Performance assessment in serious games: Compensating for the effects of randomness

Westera, W. (2014). Performance assessment in serious games: Compensating for the effects of randomness. Education and Information Technologies. Online version http://link.springer.com/article/10.1007/s10639-014-9347-3. DOI 10.1007/s10639-014-9347-3

This paper is about performance assessment in serious games. We conceive serious gaming as a process of player-lead decision taking. Starting from combinatorics and item-response theory we provide an analytical model that makes explicit to what extent observed player performances (decisions) are blurred by chance processes (guessing behaviors). We found large effects both theoretically and practically. In two existing serious games random guess scores were found to explain up to 41% of total scores. Monte Carlo simulation of random game play confirmed the substantial impact of randomness on performance. For valid performance assessments, be it in-game or post-game, the effects of randomness should be included to produce re-calibrated scores that can reasonably be interpreted as the players´ achievements.

Keywords: serious gaming; performance assessment; cut score; computer-assisted learning; monte carlo

Introduction

For many decades the engaging properties of games have been used for learning and other serious purposes (Abt 1970). These so-called serious games cover a wide range of domains, objectives, approaches and styles as to meet specific educational requirements and audiences. Because learning is their primary purpose a critical element of serious games is the assessment of learning achievements (e.g. Chin, Dukes, & Gamson 2009; Bellotti, Kapralos, Lee, Moreno-Ger, & Berta 2013; Connolly, Boyle, MacArthur, Hainey, & Boyle 2012; Shute, Ventura, Bauer, & Zapata-Rivera 2009). Our paper provides a methodology for analysing to what extent random user behaviours (guessing rather than thoughtfully deciding) affect the validity of in-game assessment data. It uses a statistical approach for making explicit to what extent the observed player performances are blurred by chance processes and explains how to compensate for this.

In many cases the assessment of game sessions is arranged as post-session summative test, interviews or questionnaires, which covers the learners' overall achievements. Generally, however, games include measures for assessing the learner's progress and failures, which are used for score assignment, level transitions, feedback and adaptation of the game play (Shute et al. 2012; Boston 2002). This fits in the trend toward formative assessment. Redeker, Punie and Ferrari (2012) describe the stepwise development from 1st generation testing in the 1990s (automated administration and scoring of conventional tests) and 2nd generation testing in the 2000s (adaptive summative testing) to 3rd generation testing from 2010 (continuous, unobtrusive, monitoring and formative assessment). Obviously, educational measurement is shifting from large numbers of students with only one observation toward an individualised approach with a large number of observations.

In-game assessment is highly relevant for serious games, because the learning-by-doing approach they generally implement and the freedom of movement that goes with it, may easily affect the effectiveness of learning. Since the instructional control in game environments is very limited, players may easily manoeuvre themselves into positions that are unfavourable for successful and efficient learning. Learning-by-doing means learning from the experiences that result directly from one's own actions, as contrasted with learning from watching others perform, reading others' instructions or descriptions, or listening to others' instructions or lectures (Reese 2011). Learning-by-doing activates learners and helps them to acquire the tacit knowledge that is intrinsically bound to the actions performed. It includes practice, discovery, inquiry, problem solving, and authentic contextual knowledge to achieve learning goals (Schank, Berman, & Macpherson 1999; Aldrich 2005; Schank 1995). Unfortunately, just doing things and having the associated experiences are not a sufficient condition for learning, because 1) doing a task may be too difficult (e.g. playing a piece of Rachmaninoff), 2) learning to do things may require doing things that don't look like the final task at all (e.g. practicing musical scales) and 3) just doing things does not necessarily lead to deep cognitive processing and the associated insights and understandings. With respect to the latter issue, research into computer-assisted instruction and simulations has shown to favour trial-and-error learning strategies that involve a lot of doing, but lack any thoughtful analysis of experiences (Vargas 1986). Likewise, games foster the tendency to act before thinking. Especially, game interactions that put little cognitive load on the users, such as interaction by direct manipulation with graphical objects, tend to induce a more implicit, trial and error learning mode (Guttormsen Schär et al. 2000). Game design patterns that induce stress, such as a time lock, time pressure or time-dependent scores are likely to promote hurried, shallow or incomplete processing. Scholars such as Schön (1983) and Kolb (1984) realised that just having the experience is not a sufficient condition for learning, but should be complemented with a thoughtful review process.

The core assertion of this paper is that in-game assessment should take into account the effects of random, thoughtless gaming behaviours and compensate for it. Our research questions are stated as follows:

(1) How can we formally describe the effects of random game play on the player´s performance score?
(2) What is the magnitude of the effects of random game play?
(3) What is the impact of random game play in practice?

First we will explain the mechanism of player-led decision taking that a wide variety of serious games are based on. Next we will provide an analytical description of decision taking that takes into account the influence of random choices and we will connect this to performance scores. We will develop an analytical model that describes the impact of randomness and we will investigate the practical impact by applying the formalism to two existing serious games. We conclude our study by discussing the outcomes.

Serious game play as a process of active decision taking

A wide range of serious games have demonstrated to provoke active learner involvement through exploration, experimentation, competition and co-operation

(Westera 2008). Aldrich (2005) distinguishes between four basic approaches to serious games:

(1) Branching stories allow the player to choose their own path through the game;
(2) Interactive spread sheets offer numerical simulations that can be manipulated by the player;
(3) Game-based models, which are derived from established entertainment games, e.g. TV-quiz formats;
(4) Virtual labs, which offer 3D environments and objects for risk-free experimentation.

Aldrich (2005) explains that in practice the boundaries between the approaches are blurred: many games combine branching mechanisms, simulations, popular formats and 3D representations. In all cases game play involves active decision taking by the players, who are challenged to achieve favourable outcomes and maximise their performances in the game. Whatever game approach is chosen, player-led decision taking is the very basis of game-based learning. It goes with active involvement, freedom of movement, problem ownership, adopting a certain role and responsibility, and the empowerment to change the game´s state (Westera 2008). Conceptually, however, playing a game is not very different from taking a multiple choice test, be it that the game style and context may easily conceal the underlying multiple choice nature. The multiple choice pattern holds for decision nodes in a branching story, the parameter selection and value assignment in simulations (e.g. buying supplies), the answering of quiz questions, as well as the experimentation design and actions in a virtual lab. More openly, many games comprise built-in multiple choice items for establishing progress or other purposes (Becker & Parker 2011). Shute et al. (2009), however, advocate stealth, unobtrusive assessment in games, because this avoids the interruption of game play which is known to negatively affect the learning process (Bente & Breuer 2009). Whatever model is favoured, player-led decision-taking is a predominant feature of serious games. Players´ decisions are the basis of progress monitoring, adaptive game play, performance assessment and providing feedback to the players, either expressed as a score, a badge, a privilege, an achievement or any other performance qualification.

Formal description of decision taking

A multiple-choice (MC) item is a closed question composed of a premise or lead question (stem) followed by a list of possible answers (alternatives) to be selected by the candidate. In its basic form the MC item is a single-answer question (single select list): only one of the alternatives is the correct answer, the other alternatives are wrong. A special case of a single-answer item is a True-False item or Yes-No item, which has only two alternatives. For interpreting the score obtained from MC questions we have to realise that selecting the right answers may be partly due to chance. The Random Guess Score (RGS) is the probability that a randomly selected answer is the right answer (Ebel & Frisbie 1991). The RGS of an MC item with m alternatives is 1/m. It means that providing the right answer is not necessarily the result of having the right knowledge, but can be a matter of being lucky. Hence, chance processes blur the significance of the observed scores. In test construction this is generally addressed by calculating a minimum score required for passing the test (the cut score, or pass mark), which takes into account the effects of chance.

As opposed to a single-answer question a multiple-answer question (multiple-response question or multi-select list) allows the candidate to select more than one answer. The chance of being successful by randomly selecting answers now depends on the boundary conditions that apply. Consider a multiple-answer question composed of m alternatives, k of which are correct answers (k≤m). In the easiest case the stem text makes explicit that the candidate has to select exactly k correct answers. Under this constraint the multiple-answer question reflects the single action of selecting k correct alternatives from a list of m alternatives. This is technically equivalent with a single-answer question that offers m' alternatives, where m' is given by the binomial coefficient ("m choose k"):

$$m' = \binom{m}{k} \tag{1}$$

A silent assumption made here is that a positive score is only assigned when exactly all k answers are selected (dichotomous scoring). If the candidate has selected too few correct answers (<k) or has given some wrong answers, no score is assigned.

If the value of k is not made explicit beforehand the chance of selecting the right alternatives by guessing is much lower, because there are more combinations of alternatives to choose from. In such cases the formula for converting a multiple-answer question with m alternatives to a single-answer question of m" alternatives is now given by:

$$m'' = \sum_{k'=1}^{m} \binom{m}{k'} \tag{2}$$

The number m" covers all possible selections of available alternatives and their combinations. Because of this technical equivalence of single-answer questions and multiple-answer questions we may restrict our analysis to the case of single-answer items.

Describing scores in sets of MC questions

Decision taking in (serious) games can be interpreted as a series of MC items to be addressed. For evaluating the performance score of a player it is needed to devise a scoring system that takes into account the weight and complexity of the items. Consider a game that can be described as a set of n single-answer MC items. Let m_i denote the number of alternatives of item i. Let x_i denote the correctness of the player's answer of item i, by setting $x_i=1$ in case of a correct answer and $x_i=0$ if the answer is wrong. In addition, let w_i be a score weight function, which takes into account the importance or complexity of item i. Then the total weighted score S achieved by the player is given by:

$$S = \sum_{i=1}^{n} x_i \cdot w_i \tag{3}$$

Obviously the minimum score (all $x_i=0$) is 0; the maximum score (to be obtained if all $x_i=1$) is given by:

$$S_{max} = \sum_{i=1}^{n} w_i \qquad (4)$$

In order to make fair judgements about the cut score, which is the minimum score required for passing the test, we have to take into account the score level that is obtained by simply answering the questions randomly (the random guess score: RGS). The RGS sets a lower bound to the performance that can reasonably be attributed to the candidate's capabilities. A normalized score S_N representing the candidate's performance, corrected for performance by chance, should thus be expressed as

$$S_N = \frac{S - RGS}{S_{max} - RGS} \qquad (5)$$

Calibrating the cut score as a 50% normalized score yields the following expression for the cut score S_c:

$$S_c = 0.50 * (S_{max} - RGS) + RGS \qquad (6)$$

Expressed as the normalised cut score, which is the cut score S_c relative to the maximum score S_{max}, we obtain:

$$S_{c,norm} = S_c/S_{max} = 0.50 * (1 + RGS/S_{max}) \qquad (7)$$

For the calculation of the RGS and the cut score we will distinguish between three different cases.

4.1. Case 1: No item completion required

In this case we assume that an item needs not be completed and corrected before continuing with the next one. Since the chance of guessing the right answer of item i is $1/m_i$, the RGS is found by replacing the score x_i with $1/m_i$ in equation (3):

$$RGS = \sum_{i=1}^{n} \frac{w_i}{m_i} \qquad (8)$$

The normalised score, the cut score and the normalised cut score can now be calculated with equations (5)-(7).

4.2. Case 2: Item completion required (with replacement)

In this case the candidate has to pass each item correctly by adjusting the selection of alternatives until the right answer is given. Such situation is quite exemplary for decision taking in serious games, for instance when closures are removed after completing a challenge or a level successfully. Here performance score is not based on the answers given, but on the number of trials (y) needed to select the right answers. We need to define a scoring function S(y), which translates observed behaviors (number of trials needed) into judgements of performances. Although different value systems may be used for defining the score, various constraints apply: for instance the score function should be representative, credible, plausible, simple and understandable and it should not conflict with common sense. Mathematical requirements include monotonousness, transitivity, proportionality and homogeneity (multiplicative scaling). Clear boundary conditions arise from the maximum score, which is assigned when y=1, and the minimum score, which approaches to zero for large values of y. For the aim of our study the expression of the score function is not critical: any score function that complies with the above-mentioned criteria would do. In its most simple form the player's score S(y) would be expressed as a reciprocal function of y:

$$S(y) = \frac{w}{y} \tag{9}$$

This formula matches all criteria explained above. Similar to the non-completion case described before, we can now determine the influence of randomness.

For calculating the RGS we need to link the score function S(y) of the item to the probability of being successful after y trials. The process of randomly selecting alternatives in a single-answer MC item is a repeated Bernoulli trial. In case a player adopts a strategy of just randomly selecting an answer, without remembering wrong answers, the repeated Bernoulli trial is an experiment "with replacement", which means that the wrong answers remain part of the set of alternatives that the player chooses from. The trials are independent and can be described by the binomial distribution. If the item has m alternatives the probability of having one correct answer in y trials (y>0) is given by:

$$P(1, y, \frac{1}{m}) = \binom{y}{1}\left(\frac{1}{m}\right)\left(1 - \frac{1}{m}\right)^{y-1} \tag{10}$$

However, we're not interested in the probability of having a correct answer in y trials, but in the probability P' of having a correct answer exactly in the yth trial and not in the previous ones. Since all draws have the same probability this means that we need to divide by the number of permutations of the y trials. This reduces equation (10) to

$$P'\left(1, y, \frac{1}{m}\right) = \left(\frac{1}{m}\right)\left(1 - \frac{1}{m}\right)^{y-1} \tag{11}$$

Equation (11) indicates the probability that the y trials that the candidate needed to answer the MC question correctly, were just a matter of chance. The RGS, which is the expectation value of the score S for the item, given the probability distribution P´, is now given by:

$$RGS = \frac{\sum_{y=1}^{\infty} S(y) \cdot P'(1, y, \frac{1}{m})}{\sum_{y=1}^{\infty} P'(1, y, \frac{1}{m})} \tag{12}$$

Since the denominator is one this can be rewritten with equations (9) and (11) as

$$RGS = \frac{w}{m} \cdot \sum_{y=1}^{\infty} \frac{1}{y} \cdot \left(1 - \frac{1}{m}\right)^{y-1} \qquad (13)$$

After summation over all items equations (5)-(7) can be used for calculating the normalised score, the cut score and the normalised cut score, respectively.

4.3. Case 3: Item completion required (without replacement)

In this case the player likewise has to pass each item correctly by adjusting the selection of alternatives until the right answer is given. Now, however, the player is supposed to learn from mistakes by remembering wrong answers given. Note that such learning strategy is not necessary related to learning content: the player could still be thoughtlessly taking decisions, while just taking into account what alternatives to exclude the next turn. Such strategy produces a repeated Bernoulli trial "without replacement", which means that the wrong answers are no longer considered to be part of the set of alternatives that the player chooses from. Consequently the trials are no longer independent and the binomial distribution no longer applies. It has to be replaced with the hypergeometric distribution, which excludes replacement. The probability of having one correct answer in y trials (y>0) without replacement is now given by:

$$P\left(1, y, \frac{1}{m}\right) = \frac{\binom{m-1}{y-1}}{\binom{m}{y}} \qquad (14)$$

The probability P' of having a correct answer exactly in the yth trial and not in the previous ones follows from induction:

$$P'\left(1, y, \frac{1}{m}\right) = P\left(1, y, \frac{1}{m}\right) - P\left(1, y-1, \frac{1}{m}\right) \qquad (15)$$

This is valid because the chance of being successful in exactly the yth trial is equal to the probability of being successful in the first y trials, minus the probability of being successful in the first (y-1) trials. Elaboration of equation (15) leads to a simple expression of P' that is independent of y:

$$P'\left(1, y, \frac{1}{m}\right) = \frac{1}{m} \qquad (16)$$

This simple outcome is the result of the fact that the probabilities of drawing different sequences of trials with one success are equal (exchangeable sequences). The outcome can also be understood as follows: the probability of being successful in the first trial is 1/m; being successful in the second trial requires failure in the first trial, which has probability (m-1)/m, and success in the second trial, which has probability 1/(m-1), rightly excluding the option chosen in the previous trial. The product of these probabilities yields 1/m. The same procedure holds for subsequent trials, all yielding a probability of 1/m.

Although the probability may remain constant for each turn, the assigned performance S(y) given by equation (9) goes down with each turn. The RGS, which is the expectation value of the score S, given the probability distribution P´, is now given by:

$$RGS = \frac{\sum_{y=1}^{m} S(y) \cdot P'(1, y, \frac{1}{m})}{\sum_{y=1}^{m} P'(1, y, \frac{1}{m})} \tag{17}$$

With equations (9) and (16) this can be rewritten as

$$RGS = \frac{w}{m} \sum_{y=1}^{m} \frac{1}{y} \tag{18}$$

After summation over all items, equations (5)-(7) yield the normalised score, the cut score and the normalised cut score.

Calculated impact and comparison

We will now analyse the impact of the approach on the RGS and the performance cut score, and make comparisons between the three different cases. For calculating the RGS and the cut score in a variety of cases we have implemented the analytical models in a SCILAB computer program (http://www.scilab.org).

For reasons of simplicity we start our analysis with considering a one-item test with m alternatives and weight factor 1. Figure 1 shows for all three cases how the RGS of a single-answer item varies with the number of alternatives m.

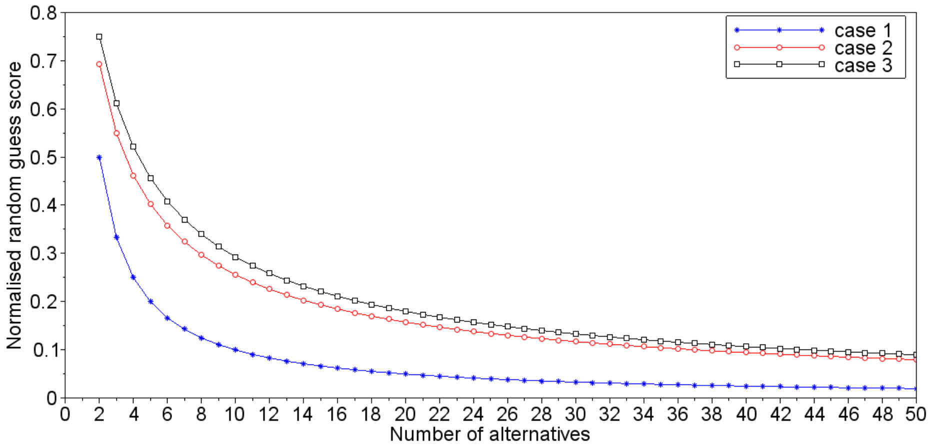


Figure 1 Normalised random guess score versus the number of alternatives for case 1, case 2, and case 3, respectively

As can be derived from equation (8) for case 1 the RGS varies with 1/m. From equations (13) and (19), which hold for case 2 and case 3, the curves cannot be qualified so easily. The numerical computations reveal shapes very similar to the ones of case 1, be it with larger magnitudes. The RGS values of case 2 and case 3 are between 2 and 4 times larger than in case 1. The RGS values of case 3 are the highest ones and are up to 14% larger than those of case 2.

The random cut scores are directly connected with the RGS (cf. equation (6)). Figure 2 displays the cut scores (which are identical with the normalised cut scores, since w=1).

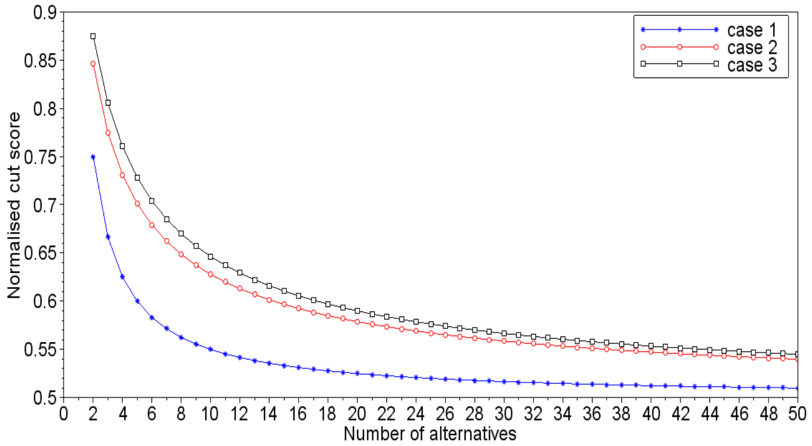


Figure 2 Relative cut score versus the number of alternatives m for each case

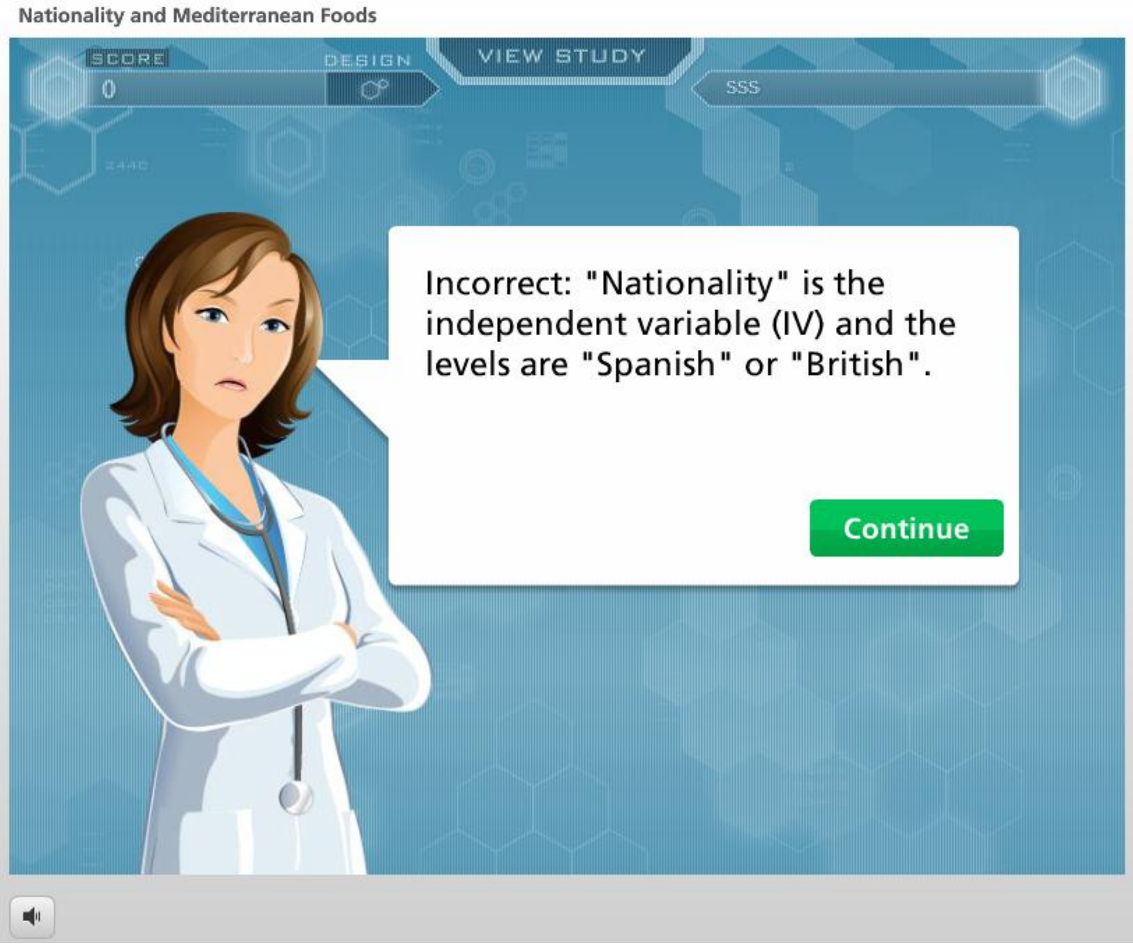
The impact of randomness on cut score is substantial. For a 6-alternatives question (m=6) the cut scores are raised to 0.57 (which is 14% above the 0.50 threshold) for case 1, and to 0.66 (+32%) and 0.69 (+37%), respectively, for the other cases. For m=12 the levels are still 0.54 (+8%), 0.61 (+21%) and 0.62 (+24%), respectively.

Random scores in existing serious games

For exploring the effects of randomness in serious gaming practice, we have analysed two existing serious games, each in the domain of higher education. The first one is a quiz-based game (CHERMUG), representing case 1; the second one is a competence-based game (DIAGNOST), representing case 2 and case 3.

6.1 The CHERMUG games (statistical methods)

The CHERMUG games comprise a set of 14 online mini-games, which have been designed to support students as they learn about research methods and statistics. These contents are of special interest for students and professionals in social sciences and health. Each of the games requires typically 10 minutes to complete. The games were developed in the CHERMUG project (Continuing Higher Education in Research Methods Using Games, http://www.chermug.eu), funded by the Lifelong Learning Programme of the European Commission. The games are freely available at http://playgen.com/chermug. In the CHERMUG games players are confronted with a short, textual scenario related to obesity problems, whereupon they have to identify variables and variable levels, propose a statistical method, interpret statistical outcomes, and so on. Figure 3 shows a screenshot of a CHERMUG game.



Incorrect: "Nationality" is the independent variable (IV) and the levels are "Spanish" or "British".

Figure 3 Screen shot of a CHERMUG game: feedback is given

An analysis of required decisions for three CHERMUG games is given in table 1. All navigational decisions were omitted.

Table 1 Decision taking in 3 CHERMUG games

Topic	Nationality and Mediterranean food	Gender and protein consumption	Type of diet and weight loss
Variables	7(2), 2(1)	6(2), 2(1)	5(2), 2(1)
Levels of measurement	4(1), 4(1)	4(1), 4(1)	4(1), 4(1)
Study type	2(1)	2(1)	2(1)
Hypothesis	2(1), 3(1), 3(1)	2(1), 3(1), 3(1)	2(1) 2(1) 3(1)

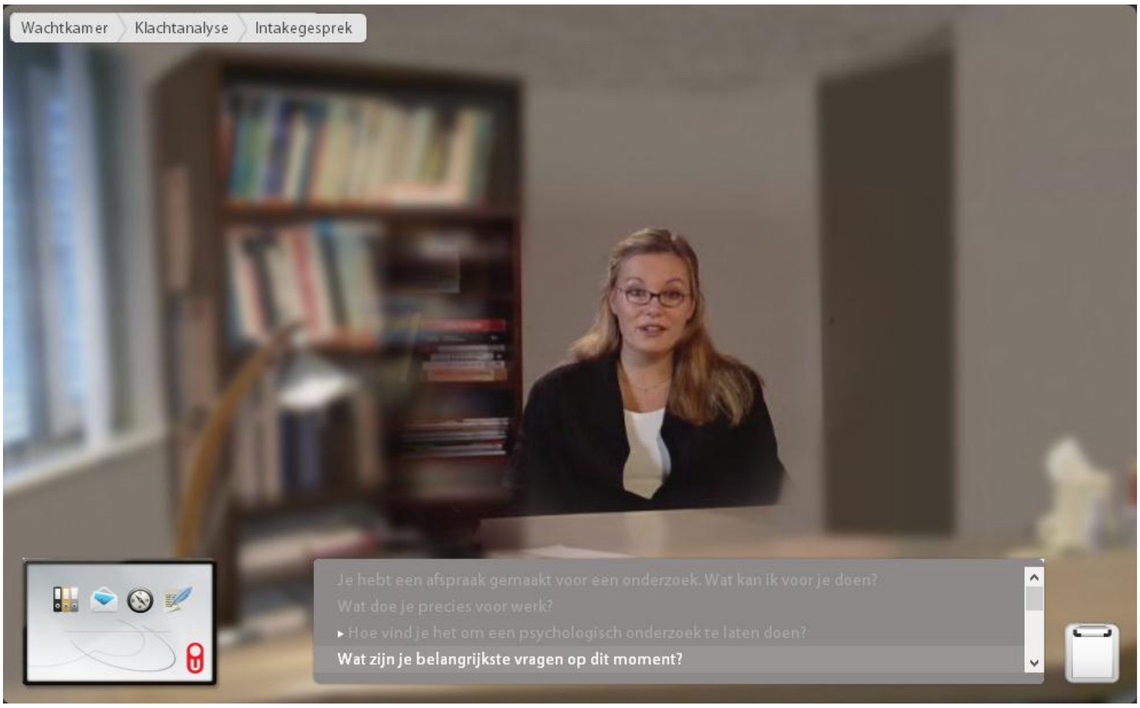
Dataset	2(1), 4(1), 2(1)
	2(1), 2(1)
		2(1), 3(1)
		2(1), 2(1), 2(1)
			2(1), 3(1)
			2(1), 2(1), 2(1)
			2(1) 2(1) 2(1)

Test selection	3(1)	2(1)	3(1)
Interpretation	2(1), 2(1)	2(1), 2(1)	2(1), 2(1)

All games comply to a standard format and mostly use single answer questions. The games are typically a case 1 example (no completion required): after each response a direct, corrective feedback is given. Special game features, e.g. based on hangman and three on a row, are applied as a motivator. As a result the number of questions to be answered depends on performance: weak students may need to answer more questions.

6.2 The DIAGNOST game (psychological diagnostics)

The DIAGNOST game is an online, competence-based game for psychology students to learn how to diagnose a client (Westera, Hommes, Houtmans, & Kurvers 2003). A screenshot is shown in figure 4. The game is an extended multimedia, branching story where the player adopts the role of psychologist and has to decide about the diagnostic approach and the clinical picture. This means: accessing resources, making interviews while asking the relevant questions, selecting relevant topics, decide on hypotheses, decide on psychological testing methods, analyse test results, drawing conclusions, composing validated advice, etcetera. The study load is about 3 hours. The game is used by students of the psychology master of science programme of the Open University of the Netherlands.



Je hebt een afspraak gemaakt voor een onderzoek. Wat kan ik voor je doen?
Wat doe je precies voor werk?
▸ Hoe vind je het om een psychologisch onderzoek te laten doen?
Wat zijn je belangrijkste vragen op dit moment?

Figure 4 Screen shot of the Diagnost game, during a video interview

In accordance with established diagnostic methodology, the game is composed of 4 stages, each of which comprises a limited set of tasks. A decision analysis excluding navigation has yielded the following pattern (table 2).

Table 2 Decision taking in DIAGNOST

Phases	Task	Decisions
Intake	Preparation of intake interview	7(5)
	Intake interview	n/a
	Help question definition	7(5)
Problem analysis	Interview situation analysis	n/a
	Indentify research questions	11(4)
Explanation	Specify hypotheses	9(2), 12(10), 6(2), 3(1)
	Specify test methods	9x3(1), 34x4(1), 2x4(2),

		7x5(1), 6(1), 3x6(2), 8(2)
	Interpret measurement results	7x2(1), 3(1), 3(2), 2x4(2), 5(2), 6(2), 2x8(4)
	Answer research questions	2(1), 3x4(2), 4(3), 5x5(3), 6(3), 6(4), 8(4)
Advice	Answer help questions	12(6)
	Feedback interview	n/a

The game frequently offers multiple-answer questions. The required number of answers k (cf. equation (1)) was always given beforehand, whereby equation (1) holds. We omitted decision taking in the interviews, because in contrast with answering MC questions decisions to view a video are irreversible. Also we excluded navigational decisions. For each of the remaining decisions the game requires completion, which makes it an example of case 2 and case 3. In extreme cases (e.g. when decision taking takes too long), the game drops the completion requirement and provides the correct answers.

Simulating random game play

We have used the data of these games as input for a Monte Carlo simulation. The three CHERMUG games were combined into one game session. Random players were simulated according to the case 1 regime. In this simulation all CHERMUG items were given the same weight: w_i =1, thereby neglecting any content-related issues. At 10,000 iterations the results were stable, variations were well within 0.1 per cent across multiple repetitions. Figure 5 displays the normalised frequency distribution of the relative scores for the 10,000 CHERMUG simulations; the normalised measures were used for reasons of convenience.

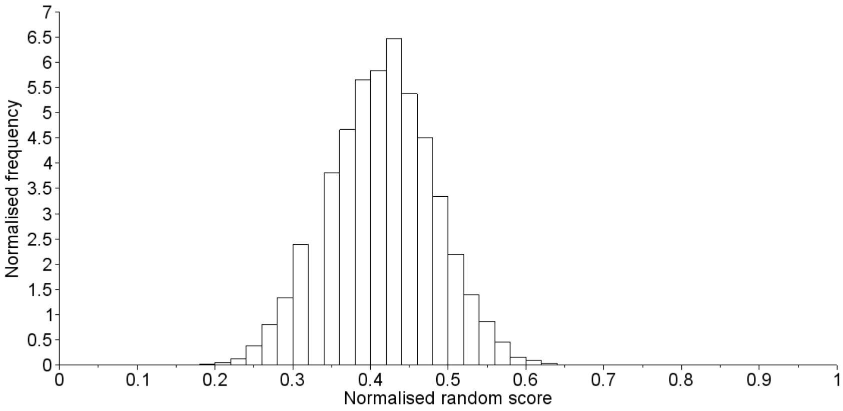


Figure 5 Distribution of random scores for 10,000 CHERMUG simulations

From figure 5 it can be read that the random score contributions have a substantial magnitude. The mean of the distribution is 0.414, explaining 41.4% of the score obtained. This mean value was found to be within 0.1 per cent of the RGS value that was calculated with equation(8), which confirms the consistency of the approach. The spread of the frequency distribution (standard deviation =0.07) reveals substantial variability across different random runs: the coefficient of variation, which is the relative standard deviation (standard deviation divided by the mean), is given by 0.16. According to equation (6) the variability of score as a result of randomness (RSG) directly translates into a variability of the cut score (the cut score is linearly related via 0.50*RSG, cf. equation(6)). Overall it can be concluded that in the CHERMUG games the disturbing effects of randomness are substantial as they explain up to 41% of the performance (not even taking the variability into account: standard deviation of 7%). Randomness raises the normalised cut score to a high level of 0.71 (which is 41% up).

Likewise, the DIAGNOST game was simulated. First we let the simulated players adopt a case 2 strategy (thoughtless play: random choices with replacement). Figure 6 displays the normalised frequency distribution of the scores for 10,000 DIAGNOST iterations. Again uniform weight functions (w=1) were used. From figure 6 it can be read that the random score contributions have a substantial magnitude.

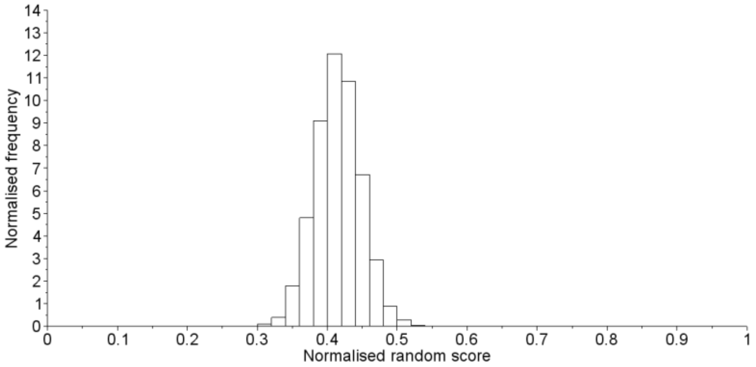


Figure 6 Distribution of random scores for 10,000 DIAGNOST simulations (case 2: completion with replacement)

The mean of the distribution is 0.415, which is within 0.1 per cent of the calculated RGS, cf. equation(13). It explains 41.5% of the score obtained. It should be noted that this value appears to be much higher than the average random success score of the set of items in the DIAGNOST case (cf. equation(8)), which is only 23.0%. Apparently, the game play mode that requires item completion, either with or without replacement, almost doubles the overall RGS score. The spread of the frequency distribution (standard deviation =0.03) reveals some variability across different random runs: the coefficient of variation, which is the relative standard deviation (standard deviation divided by the mean), is given by 0.08. The disturbing effects of randomness in the DIAGNOST game are comparable with those in the CHERMUG games. They explain up to 41% (excluding the effects of the standard deviation of 3%) of the performance. The randomness raises the cut score (ratio) to a high level of 0.71 (which is 41% up).

Similar results were found when the simulated players adopted a case 3 strategy (remembering incorrect decisions: random choices without replacement). Figure 7 shows the case 3 frequency distribution.

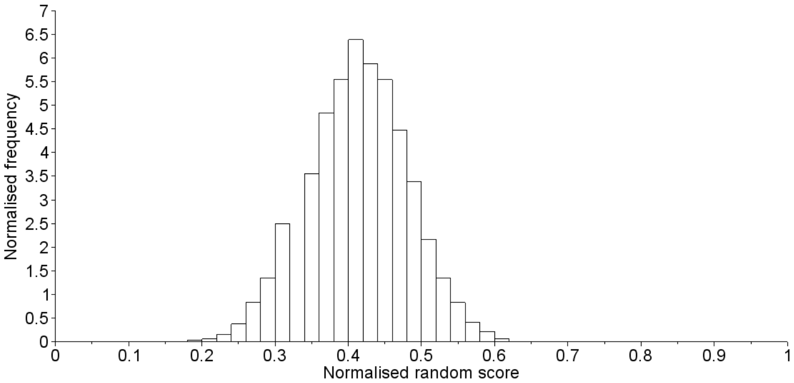


Figure 7 Distribution of random scores for 10,000 DIAGNOST simulations (case 3: completion without replacement)

The mean of the distribution is 0.414, which is within 0.1 per cent of the calculated RGS, cf. equation(18). It explains 41.4% of the score obtained. The standard deviation of the frequency distribution is 0.07, which is more than twice the spread in case 2. The coefficient of variation, which is the relative standard deviation (standard deviation divided by the mean), is given by 0.16. It means that the variability of random score in case 3 is twice the variability of random score in case 2.

Discussion and conclusion

The purpose of this study was to address the following research questions:

(1) How can we formally describe the effects of random game play on the player´s performance score?
(2) What is the magnitude of the effects of random game play?
(3) What is the impact of random game play in practice?

We have addressed research question (1) by developing and testing an analytical model that describes the influence of random effects on the player’s performance score. The model covers two types of decision taking: no item completion required (case 1), and item completion required (case 2, case 3). For required item completion we have identified two different strategies of random decision taking: in case 2 players are supposed to use a strategy of random selection with replacement, while in case 3 they use a strategy of random selection without replacement. In all cases we were able to derive analytical expressions for the random guess scores (RGS) and the resulting cut scores.

We found the RGS for a single-answer question to be different for each case. Obviously, in case 1 the RGS is inversely related to the number of alternatives, which produces a hyperbolic curve. The RGSs in case 2 and case 3 were found to display similar relationships, but they were up to 3 to 4 times larger. Normalised values typically range from 0.10 up to 0.70, which signifies a substantial impact. Case 3 offers a slightly higher RGS than case 2. This can be explained by the fact that case 2 (with replacement) corresponds with a memory-less strategy, while players who adopt a case 3 random strategy (no replacement) never make the same mistakes again and thus will demonstrate higher performances.

Also, we have studied the impact of randomness in two existing serious games (CHERMUG and DIAGNOST). Monte Carlo simulation of random game play confirmed the substantial impact of randomness on score. Random guess scores were found to be around 0.41 for all cases. This agreement between the three cases is deceptive: it appears to be pure coincidence. The simulation was tested for a wider variety of (hypothetical) game configurations and showed large differences between calculated random guess scores. By coincidence the games' branching profiles turned out to be very similar, yielding very similar results for the random guess scores. The high value of the random guess score (0.41) is the result of many low-order items (small m) that are present in the CHERMUG decision profile. When game play requires item completion (e.g. DIAGNOST) the overall RGS score appears to double as compared with the non-completion case. The spread of the random scores is substantial and varies between different games (case 1 and case 2) and different strategies (case 2 and case 3), showing a coefficient of variation up to 0.16. This affects the reliability of observed performances even more. Our study demonstrates that randomness in game play produces a non-negligible effect, the size of which depends on the types of decisions to be taken in the game. Because of the accumulative nature of scores, the effects of randomness don't fade at large numbers of items. The main conclusion based on these observations is that indicators of player performance and progression as derived from the players actions and decisions in the game may be highly inaccurate and unreliable as a result of randomness.

The model has some limitations though. First, the model assumes that players randomly select their decisions without bothering about the content. In fact, this is exactly what constitutes the RGS. But in practice players will be likely to balance different alternatives by looking into the contents and make an educated guess. Such content-related guessing, which can be quite instructive, is additional to randomly guessing. It isn't covered by the model. Hence the RGS sets a lower bound of randomness. Second, we have based the RGS on the combinatorial statistics of multiple choice questions. However, there are alternative methods for determining the RGS, either based on the content-related quality of the decisions or derived from the score distribution of the community of users (Brennan 2006). Third, the model assumes that it is possible to unambiguously determine the pattern of decisions that a player has to make in the game. This isn't always straightforward. Game sessions may display a large variability across different game runs and across different players, both with respect to the number of decisions to be taken, the type of decisions to be taken and the boundary conditions that hold. To some extent different runs of the same game may appear difficult to compare. Fourth, the model starts from the idea that decisions are either right or wrong. However, as is the case in everyday life, the boundaries between correct and wrong in a game are often blurred, or at best conditional. Some choices may be permitted, but unnecessary. Some decisions will only be correct if they are preceded or followed by a sequence of other decisions or achievements. Also, players may deliberately make unfavourable decisions simply because they want to try out things in the game and see what happens. Reverting to previous decisions may be a valuable and productive learning experience. In principle the decisions linked with navigation should be excluded, but often navigational decisions are directly related to the content of the game and make up an essential part of the learning experience. Hence, it may sometimes be difficult to judge the significance of a decision and assign a simple right or wrong. Fifth, the emphasis of our model on performance scores may disregard the importance of learning. Players who have a goal orientation toward performance rather than learning aim to demonstrate their competence to others and receive positive evaluations (high scores) (Dweck 1986; Shute et al. 2009). As a consequence, they are afraid of making mistakes, and tend to avoid or withdraw from complex tasks. Players with a learning orientation, however, show persistence in the face of failure, and display readiness of using more complex learning strategies to master the task (Farr, Hofmann, & Ringenbach 1993). This suggests that it may be wise to be reserved with using game score as a motivator for players, but use the score system as a concealed mechanism for triggering feedback. In any case player achievements have to be judged.

In all cases the assessment of a player's behaviours and performances in a serious game remains an essential component. We have demonstrated the large impact of randomness on the assessment performance, revealing RGSs increased by up to 41 per cent. In a practical context the impact may even be more severe, because student score ranges are seldom uniformly distributed between 0% and 100%, but show sharp bell-shaped distributions. E.g. the distribution of student marks on a 1-10 interval scale in pre-university schools shows an average score of 6.8 with a standard deviation of only 0.9 (Nuffic 2006). Consequently, a random effect on score of only a few per cent will have a much larger practical impact.

References

Abt, C. (1970). *Serious games*. New York, NY: Viking Press.

Aldrich, C. (2005). *Learning by Doing: the Essential Guide to Simulations, Computer Games, and Pedagogy E-Learning and other Educational Experiences*. San Francisco, CA: John Wiley & Sons.

Becker, K., & Parker, J. R. (2011). *The Guide to Computer Simulations and Games*. Indianapolis, IN: John Wiley & Sons.

Bellotti, F., Kapralos, B., Lee, K., Moreno-Ger, P., & Berta, R. (2013). Assessment in and of Serious Games: An Overview. *Advances in Human-Computer Interaction*, 1-11. http://www.hindawi.com/journals/ahci/2013/136864/#B18

Bente, G., & Breuer, J. (2009). Making the implicit explicit: embedded measurement in serious games. In U. Ritterfield, M. J. Cody, & P. Vorderer (Eds.), *Serious Games: Mechanisms and Effects* (pp. 322–343). New York, NY: Routledge.

Brennan, R. L. (2006). *Educational Measurement (Fourth Edition)*. Lanham, MD: Rowman & Littlefield Publishers.

Boston, C. (2002). The concept of formative assessment. *Practical Assessment, Research & Evaluation*, 8(9). http://pareonline.net/getvn.asp?v=8&n=9

Chin, J., Dukes, R., & Gamson, W. (2009). Assessment in simulation and gaming: a review of the last 40 years. *Simulation & Gaming*, 40(4), 553–568.

Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., & Boyle, J. M. (2012). A systematic literature review of the empirical evidence on computer games and serious games. *Computers and Education*, 59, 661-686.

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist*, 41, 1040–1048.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of Educational Measurement (Vol. 5).* New Jersey, NJ: Prentice Hall.

Farr, J. L., Hofmann, D. A., & Ringenbach, K. L. (1993). Goal orientation and action control theory: Implications for industrial and organizational psychology. In C. L. Cooper, & I. T. Robertson (Eds.), *International review of industrial and organizational psychology* (pp. 193–232). New York, NY: John Wiley.

Guttormsen Schär, S., Schluep, S., Schierz, C., & Krueger, H. (2000). Interaction for

Computer-Aided Learning. *Interactive Multimedia Electronic Journal of Computer-enhanced learning* 2(1). http://imej.wfu.edu/articles/2000/1/03/

Kolb, D. (1984). *Experiential Learning: Experience as the Source of Learning and Development*. Englewood Cliffs, NJ: Prentice Hall.

Nuffic (2006). *Fact sheet cijfers ontcijferd*. Nuffic: The Hague. http://www.nuffic.nl/bestanden/documenten/over-de-nuffic/publicaties/factsheet-cijfers-ontcijferd.pdf/view

Redeker, C., Punie, Y., & Ferrari, A. (2012). eAssessment for 21st century learning and skills. In A. Ravenscroft, S., Lindsteadt, C. D. Kloos, & D., Hernandez-Leo (Eds.) *21st Century Learning for 21st Century Skills*. Proceedings of the 7th European Conference on technology-enhanced learning EC-TEL, Saarbrücken, 2012 (pp. 292-305). Heidelberg: Springer.

Reese, H.W. (2011). The Learning-by-Doing Principle. *Behavioral Development Bulletin*, 11, 1-19. http://www.baojournal.com/BDB%20WEBSITE/archive/BDB-2011-11-01-001-019.pdf

Schank, R. C. (1995). *Engines for Education*. New York, NY: Lawrence Erlbaum.

Schank, R. C., Berman, T. R., & Macpherson, K. A. (1999). Learning by doing. In C. M. Reigeluth (Ed.), *Instructional-design theories and models: A new paradigm of instructional theory (Vol. II, pp. 161-181)*. Mahwah, NJ: Lawrence Erlbaum Associates.

Schön, D. A. (1983). *The reflective practitioner: How professionals think in action.* New York, NY: Basic Books.

Shute, V. J., Ventura, M., Bauer, M., & Zapata-Rivera, D. (2009). Melding the Power of Serious Games and Embedded Assessment to Monitor and Foster Learning: Flow and Grow. In U. Ritterfeld, M. Cody, & M. Vorderer (Eds.), *Serious Games: Mechanisms and Effects* (pp. 295-321). New York, NY: Routledge.

Vargas, J. S. (1986). Instructional Design Flaws in Computer-Assisted Instruction. The *Phi Delta Kappan*, 67(10), 738-744. Retrieved from http://www.jstor.org/stable/20403230

Westera, W., Hommes, M. A., Houtmans, M., & Kurvers, H. J. (2003). Computer-Supported Training of Psycho-diagnostic Skills. *Interactive Learning Environments*, 11(3), 215-231.

Westera, W., Nadolski, R., Hummel, H., & Wopereis, I. (2008). Serious Games for Higher Education: a Framework for Reducing Design Complexity. *Journal of Computer-Assisted Learning*, 24(5), 420-432.