdn* (*IQR*)	1 (1–3)
Played FPS, % yes	26
Computer Skills, *Mdn* (*IQR*)	3 (2–4)
Average Game Score, *M* (*SD*)	54 (22)
Game Judgment, *M* (*SD*)	76 (15)
Pre-knowledge perception, *M* (*SD*)	54 (14)
Post-knowledge perception, *M* (*SD*)	63 (11)
Pre-test, *M* (*SD*)	24 (10)
Post-test, *M* (*SD*)	50 (17)
<g> score, *M* (*SD*)	.34 (.20)

Table 2. Descriptive statistics of key demographic information and of predictor and outcome measures. Predictor and outcome measures are presented as percentages.

ing. Open questions were used to retrieve the purest answers, which were not influenced by the researcher's wording or categorization. The responses were analyzed using content analysis. The types of answers were retrieved and the accuracy was checked. For the types of answers, the literal answers were used. Only in cases of spelling errors were corrections made. For example, if a participant answered stones and another "rocks," these were considered two separate answers. However, "stone" and "stones" were considered similar.

Accuracy of the literal answers was determined using a coding protocol and magnitude coding [19]. For accuracy protocol the logic and vocabulary of the game was used as a normative model. A score was considered *very accurate* if it was literally similar to the content of the game (e.g., "pitching stone"); *accurate* if it was closely similar to the content of the game, a synonym, or a proper alternative (e.g., "pitching rock"); *slightly* accurate if it was not necessarily wrong but descriptive or when vague language was used (e.g., "stones"); and *inaccurate* if the response was simply incorrect or too vague. Test scores were calculated by summing the scores.

Game data

Each exercise resulted in game data. This game data was saved as an XML file locally and submitted to the game server after a participant finished an exercise. This game data logged every action that players made, the coordinates where this action took place, and the time it took players to complete these actions. On average, it took participants an hour to complete an exercise, and this average was consistent across all assigned exercises. The file included the final score players achieved. The average game score was calculated by averaging across all played exercises. Participants who played fewer than four exercises were excluded, as participants tended to have lower scores in the first couple of exercises and, therefore, including these participants would skew the data.

Special Research Version

A special research version of *Levee Patroller* was developed with eight exercises. The first and last exercise were scheduled to be played at the start- and end-meeting, respectively. The other six exercises were assigned to be played at home, two per week. The exercises increased in difficulty by adding more (severe) failures. For example, the first exercise included two failures, none of which were critical. The last exercise included five failures and four of those were critical. Other considerations in designing this training program were internal validity and variety. It was made sure that players would play the exact same exercises and in the same order. Participants were unable to replay previous exercises; however, at any time they could revisit the training exercise.

The variety consideration was made to keep the training interesting. This was accomplished by varying the types of failures, the environments, and the weather. Three different environments were used (Region A, B, and C). Each region had its own characteristics, including what kind of failures could appear. In terms of weather, the difference was whether or not it would rain. If it rained, it became harder to find certain failures. It also provided for additional cognitive load on players due to the constant exposure to the rain stimuli. Playing without rain is therefore arguably easier. Table 3 provides an overview of the variations between exercises for this special research version.

Procedure

The exact setup differed per water authority. In general, however, after agreeing to participate in the study, participants would be allocated to a group of 15 to 20 participants. This was the maximum number of participants that fit into most locations where the training was held and the maximum that two facilitators were able to handle. One facilitator led the training; the other was there to assist with the training.

The training started with the start-meeting. After explaining the purpose of the training, the participants filled out the pre-questionnaire, followed by the pre-test. Then they were introduced to the game by playing the training exercise. In this exercise, all aspect of the game, from navigating to reporting, were discussed step by step. Whenever participants finished the training, they were allowed to continue with their first exercise under guidance of the facilitators. When all participants finished the training level, everyone was requested to pause their activities and the facilitators would showcase how to play the game using a projector. At the end, participants received instructions on how to play at home.

After installing the game at home, participants were able to play their second exercise. This was the only exercise available to them. After completing this exercise, the third exercise would become available and the second unavailable. This was done to make sure that everyone played in the same order and just as many times. At the end of the first and second week, participants received an e-mail with a code to unlock the next two exercises.

After three weeks the same group of people convened again for the end-meeting. This end-meeting started with playing the eighth and final exercise. When everyone was done playing, participants were requested to fill out the post-questionnaire, followed by the post-test. The meeting ended with a discussion about the training.

Exercise	1	2	3	4	5	6	7	8
Region	A	B	A	B	A	B	C	C
Failures, #	2	2	3	3	4	4	5	5
Rain	No	Yes	Yes	No	No	Yes	No	Yes
Score, *M* (*SD*)	50 (32)	58 (28)	54 (25)	53 (20)	67 (19)	61 (20)	63 (19)	61 (21)
Pre-perception, *r*	.080	.098	.244*	.106	.021	.077	.080	.089
Post-perception, *r*	.287***	.352***	.532***	.405***	.300**	.431***	.513***	.492***
Pre-test, *r*	.244**	.268**	.335***	.285**	.345***	.330**	.327**	.416***
Post-test, *r*	.394***	.417***	.611***	.613***	.588***	.655***	.613***	.770***
<g> score, *r*	.293***	.336***	.512***	.531***	.455***	.531***	.478***	.629***

Table 3. The correlations with the score per exercise, *** $p < .001$; ** $p < .01$; * $p < .05$ (two-sided).

RESULTS

In general, the game-based training can be conceived as successful [8]. A large majority (80%) played at least five out of six exercises at home, which is a participation rate that exceeded expectations. Adding all the other training activities, participants spent about two full workdays on the training. The success is confirmed when considering the measures for learning outcomes. For all comparisons of pre-test/pre-knowledge to post-test/post-knowledge, we ran paired-samples *t*-tests because of the within-subjects component of these variables. Participants' perceived knowledge on levee inspection was significantly higher after the training, $t(115) = 8.64$, $p < .001$. In addition, participants' test scores were significantly higher, $t(124) = 19.21$, $p < .001$. Additionally, we calculate <g> scores [12] for each participant, a score used to assess learning between pre- and post-tests. Results suggest that on average participants received 34% of the possible score increase on the post-test, suggesting the game as an educational tool was successful. Much to our surprise, no differences were found on how participants performed on responding to real versus virtual pictures, on the pre-test as well as the post-test. This suggests that players are able to transfer the knowledge and skills they gained from the game and apply this to other contexts. In the remainder of this results section, we discuss the results in the context of game performance, which is the focus of this paper.

Game Score

The first step was to consider the overall (average) game score players gained across all exercises. The average was chosen giving consideration to the variations among the exercises. Overall game score was positively correlated moderately with the pre-test, $r = .40$, $p < .001$; strongly with post-knowledge perception; $r = .54$, $p < .001$; and also strongly with the post-test, $r = .71$, $p < .001$. No correlation was found with the pre-knowledge perception. The moderate relationship with the pre-test may suggest that prior knowledge helps to perform well in the game; however, this idea is not supported when considering perceived knowledge.

The use of self-assessment, such as the knowledge perception measure, and its relationship with test scores has been a debate for decades [4, 25]. In general, moderate correlations are to be expected (i.e., Cohen's criteria are used, where .50 is the cut-off between medium and large effect sizes [5]). In case of *Levee Patroller*, the pre-knowledge perception indeed relates moderately to the pre-test, $r = .40$, $p < .001$. Post-training has a similar relationship with post-knowledge perception $r = .40$, $p < .001$. This consistency aligns with the literature.

With this in mind, the strong relationship between the average game score and the test score was interesting. Considering that the game scores were achieved before participants filled out the post-test, it suggests that the game score was a good predictor of test performance. This would support the argument that game scores could fulfill a role in assessmentif designed well. In fact, game scores have a strong correlation with perceived knowledge as opposed to the medium correlation between this variable and the test score. This relationship may have been influenced by the feedback in the game. Sitzmann et al. [25] state that "If learners receive feedback on their performance, they should modify their self-assessments to be more aligned with their actual knowledge levels" (p. 172). Therefore, the feedback the game provided may have influenced their sense of knowledge about levee inspection. This is an important aspect of gamifying learning because it provides an interesting approach to not only informing educators/employers about the knowledge and skills of the individual, but the individual is better able to assess what they do or do not know. Having a better understanding of one's own skills and limitations allows the learner to better focus their efforts in their education.

Interestingly enough, game scores only appear to correlate moderately with how participants appreciate the game, $r = .35$, $p < .001$. This is an important finding. Having fun and being engaged with the game does relate to better game performance; however, the moderate relationship suggests that these highly valued "player experiences" are not necessary to adequately gauge understanding.

Based on this analysis, we can conclude that Hypothesis 1 is supported. The score is positively related to the results on the learning outcomes. To understand the role of scores better, we investigated them as well per exercise.

Per exercise

A problem with using the average game score is that it does not give a good sense on how performance develops over time. However, because the region layouts were different and each level contained different failures and weather conditions (a necessary design implementation to keep players engaged), we did not expect comparisons in scores across levels to be meaningful. For example, Region B contained a maze of levees as opposed to Regions B and C, making it more difficult

to navigate the environment and to inspect each levee. On levels with rain, it would be more difficult to assess failure signals. A significant non-parametric Friedman test, $\chi^2(7) = 46.06$, $p < .01$, reflected a difference in game scores between levels. However, a non-significant effect of trial on game score ($p > .05$) was observed, which suggests that although there were differences observed between levels, the differences could not be accounted for by learning alone. For this reason, we considered the scores per exercise and how they relate to the main learning outcome measures. Table 2 provides an overview of the results.

The results of the Friedman test suggest there were significant differences between exercises. A peak happened on the fifth exercise. A possible explanation is that it is easier to find failures in Region A as opposed to Region B. Participants confirmed this idea and complained in particular about getting lost in Region B. In the last two exercises players were introduced to an entirely new region and had to deal with five failures. Participants reported to struggle with this increase in failures. We would expect that if a ninth exercise were added, the average scores would be equal or higher to those in the fifth exercise. At that point, the participants would have been acclimatized to the new region.

A smaller peak happened in the second exercise, which was in the arguably more difficult Region B. A possible explanation for this peak is that the facilitators used this scenario as an illustration during the start-meeting of how to play the game, after all participants concluded the training level. This was done on purpose to make sure that the participants would be comfortable in playing their first exercise at home. The participants were not informed about this and they also did not get to see the exercise from the beginning to the end. Nevertheless, this may have influenced the results.

When considering how the scores per exercise related to the learning outcome measures, we found interesting patterns. First, except for the third exercise, no exercise score related to pre-knowledge perception. The exception with the third exercise may be explained by that all players first had to learn how to play and their scores may not have reflected their actual knowledge. With "learning how to play" we refer to getting used to the control scheme of how to navigate the environment, knowing how to deal with failures using the tools available to the player, etc. With this argument, however, we would need to adjust our idea of how performance improved over time. Only a subtle improvement may have occurred.

The correlations with post-knowledge perception and the pretest fluctuate per exercise, making it hard to draw any conclusions. Both seem to increase and stabilize after the second exercise. This provides possible evidence that participants needed at least the first two exercises to learn how to play the game. The correlation with the post-test, on the other hand, provides an intriguing pattern. The strength of the relationship appears to increase over time, with a clear peak at the final exercise, $r = .77$, $p < .001$. This tells us that the player performance at the final exercise is a good measurement of learning, even better than the average score.

Per learning objective

The game scores were designed using the learning objectives that were defined in consultation with the users (see Table 1). The questions for each picture on the tests were also designed using these objectives. Therefore, a possibility exists to compare how performance in the game on the specific learning objectives relates to the items on the test that aim to measure learning on each of these objectives (e.g., how does the game score on reporting accuracy relate to the accuracy in reporting pictures on the test?). For the game performance we chose the players average score on each learning objective; for the test we summed the accuracy on each specific question across all the pictures. In this analysis, we are only considering the post-test results. The learning objective "Taking measures" has been excluded from this analysis. Participants only had to respond to this question if they indicated that the depicted failure was critical. Due to this, there are far too few answers to analyze this relationship.

For observing failures, two measures were implemented in the game: finding the failures and reporting the signals. There was a significant positive relationship between noting what failure was depicted on the post-test and finding the failures in the game, $r = .38$, $p < .001$; as well as for identifying the signals in the game, $r = .41$, $p < .001$. For reporting failures, two measures have been implemented as well and were significantly correlated with post-test scores for reporting: reporting the location, $r = .59$, $p < .001$; and creating a report associated with each signal, $r = .51$, $p < .001$. These are the core responsibilities of patrollers. Improvements are especially desired in these areas. The relationships may not have been as strong because many participants had trouble with the interface in reporting signals and reports. This issue was so prevalent that small changes were made in the interface during the training to prevent players from making more errors.

For the learning objectives assessing and diagnosing, the implementation in the game is far more straightforward. With diagnosing this is reflected in the correlation between the in-game and post-test measure for diagnosing, $r = .72$, $p < .001$. For assessing, however, the relationship between in-game and post-test scores was not as strong, $r = .34$, $p < .001$. Considering that participants had three options in assessing failures (i.e., reportable, severe, and critical), this relationship could be considered random. An explanation for this relationship is that participants had trouble with assessing failures. The game did not provide enough heuristics to help them. This problem was not necessarily due to the design. In reality this ambiguity exists and there are no clear guidelines on how to judge the severity. In a session with experts, hardly any consensus was reached on how to assess any of the pictures.

In addition to the score per exercise, the consideration per learning objective has revealed that the role of game performance is complex. We are now considering possible influential factors in our analysis, such as skills and age.

Prior Skills

We hypothesized that a few variables may influence performance in a game environment, in particular computer and game experience. This possible influence is of importance in

considering the use of computer environments such as games for assessment. The results may not reflect actual knowledge and skills, simply because the user is lacking in these areas. On the pre-questionnaire, participants were asked to indicate their computer skills on a 5-point Likert item (from not skilled to very skilled). For *Levee Patroller* it turns out that computer skills relates uniquely with the average game score, $r = .57$, $p < .001$. Possible moderating/mediating variables [2] such as game skills and age were controlled for using partial correlations. Therefore, computer skills had a strong relationship with performance, which supports Hypothesis 2.1.

On the same pre-questionnaire, participants were asked how often they played digital games on a 5-point Likert item (from rarely to daily). The amount of gameplay was used as a proxy for measuring game experience, assuming that if people play more, they will have acquired more skills in how to play games. They were also asked if they ever played a first-person shooter (FPS; Yes or No). This question was asked because *Levee Patroller* uses the same perspective and control scheme as this game genre. Analyses show that the relationships between performance and both measures were mediated by computer skills, hence Hypothesis 2.2 was not supported.

It should be kept in mind that this population sample has not been exposed much to games. About 53% indicated to rarely play games and 23% mentioned to have ever played a FPS. Similarly, the variance in computer skills may be more extreme in this population sample than in others, due to the large variety of backgrounds of participants, with some extremes, such as participants who did not ever own a computer. Nevertheless, based on these results we should conclude that performance is likely not a function of game experience, but rather a function of computer skills.

Age

Using partial correlations, controlling for factors such as computer skills, we found that age has a similar but reverse correlation with game performance, $r = -.57$, $p < .001$. This suggests that there is something other than the skills to control the game for why older people gain less from the game. Research has suggested that age differences may due to a decrease in cognitive functioning as we age and games can actually be used to counter this effect [1]. With *Levee Patroller*, we did not observe this trend. This suggests that in developing games for a wide age range, the cognitive functioning of older participants needs to be considered, see also [9]. With these results, however, we found support for Hypothesis 3: younger participants do achieve higher game scores compared to older participants, even after controlling for factors such as computer skills and experience playing games.

DISCUSSION

The premise of this paper was to explore the role of game performance through the results of a game called *Levee Patroller*. The results show promise for using game scores as an assessment measure, and therefore the potential of games as a tool for assessment of learning, behavior, or anything else that can be measured using games [21]. In fact, assuming test scores are more objective than self-assessment, the game score serves as a more accurate predictor than self-assessment. A few topics are worth discussing for future research, including the validity of the results, population sample, design of the training, and design of learning objectives.

Validity

Our results are largely based on a comparison between the game scores and the results on a test that had not been standardized and validated. This begs the question to what extent the results are empirically valid. First, it is important to discuss the game scores. In *Levee Patroller* the scoring is contingent on whether players find the failures. Only when a failure is found, players can get scores on any of the criteria for that failure. As one participant noted in one of the first exercises, "Found nothing, learning nothing." Finding the failures is a difficult task (especially in Region B) and therefore not trivial. This means that the score is to some extent based on mere luck. The timing of finding the failures is important too. If players find a failure at a later stage of the game, they cannot get any points for observing this failure in its earlier phase(s).

However, it can be expected that this element of luck will decrease over time. Players will know where to look for failures and how they deal with failures efficiently. With this in mind, we can argue that this is actually part of the learning process. Knowing where to find failures is a skill. For this reason, the element of luck might actually just reflect the learning objectives. Nevertheless, it could be that some of the variation that is not explained by the game score is due to the element of luck in the game. Considering that luck and chance play an important role in many games, this issue should be kept in mind for using game scores as a performance measure.

The test has been tailor-made for this particular study. This was out of necessity and its design is a common practice for assessing software tools [3]. One significant difference is the use of open questions and manual coding to derive the test scores. This is arguably more error prone compared to the use of, for example, multiple-choice questions. On the other hand, through this means we retrieved more "honest" answers, answers that were written by the participants themselves. In addition, the coding process involved using the literal answers and a strictly defined coding protocol based on the normative model from the game, which was co-developed with subject-matter experts. The coding between raters was almost identical, further proofing that the coding process was unambiguous. Therefore, it could be argued that the derived test scores reflect a participant's knowledge level better than, for example, a test with just multiple-choice questions.

Population Sample

The population sample used in this training is very specific. It concerned an audience with very little computer skills, and as it turns out, these skills have a major impact on what players gain from playing. This begs the question whether these kinds of environments should be used for audiences with little to no computer skills. In our opinion, they should be used. The training was overall a success, also for participants with difficulties. Nevertheless, more attention is needed in the design

for people with little computer skills, just as much attention is needed to accommodate for older players.

Fun and Engagement

One interesting key aspect for further investigation is the role of game judgment. In the case of *Levee Patroller*, it turned out that a relatively weak relationship exists between judgment and game scores. Game scores are strongly related to the learning outcomes and therefore fun and engagement seem less important than game designers often argue for. A common design philosophy among game designers who develop games for impact is "fun first" [17]. Our results speak against this. A recent study confirmed that fun is not a good predictor of learning success in serious games [10].

What may explain this is to distinguish between the concepts of engagement and motivation [7]. It can be motivating to learn something more efficiently and better with the help of an educational game [10]. Many levee patrollers had never experienced or even seen the content in the game, and therefore, they were intrinsically motivated to play this game. In fact, one volunteer was 65 years old, volunteered to be a levee patroller since he was 18 years old, and mentioned that he never experienced a failure in his life. He and others did not need to be engaged to learn.

Of course, these results stem from a target group who has little affinity with games to begin with. It should be pointed out that the game still had usability flaws at the time of deployment, which clearly led to a dissatisfied experience for some [8]. Also, the importance of fun likely depends on the context in which games are used. Future research should further address the relationship between fun and scores.

Training Design

Upon dissecting the initial results of the strong predictive value of game scores, it turns out that the role of game performance is quite complex. Many design decisions, in the design of the game and in the design of the training, play a role in what kind of game performance is achieved and its meaning. In the future designers/scholars should keep this in mind and possibly create cleaner designs that will help to achieve more robust measures of the usefulness of game performance as a measure of assessment. Most likely trade-offs will have to be made. In the design of the game-based training with *Levee Patroller*, one of the explicit requirements was to include variety. This variety is what may have confounded the results.

It should be noted that dissecting the gameplay results was a useful exercise in understanding the impact of the training. For example, through this it became clear that the game might not teach anything about assessment. These insights, in turn, can help to improve the training. This suggests that the use of game data is not only useful for assessing the player; it is also useful for assessing the game itself. With new methods and techniques, more possibilities will become available for analyzing game data, see [20].

In designing a training, designers should further keep in mind that players first need to "learn how to play before they can learn from the game." For the first couple of tries, players' game scores may not reflect their knowledge at all.

Learning Objectives Design

The case of *Levee Patroller* was explored to look into the role of game performance because its game goals are directly coupled to the learning objectives. The hypothesis was that due to this direct link, the game scores would potentially be a good measure of assessing learning. In this paper, we have not studied what happens if the game scores are not aligned with the learning objectives. The logical hypothesis would be that in that case the game score is not suitable as a measure of assessment. Although further research would need to clarify under what circumstances the game score is suitable, this research highlights that if the learning objectives are aligned, it can be a measure of assessment.

To design game scores that predict actual performance, designers/scholars will first need to identify the learning objectives and then use these to design the activities in the game environment as well as the criteria for determining how players performed. To assess player performance, a normative model is needed. The complications of assessing levee failures illustrates the difficulty in establishing such a model.

Limitations

We acknowledge that this study is limited, largely because we are discussing the use of scores in one game created for a very specific audience; however, any game is limited by its design and the context of its use. Because of the specific nature of *Levee Patroller*, results from this study may not transfer to other contexts or to other types of learning. However, we believe that this approach for designing learning-centered game scores is transferable and believe that providing an in-depth description for how this was done in *Levee Patroller* will allow future researchers to apply a similar framework to other games and, in doing so, validate the methodology. Because our work is closely related to the ECD framework [15], which a few scholars have implemented in the context of games [14, 22, 24], and is supported by related work on IM and SE, we have reason to believe that our approach is content-agnostic.

CONCLUSION

This research shows that there can be a goal to scoring in games. If designed well, by aligning the learning objectives with the game goals, game performance can play a role to assess player learning. These findings can extend beyond the use of educational games. Games can be used to assess practically anything, from a person's personality to the use of software tools [7]. In addition, the results have implications for the design of gamification, as we provide a framework for a meaningful implementation, one that stimulates intrinsic motivation. Future research should illustrate how the design of game goals can be most effectively done in these other contexts for assessment purposes.

In addition to this main finding, which will help foster the recent movement of using game-based assessment, this research has three contributions. First, it specifies a design approach for creating learning-centered game scores that are

predictive of player learning. Second, it details a validation approach using a unique mixed-methods study where players play at home. Third, it highlights how game data and other methods can be used to analyze game scores. Specifically, this research shows the importance of dissecting player behavior using game data. This process will reveal insights that may otherwise go unnoticed.

ACKNOWLEDGMENTS

We would like to thank Deltares and TU Delft for their support, and in particular the game teams at both institutes. Our gratitude extends to the participating water authorities and the patrollers who were willing to participate in this study.

REFERENCES

1. Anguera, J., Boccanfuso, J., Rintoul, J., Al-Hashimi, O., Faraji, F., Janowich, J., Kong, E., Larraburo, Y., Rolle, C., Johnston, E., and others. Video game training enhances cognitive control in older adults. *Nature 501*, 7465 (2013), 97–101.
2. Baron, R. M., and Kenny, D. A. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology 51* (1986), 1173–1182.
3. Brennan, K., and Resnick, M. New frameworks for studying and assessing the development of computational thinking. In *Proceedings of AERA* (Vancouver, Canada, 2012).
4. Brown, K. G., Sitzmann, T., and Bauer, K. N. Self-assessment one more time: With gratitude and an eye toward the future. *Academy of Management Learning & Education 9*, 2 (2010), 348–352.
5. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2 ed. Lawrence Erlbaum, Hillsdale, NJ, 1988.
6. Fishman, B. J., and Deterding, S. C. Beyond badges & points: Gameful assessment systems for engagement in formal education. In *Proceedings of Games+Learning+Society* (Madison, WI, 2013).
7. Harteveld, C. *Triadic game design: Balancing reality, meaning and play*. Springer, London, UK, 2011.
8. Harteveld, C. *Making sense of virtual risks: A quasi-experimental investigation into game-based training*. IOS Press, Amsterdam, the Netherlands, 2012.
9. IJsselsteijn, W., Nap, H. H., de Kort, Y., and Poels, K. Digital game design for elderly users. In *Proceedings of the 2007 conference on Future Play*, Future Play '07, ACM (New York, NY, USA, 2007), 17–22.
10. Iten, N., and Petko, D. Learning with serious games: Is fun playing the game a predictor of learning success? *British Journal of Educational Technology* (2014).
11. Kapp, K. M. *The gamification of learning and instruction: game-based methods and strategies for training and education*. John Wiley & Sons, 2012.
12. Mayo, M. J. Games for science and engineering education. *Communications of the ACM 50*, 7 (2007), 30–35.
13. Meister, J., Ed. *Pillars of e-learning success*. Corporate University Exchange, New York, NY, 2002.
14. Mislevy, R., Oranje, A., Bauer, M., Davier, V., Hao, J., Corrigan, S., Hoffman, E., DiCerbo, K., and John, M. Psychometric considerations in game-based assessment. *GlassLab Report* (2014).
15. Mislevy, R. J. Evidence and inference in educational assessment. *Psychometrika 59*, 4 (1994), 439–483.
16. Nicholson, S. A user-centered theoretical framework for meaningful gamification. In *Proceedings of Games+Learning+Society* (Madison, WI, 2012).
17. Prensky, M. *Digital Game-Based Learning*. McGraw-Hill, New York, NY, 2001.
18. Ryan, R. M., and Deci, E. L. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist 55*, 1 (2000), 68.
19. Saldaña, J. *The coding manual for qualitative researchers*. No. 14. Sage, Thousand Oaks, CA, 2012.
20. Seif El-Nasr, M., Drachen, A., and Canossa, A., Eds. *Game analytics: Maximizing the Value of Player Data*. Springer, London, UK, 2013.
21. Shaffer, D. W., and Gee, J. P. The right kind of GATE: Computer games and the future of assessment. In *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research*. Information Age Publishing, Charlotte, NC, 2012.
22. Shaffer, D. W., Hatfield, D., Svarovsky, G. N., Nash, P., Nulty, A., Bagley, E., Frank, K., Rupp, A. A., and Mislevy, R. Epistemic network analysis: A prototype for 21st-century assessment of learning. *International Journal of Learning and Media 1*, 2 (2009), 33–53.
23. Shute, V., and Ventura, M. *Stealth assessment: Measuring and supporting learning in video games*. MIT Press, 2013.
24. Shute, V. J. Stealth assessment in computer-based games to support learning. *Computer games and instruction 55*, 2 (2011), 503–524.
25. Sitzmann, T., Ely, K., Brown, K. G., and Bauer, K. N. Self-assessment of knowledge: A cognitive learning or affective measure? *Academy of Management Learning & Education 9*, 2 (2010), 169–191.
26. Takatalo, J., Häkkinen, J., Kaistinen, J., and Nyman, G. User experience in digital games: Differences between laboratory and home. *Simulation & Gaming 42*, 5 (2011), 656–673.
27. Washbush, J., and Gosen, J. An exploration of game-derived learning in total enterprise simulations. *Simulation & Gaming 32*, 3 (2001), 281–296.
28. Zimmerman, B. J. Self-efficacy: An essential motive to learn. *Contemporary educational psychology 25*, 1 (2000), 82–91.