Journal of Universal Computer Science, vol. 27, no. 8 (2021), 854-876
Doi: http://dx.doi.org/10.3897/jucs.78248

Assessing Repetitive Trials in Serious Games

Wim Westera
(Open University of the Netherlands, Heerlen, The Netherlands
https://orcid.org/0000-0003-2389-3107, wim.westera@ou.nl)

Abstract: Players in serious games may often need multiple trials for correctly completing a game task. Therefore, the number of attempts should be reflected in the score. This article presents three computational score models that take into account the number of attempts that a player makes to be successful. The models, which are extensions of test theory, quantify the random contributions to the scores that need to be removed. They also describe the influence of prior knowledge used for elimination of incorrect options, and take into account that the decision options in a node may not be equally plausible. In a series of simulation studies the score outcomes of the models are compared under various conditions. Results show that the number of trials cannot be ignored as they have a strong impact on the performance scores to be assigned. Neglecting the number of trials leads to inaccurate scores that significantly overrate the observed performances, occasionally up to 100% or even more. The effects depend on the number of decision options, the presence of obvious, correct or incorrect options given the player´s knowledge level and, to a lesser extent, different plausibility levels of the options to decide upon. The practical feasibility is high, because a simple score formula largely solves the problem.

Keywords: games, serious games, learning, multiple choice, assessment, score, item, trial, retrial, computational, model, simulation
Categories: L0, L1, L2, L3, L5
DOI: 10.3897/jucs.78248

1 Introduction

Digital games are increasingly deployed in education and training. Such “serious games” [Abt, 70] are valued for their potential to engage and motivate and to provide a challenging, meaningful learning context, while they benefit from the latest multimedia technologies. A critical element of serious games is the monitoring of learning achievements (e.g. [Chin, 09], [Bellotti, 13], [Connolly, 2012], [Shute, 2009], [Westera, 2019]). Digital games offer the opportunity to track the player´s successes and failures during the game and base the player´s performance assessment on those data traces. Accordingly, most games include progress measures, which are used for score assignment, level transitions, feedback or adaptation of the game play [Boston, 2002] [Shute et al., 2013] [Westera, 2015]. This on-the-fly assessment fits well in the trend toward formative assessment [Redeker et al., 2012]. It marks a shift from post-practice assessment based on a single moment of observation toward an individualised approach with a large number of observations over time. The need for timely in-game performance tracking and associated feedback is further indicated by the unfavourable fact that games tend to favour trial-and-error learning strategies and free exploration [Vargas, 1986]. As many game interactions are driven by direct manipulation of graphical objects, which inherently put little cognitive load on the users, they tend to reinforce an shallow, impulsive, trial and error learning mode

[Guttormsen Schär et al., 2000]. The same holds for popular game design elements that induce stress, such as a time lock or time-dependent scores, which are likely to promote hurried, shallow or incomplete information processing. As such, failure during the process of learning needs not be a problem, because failure is a crucial factor for any learning to occur. Still, it calls for timely feedback on performance. To this end, the question arises how to describe the performance of players that frequently redo parts of the game before finally succeeding: if two players both succeed at successfully completing a game, but one player needed more tries than the other, how should this be reflected by the respective performance scores? In many cases, games use ad hoc metrics, as simple as playing time spent or the number of trials that were needed, but these are neither theoretically grounded nor empirically validated. Classical test theory [Spearman, 1904] and Item Response Theory [Lord, 1980] do provide formalised solutions for the case of repetitive failures, but these only describe the effects of random guessing by ignorant subjects (e.g. monkeys) devoid of thoughtful considerations, which does not hold for human individuals. Dynamic approaches such as Computerised Adaptive Testing [Weiss et al, 1987] and personalisation approaches [Maseleno et al., 2018] do allow to adjust task difficulty to observed player skills, but do not explain how performance should be calculated when multiple efforts are needed to succeed.

Our research question reads: how do repetitive efforts during game play translate into the indicators of performance? This paper will present three formalised score models that take into account the revisiting op decision nodes after failure. The first model extends the common test-theoretical approach by accounting for the number of trials. Model 2 offers a further extension by including the possibility that the player simply knows the right decision or knows that one or more options are incorrect and can be removed beforehand. In addition Model 3 also includes the possibility that the various options aren't equally plausible. The models are implemented in a computer program that simulate game play, which allows to evaluate and compare the models under various conditions, such as different numbers of options, and the effects of early eliminations.

2 Model development

2.1 Player-led decision taking in serous games

By principle, all games involve active decision taking by the players, who are challenged to achieve favourable outcomes and maximise their performances. A game can thus be represented as a network of decision nodes. These decision nodes represent challenges or activities that have multiple potential outcomes. Rather than micro-level decisions (single mouse clicks) the nodes represent meso-level problems, challenges or activities possibly requiring a set of connected considerations by the player, viz. meaningful pursuits directly originating from the game scenario. Examples would be fighting a monster, writing a note, talking to a character, tracking a hidden object, moving objects, buying supplies, navigating, etcetera, all requiring active consideration and decision making by the player. The edges of the network represent the pathways between the game activities. During gameplay, the player

moves from node to node, while performance is based on the achievements in each node, related to the successes and failures experienced. Different player decisions in a node may lead to different outcomes that open up different pathways in the network. Accessibility of the nodes may change over time: not all nodes need to be accessible from the start, but they may gradually open up when precedence conditions are met. Such conditions may be related to the logic of storytelling, the causal order of events, time spent, the interdependence of content, and the player's achievements so far. Likewise, some pathways may get blocked during the game. Game play is thus described by a pathfinding process through the dynamic network of decision nodes. Given this network structure and the principle of player-led decision taking, playing a game is conceptually equivalent with taking a multiple choice test, where the nodes correspond with the test items. This equivalence is manifest in games that are structured as branching stories or quizzes [Aldrich, 2009], but it also holds for many other cases where the game style and context may conceal the player-led decision structure.

2.2 The process of decision taking

A serious game can be represented as a network of N decision nodes, each of which offer a number of decision options [Westera, 2018]. For reasons of simplicity we assume that the decision nodes are dichotomous and single select, which means that only one of the options in the node is correct, while the remaining options are incorrect. Consider a node with m options to decide upon. How does this decision process look like? In some cases the players may immediately spot the correct decision, simply because they possess sufficient knowledge and skills to decide rightly. If they cannot, they will check if they can cross out some of the options (q) as being incorrect, again relying on their knowledge and skills. Thereby they reduce the pool of options to choose from to (m-q). Because the players are uncertain about the remaining options, they have to take a chance and may need multiple trials to get it right. After each failure, the pool of options is reduced further, which is enforced by the assumption that the player doesn't make the same mistake again. The decision tree of such process is depicted in [Fig. 1].

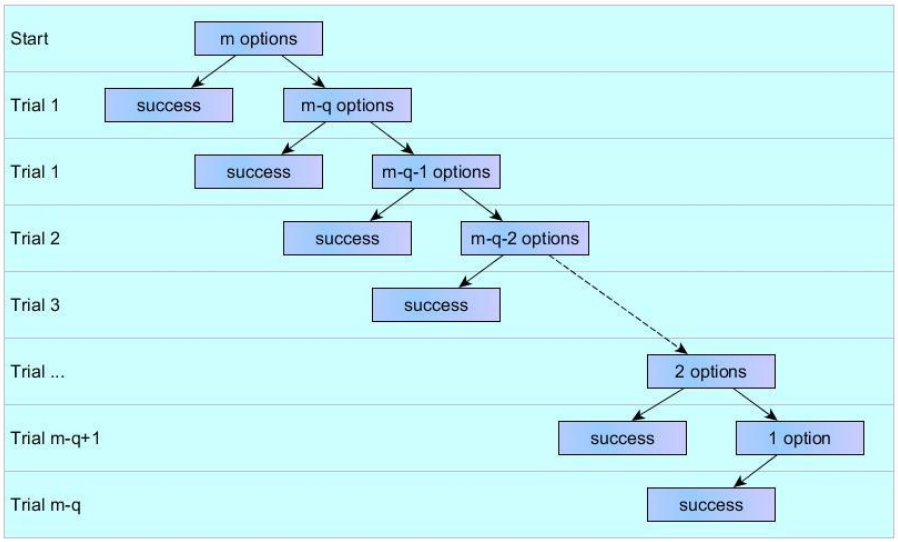

Figure 1: The decision tree of a dichotomous and single select node with m options and q prior eliminations.

Given this process, the next sections present three score models that do take into account the effects of the player needing multiple trials to pass a node.

2.3 Model 1: The extended test-theoretical model

2.3.1 Single visit

Test theory introduces the Random Guess Score (RGS), which is the score level that is obtained by simply taking decisions randomly. The Random Guess Score sets a lower bound to the performance that can reasonably be attributed to the candidate's capabilities [Gronchi et al., 2021]. In principle, the Random Guess Score of a test item, which is denoted as RGS, equals the sum of the probability for each option of the item multiplied by the assigned score of the associated outcome [Draaijer et al., 2018]. In the case of dichotomous and single select decision nodes the Random Guess Score in a node with *m* options reduces to:

$$RGS = \frac{1}{m} \tag{1}$$

Since the Random Guess Score is not the result of the player's performance, the player's score *S* needs to be corrected by removing this random contribution.

$$S = 1 - \frac{1}{m} \tag{2}$$

2.3.2 Revisits

A slight extension of the test-theoretical model allows to cover the effects of repetitive trials. After a failure in a node, the player may revisit the node, while taking up a different approach to avoid making the same mistake again. Consequently, in the second trial there are only *(m-1)* options left. With a slight adjustment of equation (3) the RGS after *t* tries in a node can be written as:

$$RGS(t) = \frac{1}{(m-t+1)} \tag{3}$$

This formula holds for all $t \leq m$. Accordingly, the score *S* from equation (4) can be adjusted to include the number of trials *t* in the node.

$$S(t) = 1 - \frac{1}{(m-t+1)} \tag{4}$$

2.3.3 Issues with the basic test-theoretical model

Although the test-theoretical approach is well-established, it suffers from two fundamental weaknesses. First, it does not consider the full decision tree displayed in [Fig. 1]. In particular, it corrects the score for a random contribution even though players may simply know the correct decision without any guesses made. Second, it is assumed that the *m* decision options in a node are equally plausible. But if that were really the case, the recorded decision in the node would not be informative for distinguishing between weak and strong performances: a node with equally plausible options could best be removed from the performance metric (the item's discrimination index is said to be equal to zero). In practice, however, the diverse options in a node have, and actually they should have, different degrees of plausibility to be of any value for performance assessment: some options being more plausible than other options.

The next model, which is Model 2, extends the basic test-theoretical model by including the possibility that the player simply knows the right decision or knows that one or more options are incorrect and can be removed beforehand. In addition Model 3 includes the possibility that the various options aren't equally plausible.

2.4 Model 2: The knowledge-based decision model

2.4.1 Single visit

This model extends Model 1 by incorporating the possibility that the player may simply know the right answer, so that no guessing is involved at all. Second, there is a possibility that the player can tell in advance that one or more of the options are incorrect and can be crossed out, thus changing the odds. These starting points call for a probability analysis of the decision process [see Fig. 1], which is presented below.

We define the following four disjunctive events, in accordance with the decision tree of [Fig. 1]:

1. *K* = the player simply knows the correct decision.

2. $R_{j1, \ldots, jq}$ = the player doesn't know the correct decision, but knows that q specific options ($j_1, \ldots, j_q$) are incorrect so that they can be removed from the pool of options to decide upon.
3. G = the player does not know the correct answer and cannot identify options to be eliminated. So, the player has to make an educated guess from the pool of m options.
4. C = the player selects the correct decision.

The probability *P(C)*, which is the chance that the player makes the correct decision can now be decomposed by accumulating the conditional probabilities of its constituents:

$$P(C) = P(C|K).P(K) + \sum_{\substack{all\ subsets \\ 1 \le q < (m-1)}} P\left(C \middle| R_{j_1, \ldots, j_q}\right).P_E(q) + P(C|G).P(G) \quad (5)$$

P(C) accumulates the probabilities of the full set of eventualities associated with taking the right decision *C*. All components from the right hand side of equation (5) can now be elaborated.

P(K)

The probability *P(K)* that players know and recognise the correct decision is directly related to their personal knowledge, skills and experiences. But these are hard to specify. However, in well-tuned serious games the players' knowledge level is cautiously reflected in the difficulty of the game challenges. Decision taking should be doable [Gee, 2003], that is, it should be neither too straightforward nor too difficult [Nyamsuren et al., 2017], otherwise players will either get bored or frustrated [Csikszentmihalyi, 1991]. According [Eggen et al., 2006] and [Klinkenberg et al., 2011] an average success rate of typically *P(C)=0.75* provides a good balance winning and losing as to keep a player motivated. Successes, however, are mainly the result of learning new things, an educated guess, or just pure chance, and for the remainder on knowing. The probability that the player knows the correct decision is likely to be small, typically in the range of 0.25 or below, to prevent the game play from being degraded to a blanks exercise.

P(C|K)

The probability of taking the right decision, when one knows the right decision can be simply set to 1:

$$P(C|K) = 1 \quad (6)$$

because if one knows the right decision then one will take the right decision.

$P_E(q)$ and $P(R_{j1, \ldots, jq})$

$P_E(q)$ denotes the probability that the player is able to identify and eliminate q options from the (*m-1)* options that are incorrect. The probability that the player is able to

identify exactly one incorrect option is thus be written as $P_E(1)$. As in the case of knowing the correct decision, this probability is likely to be small, typically in the lower range smaller than 0.25. In fact, the ability to spot one incorrect decision should not be very different from spotting one correct decision. As it uses essentially the same knowledge base of the player we may assume

$$P_E(1) \approx P(K) \tag{7}$$

If we assume equal probabilities of the (m-1) incorrect options, the probability $P(R_j)$ that the player identifies not just one of the options but exactly option j as an incorrect option can be written as

$$P(R_j) = P_E(1) \cdot \frac{1}{(m-1)} \approx P(K) \cdot \frac{1}{(m-1)} \tag{8}$$

Consequently, the probability $P(R_{j1,j2})$ that both option j_1 and option j_2 are recognised as the only two incorrect options can now be written as

$$P(R_{j1,j2}) = 2! \cdot P(K)^2 \cdot \frac{1}{(m-1)} \cdot \frac{1}{(m-2)} \tag{9}$$

The generalised expression of the probability $P(R_{j1,j2,\ldots,jq})$ that exactly the options $j_1,j_2,\ldots,j_q$ are eliminated can be written as

$$P(R_{j1,j2,\ldots,jq}) = q! \cdot P(K)^q \cdot \frac{1}{(m-1)\cdot(m-2)\cdot\ldots\cdot(m-q)} \tag{10}$$

$P(C|R_{j1,j2,\ldots,jq})$

Once that the options $j_1,j_2,\ldots,j_q$ have been eliminated from the pool, the player has to guess from the remaining *(m-q)* ones, labelled $j_{q+1}, \ldots, j_m$. The probability $P(C|R_{j1,j2,\ldots,jq})$, which is the chance to select the correct option from the remaining *(m-q)* candidates is given by

$$P(C|R_{j1,j2,\ldots,jq}) = \frac{1}{(m-q)} \tag{11}$$

P(G)

The probability *P(G)* for the occasion that the player neither knows the correct decision nor finds options to be eliminated can be easily resolved from the sum of probabilities being one:

$$P(G) = 1 - P(K) - \sum_{\substack{all\ subsets \\ 1 \le q < (m-1)}} P\left(C \middle| R_{j_1,\ldots,j_q}\right) . P_E(q) \tag{12}$$

P(C|G)

Finally, the probability *P(C|G)*, which is the chance to select the correct option from the full set of *m* candidates is given by

$$P(C|G) = \frac{1}{m} \tag{13}$$

Herewith all components of equation (5) are covered and the probability *P(C)* of taking the correct decision can now be fully evaluated.

Once the player has successfully passed the node, given the probabilities explained above, the score can be expressed as an extension of equation (2) by incorporating the effects of knowing the solution and the capability to remove one or more incorrect options from the pool. By extending the score expression of Model 1, the score of Model 2 for a single trial can be written as

$$S = 1 - \frac{\gamma}{(m-q)} \tag{14}$$

with $\gamma=0$ when the player knows the correct decision (no guessing involved) and $\gamma=1$ otherwise (correction for guessing needed).

$$P(\gamma = 0) = P(K) \tag{15}$$

2.4.2 Multiple trials

As in the previous model, the player may revisit the node after an incorrect decision to give it another try. When the player failed to immediately recognise the correct decision in the first trial, the player will neither be able to do so in the next try, given the same knowledge base. The eliminations identified in the first trial would still hold, leaving the smaller pool of remaining options. Each time when a wrong is decision taken, it is removed from the pool. In the second and subsequent trials the player just makes an educated guess from the remaining options. The score formula for a success in the *t-th* trial thus reads:

$$S(t) = 1 - \frac{\gamma}{(m-q-t+1)} \tag{16}$$

2.5 Model 3: The knowledge and plausibility model

2.5.1 Single visit

This model extends Model 2 by also taking into account that the *m* decision options in a node may not be equally plausible. By design, each option *j* in a node should have a plausibility value assigned that is given by w_j, a number between 0 and 1 that indicates to what extent the player may accept the option as a credible solution. The plausibility concept reflects the deceptive appearance of competing options: to offer sufficient challenge in a serious game, the correct option should never be too obvious, but should be flanked by competing options that may seem quite okay, but include subtle error or obscured pitfalls. The plausibility values act as non-uniform weight factors for the respective options, assigning plausible options a higher value than less plausible ones. Consequently, generic score formulas that assume even probabilities

of the options are no longer valid. Instead, various ingredients of *P(C)* in equation (5) should be further detailed.

P(Rj1, ….,jq)

The probability $P(R_j)$ that the player identifies option j as an incorrect option is proportional to its implausibility value, which is $(1\text{-}w_j)$. Let for reasons of convenience, the correct option, which is not among the candidates to be eliminated, be located at $j=1$. Then we can write

$$P(R_j) = P(K) \cdot \frac{(1-w_j)}{\sum_{k=2}^{m}(1-w_k)} \tag{17}$$

The probability $P(R_{j1,j2})$ that both option j_1 and option j_2 are recognised as the only two incorrect options can now be written as

$$P(R_{j1,j2}) = 2! \cdot P(K)^2 \cdot \frac{(1-w_{j1})}{\sum_{k_1=2}^{m}(1-w_{k_1})} \cdot \frac{(1-w_{j2})}{\sum_{\substack{k_2=2 \\ k_2 \neq j_1}}^{m}(1-w_{k_2})} \tag{18}$$

The generalised expression of the probability $P(R_{j1,j2,\ldots,jq})$ that exactly the options $j_1,j_2,\ldots,j_q$ are eliminated can be written as

$$P(R_{j1,j2,\ldots,jq}) = q! \cdot P(K)^q \cdot \frac{\prod_{j_i=j_1}^{j_q}(1-w_{ji})}{\sum_{k_1=2}^{m}(1-w_{k_1}) \cdot \sum_{\substack{k_2=2 \\ k_2 \neq j_1}}^{m}(1-w_{k_2}) \cdot \ldots\ldots \cdot \sum_{\substack{k_q=2 \\ k_q \neq j_1,\ldots,j_{q-1}}}^{m}(1-w_{k_q})} \tag{19}$$

This may seem a complex formula, but it is fully denumerable and thus it can be easily evaluated by a computer.

P(C|Rj1,j2,…,jq)

Once that the options $j_1,j_2,\ldots,j_q$ have been eliminated from the pool, the player has to guess from the remaining *(m-q)* ones, labelled $j_{q+1}, \ldots,j_m$. Although the player has demonstrated not to be able to tell for sure whether an option is right or wrong, some options are more likely than other options because of different plausibility values. Consequently, the player's decision will be an educated guess guided by the respective probabilities w_j. As defined before, the correct option is located at $j_k=1$ with plausibility w_1. Then the probability $P(C|R_{j1,j2,\ldots,jq})$, which is the chance to select the correct option from the remaining *(m-q)* candidates $j_{q+1}, \ldots,j_m$ is simply given by

$$P(C|R_{j1,j2,\ldots,jq}) = \frac{w_1}{\sum_{j_k=jq+1}^{jm} w_{j_k}} \tag{20}$$

P(G)

The probability *P(G)* for the occasion that the player neither knows the correct decision nor finds options to be eliminated can be easily resolved from the sum of probabilities being one:

$$P(G) = 1 - P(K) - \sum_{\substack{all\ subsets \\ 1 \le q \le (m-1)}} P\left(C \middle| R_{j_1,\ldots,j_q}\right) . P_E(q) \qquad (21)$$

P(C|G)

Finally, the probability *P(C|G)*, which is the chance to select the correct option from the full set of *m* candidates, while taking into account their respective plausibility values w_j, is given by

$$P(C|G) = \frac{w_1}{\sum_{j=1}^{m} w_j} \qquad (22)$$

Herewith all components of equation (5) are covered and *P(C)*, the probability of taking the correct decision, can now be fully evaluated. In accordance with item-response theory the score needs to be adjusted for constituents that cannot reasonably be attributed to the player. For this case, we cannot apply the generic random accounts of *1/(m-q)* (see equation (14)), but we should apply a weighted adjustment of the score, based on plausibility values. Successes will be easy when w_1 is relatively large: a w_1 close to one is like a give-away. Successes will be difficult when w_1 is relatively small. The score formula can now be specified as follows:

$$S = 1 - \frac{\gamma \cdot w_1}{\sum_{\substack{All\ j_k \\ in\ the \\ pool}} w_{j_k}} \qquad (23)$$

with $\gamma=0$ when the player knows the correct decision (no guessing involved) and $\gamma=1$ otherwise (correction for guessing needed).

2.5.2 Revisits

Equation (23) also holds for multiple trials. The eliminations possibly identified in the first trial would still hold, leaving a smaller pool of remaining options. In the second and subsequent trials the player makes an educated guess from the remaining options and while taking the plausibility values into account. Each time when a wrong decision is taken, it is removed from the pool.

3 Model investigations

To investigate the score models presented above, we have implemented the models in a SCILAB computer program (http://www.scilab.org) and carried out a number of

simulation runs. To represent the game, a network of N evenly rated game nodes was generated. Baseline parameters of the simulation model are explained and summarised in [Tab. 1].

Description		**Parameter**	**Value**
Number of game nodes		N	40
Number of options in the nodes:			
A	Randomised by sampling from a normal distribution	Mean m	6
		Standard deviation sd	2
		Removal below cut off	<3
B	Fixed	M	10
Plausibility:			
C	Random draw from a uniform distribution	w_{ij} from $U(0,1)$	$0 \leq w_{ij} \leq 1$
D	Equidistant linear distribution	$w_{ij}=0.9-0.8*(j-1)/(m-1)$	$0.1 \leq w_{ij} \leq 0.9$
		w_{max}	0.9
		w_{min}	0.1
Number of players		n	1, 100
Probability of knowing the correct decision		$P(K)$	0.25
Probability of knowing 1 option that can be removed		P_{R1}	0.25

Table 1: Baseline simulation parameters

The parameter settings in the right-side column were defined by taking into practical significance and representativeness, and limitations with respect to simulation processing time. The number of options in node i, given by $j=1, \ldots, m_i$, was either randomly generated (A) or predefined as a constant (B). For case A the random number of options in each node was drawn from a normal distribution with mean 6 and standard deviation ($sd=2$). By rejecting and redoing values of m_i smaller than 3, the number of options typically ranged from 3 to 12.

Likewise, plausibility values of Model 3 are either randomly generated (C) or predefined (D). The random versions of plausibility (C) were drawn from a uniform distribution. The predefined values (D) are at an equidistant interval between a minimum value ($w_{min}=0.1$) and a maximum value ($w_{max}=0.9$). The correct decision option in each node was assigned a score of 1 point, while wrong decisions prompted the players to retry until the node is successfully passed. To simplify comparison, all players move through the network of nodes in a fixed order. In each node, the player's score is then updated in accordance with the score models and the respective node is removed from the set of accessible nodes. In Model 2 and Model 3 the statistics are produced by integrating over q, the number of incorrect cases identified, with q running from q=1 to q=(m-1), cf. equation (12) and equation (21), respectively. The average value of q is largely determined by the preset probability P(PR1) (see equation (7) and equation (9)) and the number of options m.

3.1 Study 1: How scores of the models differ

[Fig. 2] shows the progression of the normalised cumulative scores of the three models in a single run with random options and random plausibility (A,C in [Tab. 1]) . Also the ideal score (all decisions are right in their first go) is indicated.

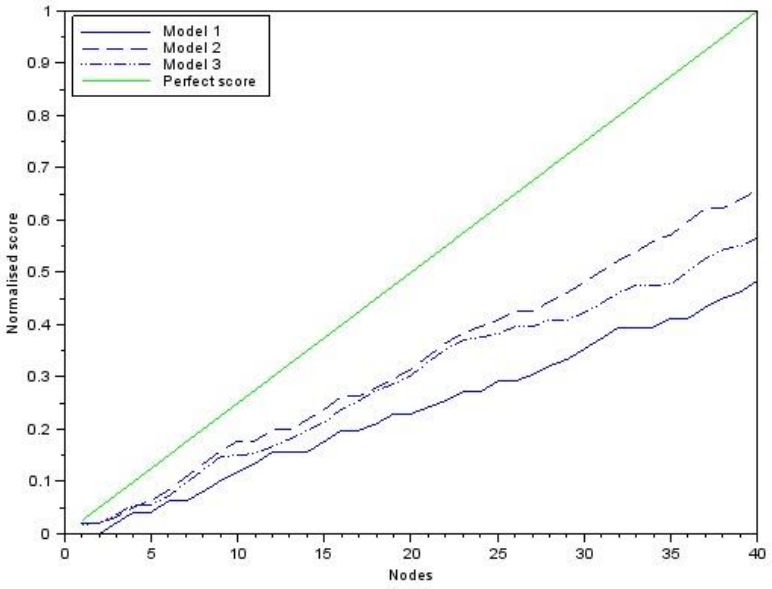

Figure 2: Score progression of models 1, 2 and 3 as well as the perfect score for a single player as a function of progressions along the game nodes.

The variability of the three scores in this single run is caused by the randomly generated number of options in each game node, the probabilistic process of decision taking and the randomly generated plausibility values (Model 3), respectively. [Fig. 3] shows the mean cumulative scores for a run with 100 players.

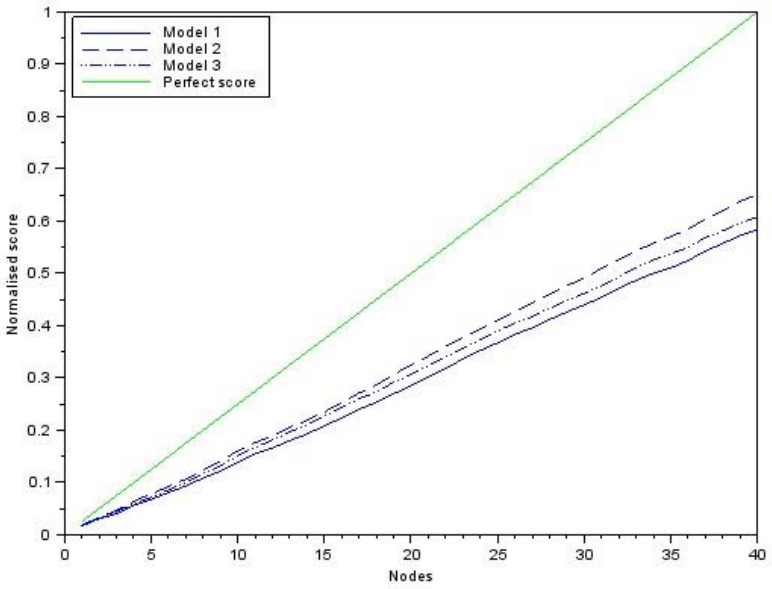

Figure 3: Score progression of models 1, 2 and 3 as well as the perfect score for a 100 players.

The variability almost fades with this large number of players, but even so a slight ripple remains, which can be attributed to the random distribution of options in the game. From [Fig. 3] it can be seen that the models produce different score levels and that all scores are substantially lower than the perfect score.

In order to accommodate easy replication, all simulations in the remainder are based on a fixed number of options (*m=10*) (case B) [see Tab 1] and a predefined set of plausibility values in accordance with an equidistant linear distribution (case D) [see Tab 1]. The overall relative scores of the three models and the mean number of trails for this fixed and linear distribution approach (case B,D) are presented in [Tab. 2].

	Normalised score	**Standard error of the score**	**Mean trials per node**	**Standard error of trials per node**
Perfect score	1.00	0	1.00	0
Model 1	0.709	0.001	5.50	0.05
Model 2	0.780	0.001	4.18	0.05
Model 3	0.749	0.001	3.01	0.03

Table 2: Scores and number of trials of the three models and standard errors.

All scores in [Tab. 2] are represented as values relative to the perfect score. The score of Model 1 displays the largest gap with the ideal perfect score: as much as almost 29 % of the score is lost both as a result of random processes that cannot be attributed to the player and reductions applied because multiple trials were needed. Since no knowledge-based selections or eliminations are taken into account in Model 1, retrials are more often needed than in Models 2 and 3, respectively. The extra information from the plausibility values in Model 3 help to substantially reduce the number of efforts. However, differences in plausibility make it easier to discriminate and to decide between the options, which in the end is compensated for by extra reductions in the scores.

3.2 Study 2: How scores decrease with the number of efforts

A baseline simulation was run with the equidistant linear distribution of plausibility values. [Fig. 4] shows the assigned scores in each model as a function of the number of trials needed to pass the nodes.

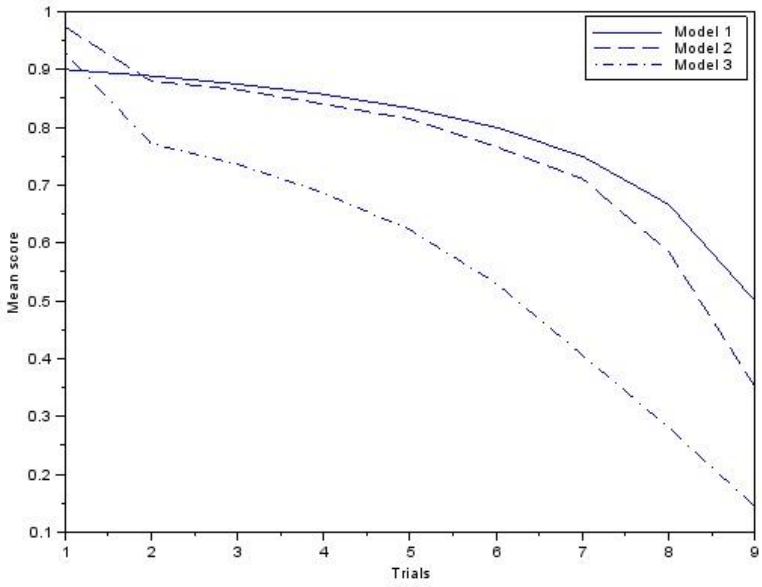

Figure 4: Dependence of the normalised score on the average number of trials that the players needed.

Given the substantially decreasing scores in [Fig. 4] it can be concluded that the number of trials cannot be ignored in the score. Neglect of the number of trials would

unjustly assign a score of 1 point or, at best, a score of 0.9 when a correction for the static Random Guess Score is applied (*1/m = 0.1*). The curve of Model 1 starts at the 0.9 level (standard Random Guess Score applied). At subsequent trials, the score gradually drops since ever less options remain while the random contribution increases (cf. equation (4)). The curves of Model 2 and Model 3 have starting scores well above 0.9, because in both models it is assumed that players sometimes may know the correct decision without guessing and thus obtain the full score of 1 point. However, overall the scores of Model 2 and Model 3 are well below also those of Model 1. This is the result of eliminations in the first trial, which reduces the pool of options and thus increases the random component that needs to be corrected for. For Model 3 the more diverse range of plausibility values makes it easier to eliminate options, which reduces the part of the score that can be directly attributed to the player. This is further explained by the data in [Tab. 3] and [Fig. 5].

	First hits per node	**Knowledge-based first hits per node**	**Random first hits per node**	**Eliminations per node**	**Trials per node**
Perfect score	1.00	0.00	1.00	0	1.00
Model 1	0.10	0.00	0.10	0	5.50
Model 2	0.33	0.25	0.08	0.33	4.18
Model 3	0.39	0.25	0.14	0.63	3.01

Table 3: First hits, elimination and number of trials for the various models.

[Tab. 3] shows the overall fractions (per node) of the first hits, eliminations and the number of trials required. The data confirm that Model 2 and Model 3 account for more early successes and allow for eliminations. In those respects, Model 3 outperforms Model 2, which is ascribed to the larger discrimination index associated with the non-uniform distribution of plausibility values in Model 3, which increases the probability of eliminations (cf. equation (19)). Consequently, the pool of remaining options in model 3 is commonly smaller, so that less trials are needed to be successful, as is displayed in [Tab. 3].

[Fig. 5] shows how often *q* eliminations are made (*q=1,..,m-1*) for each node.

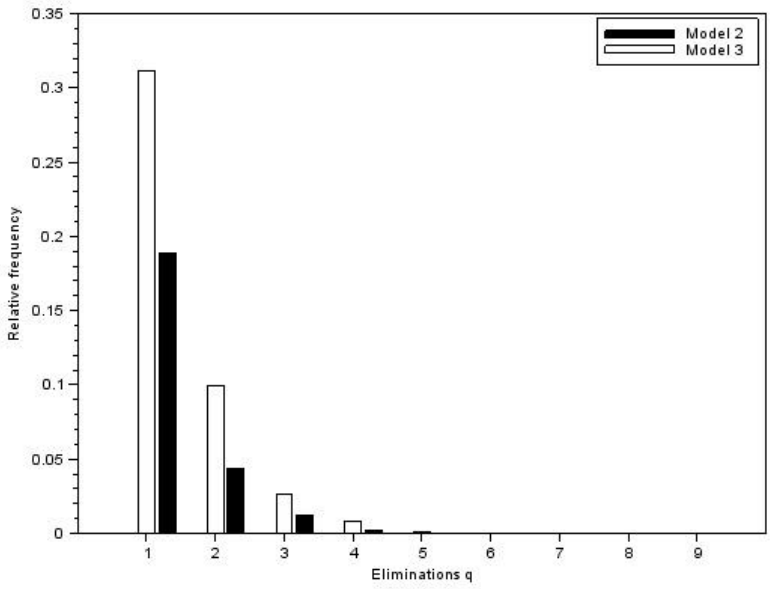

Figure 5: Relative frequencies of eliminations q for Model 2 and Model 3.

Both models display that the frequencies of eliminations rapidly decrease with q. Model 3 allows for about twice as many eliminations as Model 2, which is in agreement with the data in [Tab. 3].

3.3 Study 3: Dependence of scores on the number of options

To investigate how the scores of the models depend on the number of options (m) in the nodes, the latter was varied step by step from $m=3$ to $m=20$ in the linear baseline set-up (B,D), while each time the score was reiterated. [Fig. 6] shows the mean scores of the three models in the baseline set-up.

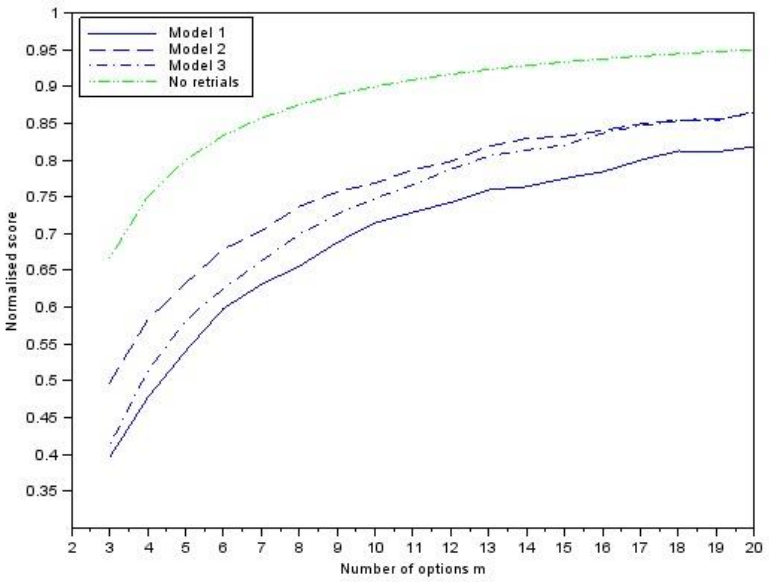

Figure 6: Mean scores as a function of the number of options in the decision nodes (m).

The three lower curves incorporate the effects of retrials and need to be compared with the static score level of 1.0 or at best with the upper curve, which takes into account random effects of a single trial (cf. equation (2)) and is commonly used in practice. By neglecting the influence of retrials the assigned scores are easily overrated, up to 70% at small values of *m*. Although the three Model curves display similar patterns, the scores of Model 2 and Model 3 are structurally higher than those of Model 1. Apparently, the positive effects of knowing the correct decision compensate the negative effects of reducing the pool by eliminations.

[Fig. 7] shows how the number of trials changes for each model. Standard errors are all well below 0.1 %.

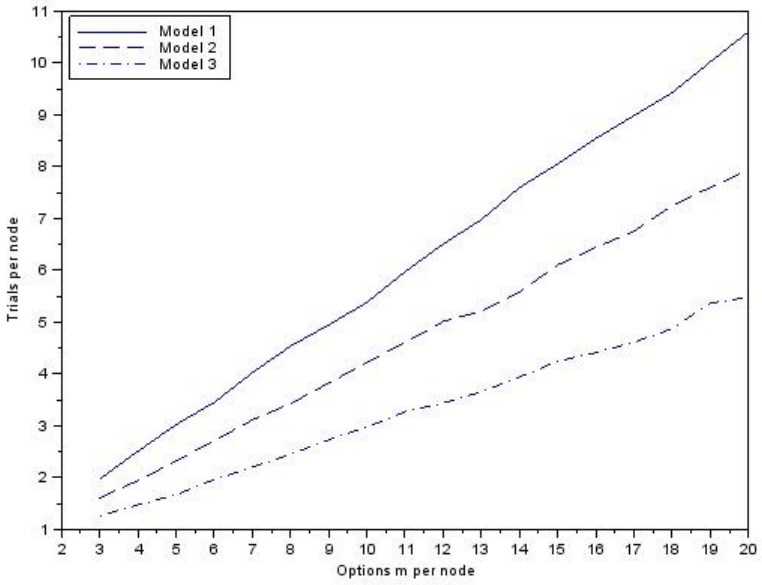

Figure 7: Mean number of efforts needed for a pass in Models 1, 2 and 3 as a function of the number of options in the nodes.

For all models the mean number of trials increases more or less proportionally with the number of options in the nodes. This is in accordance with the reduced chance of a lucky guess at increasing numbers of options. Model 3 requires the lowest number of trials, which can be explained by the favourable effects of non-uniform plausibility values. Model 1 requires the most trials as it does not account for knowledge-based decisions.

3.4 Study 4: The influence of prior knowledge on the score

So far, we have assumed that the probability to know the correct decision $P(K)=0.25$. The same value was used to eliminate an option ($P(R_1)$ cf. equation (7)). In this study we have varied $P(K)$ (and $P(R_1)$ accordingly) from 0.0 to 0.5 with 0.05 increments, while every time running the linear baseline set-up to determine the mean scores. [Fig. 8] shows how the scores vary with $P(K)$.

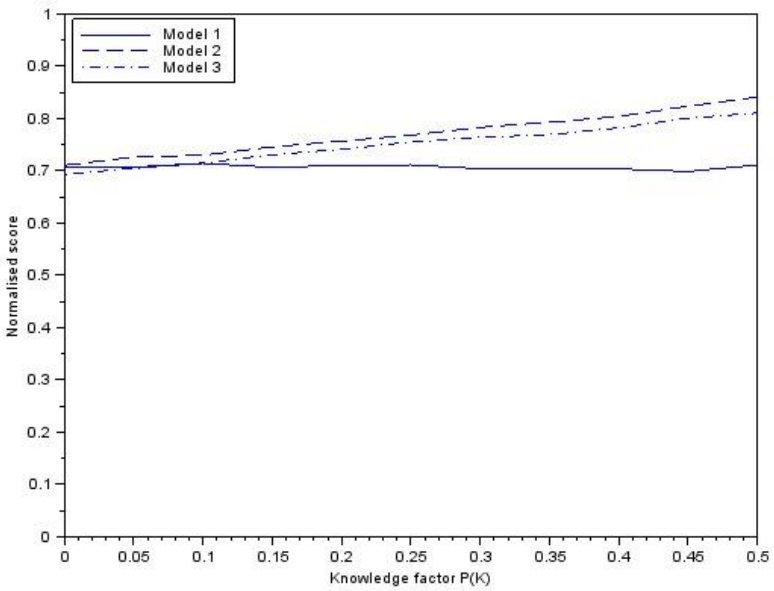

Figure 8: Scores as a function of the knowledge factor P(K) for the three models.

Differences between the three models gradually increase with larger *P(K)*. Eventually, the effects are substantial, showing higher scores in Model 2 and Model 3. This can be largely attributed to the many first hits that yield the full score of 1 point in these models.

3.5 Study 5: The influence of plausibility values on the score

So far, the plausibility values in the equidistant baseline model (B, D) have been kept unchanged. In this final study we used the linear baseline set-up while applying eight different linear distributions of the plausibility covering different ranges, as indicated in [Fig. 9], using the lower bound w_{min} of the plausibility as a parameter, while keeping the upper bound w_{max} fixed at 0.9..

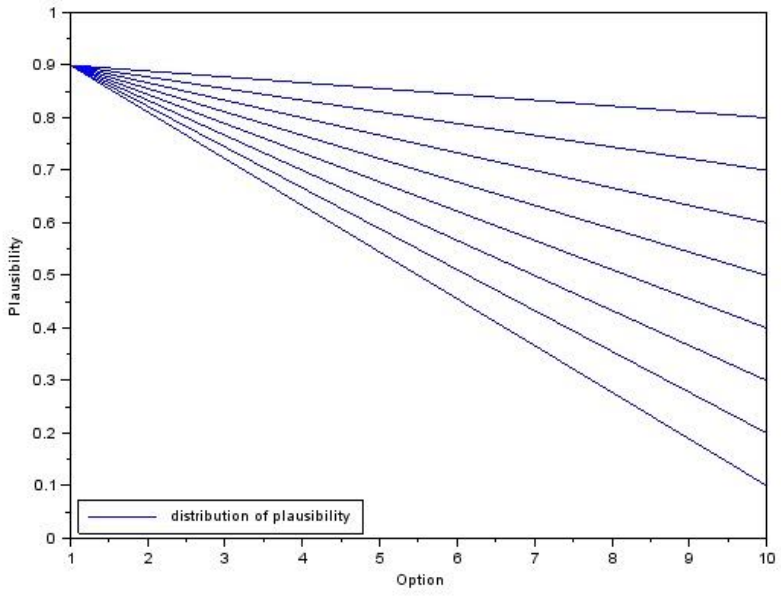

Figure 9: Eight distributions of plausibility over the nodes' options.

[Fig. 10] shows the scores for these eight plausibility distributions against the minimum plausibility bounds on the horizontal axis. The light curves correspond with knowledge probability P(K)=0.25. As a result, Models 2 and 3 show higher scores, in accordance with the findings in Study 1 and Study 2. To isolate the effects of plausibility differences from the effects of knowing the correct answer and knowing which alternatives to eliminate (Models 2 and 3), the dark lines in [Fig. 10] represent the scores for *P(K)=0*. It is can be concluded that the resulting scores in [Fig. 10] for *P(K)=0* show only minor differences given the sets of plausibility values of [Fig. 9]. This means that the effects of plausibility differences is small compared to random contributions from *P(K)* and the trial-based score reductions that hold for all models.

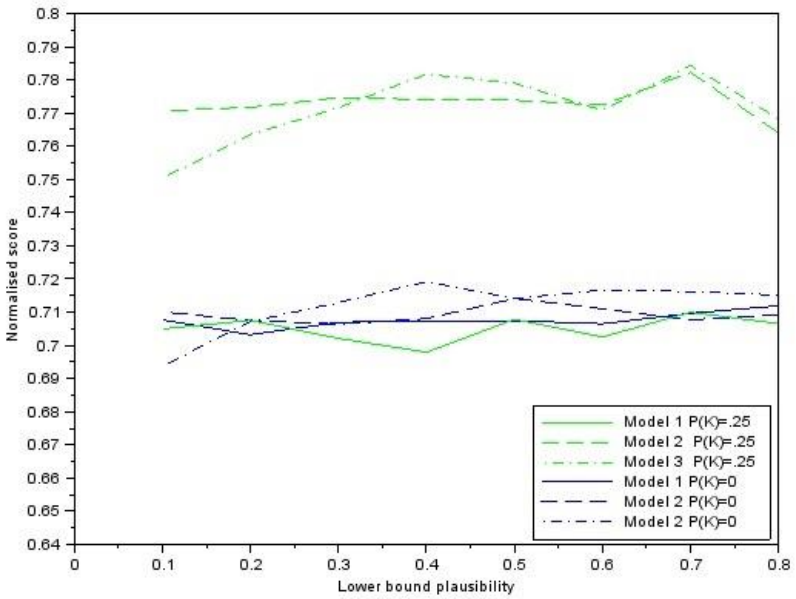

Figure 10: Scores as a function of eight different plausibility distributions from [Fig. 9].

4 Discussion and Conclusion

The studies presented above have demonstrated that the number of trials has a strong impact on the performance scores to be assigned: scores should be corrected for the random effects that increase by every trial. Neglecting the number of trials, which is quite common in serious games, leads to inaccurate scores that significantly overrate the observed performances, occasionally up to 100% or even more. The effects depend on the number of decision options, the presence of obvious, correct or incorrect options given the player´s knowledge level and, to a lesser extent, different plausibility levels of the options to decide upon. Biggest corrections are required for Model 1. Model 2 and Model 3 account for a reduction of the random component, but they make big demands to the game design, in particular with respect to estimating the odds on first hits and eliminations, and the need the assign plausibility values. From a practical perspective, Model 1 is the most favourable approach as it can be easily implemented by including the options in a node and the number of trials (see equation (4)).

Some shortcomings of the models should be considered. First, the models assume that decision taking is a dichotomous and single select process. However, in many training environments multiple solutions are allowed, often rewarded with non-binary,

partial-credit scores. In such polytomous scoring cases, the calculation of the Random Guess Scores becomes more complex [Draaijer et al., 2018]. Second, the models neglect the possibility that the player may never succeed at passing each node. If so, a decent serious game would provide guidance, hints or feedback, which in the end should be incorporated in the scores. Third, it was assumed that players, never make the same mistake again. This is a simplification that may not be applicable in all situations. Fourth, gaming the system is a common style of playing, where players try to circumvent the obstacles and challenges in the game or deliberately make mistakes to explore their effects in the game [Baker et al., 2004]. Such players do not act as regular players that want to address the game challenges seriously and learn from these, but rather aim to discover the bypasses in the game. Consequently, their assessments and learning fail. Fifth, although each serious game design should be tuned to the knowledge level of its target group, it will be hard to accurately estimate the probability *P(K)* for each player separately. Even though it may be wise to include some easy challenges (exposing obviously correct and/or incorrect alternatives) for motivational purposes, they should best be excluded from the performance measurement. Finally, defining the plausibility values may be a difficult task, because the plausibility of options as perceived by players may be dependent on the preceding events and their individual trajectories in the game. Still, in most cases the effects of plausibility differences has shown to be small.

Altogether, this study has shown that incorporating the number of trials in the score assignment is essential, and that a formula as simple as equation (4) from Model 1 offers an efficient solution that compares with more complex solutions from Model 2 and Model 3. A final comment is that the considerations, although they focussed on serious games, have significance in the wider field of multiple choice testing and assessment.

References

[Abt, 1970] Abt, C.: “Serious games”; New York, NY; Viking Press (1970).

[Aldrich, 2009] Aldrich, C.: “The Complete Guide to Simulations and Serious Games: How the Most Valuable Content Will Be Created in the Age Beyond Gutenberg to Google”; San Francisco, Pfeiffer (2009).

[Baker et al., 2004] Baker, R.S., Corbett, A.T., Koedinger, K.R., Wagner, A.Z.: “Off-Task Behavior in the Cognitive Tutor Classroom: When Students Game The System”. In: Proc. ACM CHI 2004: Computer-Human Interaction, 383-390 (2004)

[Bellotti et al, 2013] Bellotti, F., Kapralos, B., Lee, K., Moreno-Ger, P., Berta, R.: “Assessment in and of Serious Games: An Overview”; Advances in Human-Computer Interaction, 1, 1 (2013), 1- 11. http://www.hindawi.com/journals/ahci/2013/136864/#B18 , https://doi.org/10.1155/2013/136864

[Boston, 2002] Boston, C.: “The concept of formative assessment”; Practical Assessment, Research & Evaluation, 8, 9 (2002); https://scholarworks.umass.edu/pare/vol8/iss1/9/ , https://doi.org/10.7275/kmcq-dj31

[Chin at al., 2009] Chin, j. Dukes, R., Gamson, W.: “Assessment in simulation and gaming: a review of the last 40 years”; Simulation & Gaming, 40, 4 (2009), 553–568. https://doi.org/10.1177/1046878109332955

[Csikszentmihalyi, 1991] Csikszentmihalyi, M.: “Flow: The psychology of optimal experience”; New York, Harper Perennial (1991).

[Connolly et al., 2012] Connolly, T.M., Boyle, E.A., MacArthur, E., Hainey, T., Boyle, J. M.: “A systematic literature review of the empirical evidence on computer games and serious games”; Computers and Education, 59, 2 (2012), 661–686.

[Draaijer et al., 2018] Draaijer, S., Jordan, S., Ogden, H.: “Calculating the Random Guess Score of Multiple-Response and Matching Test Items”; in Ras, E., Guerrero Roldán, A. (Eds.): “Technology Enhanced Assessment”. TEA 2017. Communications in Computer and Information Science, 829, 210-222; Cham, Springer (2018). https://doi.org/10.1007/978-3-319-97807-9_16

[Eggen et al., 2006] Eggen, T.J., Verschoor, A.J.: “Optimal testing with easy or difficult items in computerized adaptive testing”; Applied Psychological Measurement, 30, 5 (2006), 379-393.

[Gee, 2003] Gee, J.P,: ”What Videogames Have to Teach Us About Learning and Literacy”; New York, Palgrave Macmillan (2003).

[Gronchi et al., 2021] Gronchi, G., Sloman, S.A.: “Regular and random judgements are not two sides of the same coin: Both representativeness and encoding play a role in randomness perception”; Psychon. Bull. Rev. 28, 1707–1714 (2021), https://doi.org/10.3758/s13423-021-01934-9

[Guttormsen Schär et al., 2000] Guttormsen Schär, S., Schluep, S., Schierz, C., Krueger, H.: “Interaction for Computer-Aided Learning”; Interactive Multimedia Electronic Journal of Computer-Enhanced Learning 2, 1 (2000). http://imej.wfu.edu/articles/2000/1/03/

[Klinkenberg et al., 2011] Klinkenberg, S., Straatemeier, M., Van der Maas, H.L.J.: “Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation”; Computers & Education, 57, 2 (2011), 1813-1824.

[Lord, 1980] Lord, F.M.: “Applications of item response theory to practical testing problems”; Hillsdale, NJ: Lawrence Erlbaum (1980).

[Maseleno et al., 2018] Maseleno, A., Sabani, N., Huda, M, Ahmad, R., Jasmi, K.A., Basiron, B.: “Demystifying Learning Analytics in Personalised Learning”; International Journal of Engineering & Technology, 7, 3 (2018), 1124-1129.

[Nyamsuren et al., 2017] Nyamsuren, E., van der Vegt, W., Westera, W.: “Automated Adaptation and Assessment in Serious Games: a Portable Tool for Supporting Learning”; in Winands, M., van den Herik, H.J., Kosters, W. (Eds.), “Proceedings of the Fifteenth International Conference on Advances in Computer Games”, ACG 2017: 15th International Conferences, ACG 2017, Leiden, The Netherlands, July 3–5, 2017, Revised Selected Papers. Lecture Notes in Computer Science, 10664, 201-212; Cham, Springer International Publishing AG (2017).

[Redeker et al., 2012] Redeker, C., Punie, Y., Ferrari, A.: “eAssessment for 21st century learning and skills”; in Ravenscroft, A., Lindsteadt, S., Kloos, C.D., Hernandez-Leo, D. (Eds.): “21st Century Learning for 21st Century Skills”; Proceedings of the 7th European Conference on Technology-Enhanced Learning EC-TEL, 292-305; Heidelberg, Springer (2012).

[Shute et al., 2009] Shute, V., Ventura, M., Bauer, M., Zapata-Rivera, D.: “Melding the power of serious games and embedded assessment to monitor and foster learning: flow and grow”; in Ritterfeld, U., Cody, M., Vorderer, P. (Eds.): “Serious Games: Mechanisms and Effects”, 295–321; Mahwah, NJ, USA, Routledge, Taylor and Francis (2009).

[Shute et al., 2013] Shute, V.J., Ventura, M.: “Measuring and supporting learning in games: Stealth assessment”; Cambridge, MA: The MIT Press, 2013.

[Spearman, 1904] Spearman, C.: The proof and measurement of association between two things”; American Journal of Psychology, 15 (1904), 72-101.

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

[Weiss et al, 1987] Weiss, D.J., Vale, C.D.: “Adaptive Testing”; Applied Psychology 36, 3-4 (1987), 249-262. https://doi.org/10.1111/j.1464-0597.1987.tb01190.x

[Westera, 2015] Westera, W.: “Games are motivating, aren´t they? Disputing the arguments for digital game-based learning”; International Journal of Serious Games, 2 (2015); Online publication at http://journal.seriousgamessociety.org/index.php/IJSG/article/view/58

[Westera, 2018] Westera, W: “Simulating serious games: a discrete-time computational model based on cognitive flow theory”; Interactive Learning Environments, 26, 4 (2018), 539-552. https://doi.org/110.1080/10494820.2017.1371196[Westera, 2019] Westera, W.: “Why and How Serious Games can Become Far More Effective: Accommodating Productive Learning Experiences, Learner Motivation and the Monitoring of Learning Gains”; Educational Technology & Society, 22, 1 (2019), 113–123.