Improving the scoring of Serious Educational Games as Assessments

Walsh, Clare <cew2g15@soton.ac.uk>, University of Southampton, UK,
Bokhove, Christian, University of Southampton, UK, <C.Bokhove@soton.ac.uk> and
White, Su, University of Southampton, UK, <su@soton.ac.uk>

Key words: Assessment Big Data Serious Educational Games Scoring Validation

Abstract.
The demand for alternative forms of testing is growing. The 21st Century Initiative highlighted a growing desire from industry to introduce modern workplace skills in the school curriculum. Complex or procedural problem solving, systems management and collaboration are in high demand in the workplace. However, until these skills can be tested, their impact on education is likely to be minimal. Serious Educational Games (SEGs) provide a format to carry out these kinds of complex tasks, but we do not yet know how to score gaming data fairly. The way games are currently scored has been influenced by games industry practices which do not have to meet the same levels of accountability for accuracy as formal testing. The data produced during gameplay is unlike the kinds of data that assessors are used to analyzing. Without a convincing argument for the technical adequacy of the test from an assessment perspective, we cannot make a serious case for their widespread use in education. This paper identifies key issues around the conceptualization of missingness, time and iteration affecting the scoring through a case study of an educational gaming data set. It also provides an initial estimation of the fairness of the game scores.

Games provide rich evidence of learning activity, but this is not necessarily the same as evidence of participants' learning [1] and we need to develop methodologies to identify and separate the two. Within the field of games analytics, alternative approaches to scoring cognitive skills in Serious Educational Games (SEGs), such as Bayesian Analysis and other types of Machine Learning have been proposed, as they are able to deal with large data sets, missing data, co-dependent variables and data that changes dynamically [2]. In Bayesian terms, 'fairness' is often reduced to an estimated overall likelihood that final derived score is accurate, but the errors may have occurred before that stage.

1.1 The Background

Classical Test Theory established the idea that any test-derived score reflects the true score plus an error term [3]. Designing fair tests is very challenging. The first barrier to accurate assessment is the problem of difficulty of the tasks. In games analytics, ability is often assessed through a frequency tally of tasks that have been completed correctly. Such an approach requires the test designer to assume that the increase in difficulty for each task was exactly the same every time. This is simply not possible. The leap in cognitive ability will vary from barely any increase in challenge, to a significant jump. 'Difficulty' cannot be guessed by looking at the task, but it can be measured as a property of the task and the player's interaction with it.

The psychometric approaches to estimate difficulty have been developed over many decades, but gained more widespread application in mainstream testing in the 1990s, once computational analysis made uncovering error terms more practical. Rasch introduced a Log-Odds approach to modelling score data [4]. By distributing the scores around a mean of zero, a probability-based Log-odds estimate converts any raw score, which is ordered but randomly spaced data, into ordered interval data. In other words, the spaces between learners' scores are no longer random. It also creates an estimation of difficulty and ability that is independent of the assessment tool. This independence criteria allows the test designer to remove or add tasks with minimal impact on scores, which is important for test design purposes.

A value of difficulty for each item can be escalated to the model in many Bayesian software applications. An ordered value for these items could be established inductively, in response to the observed outcomes, in much the same way as Rasch modelling. Alternatively, it could be deductively established, or assigned in a similar way once a difficulty value for an item has been anchored, or fixed, in Rasch terms. The Bayes models in software often instantiate those values, or update them, in response to observed outcomes [2]. This still does not address the concern about intervals or that there may be considerable bias in the original data set, though, which practitioners using more established Rasch approaches have experience of uncovering.

Much of the bias comes from the fact that test writers are not always good at writing fair questions, and test-takers do not always behave predictably. Problems in test design include but are not limited to lucky guessing, poor task design and wording or time management issues. Once standardized scores for the difficulty of the task and the ability of the learner have been identified, a pattern of predicted behavior emerges. If Question 13 had a difficulty level of '+2.5' and Student B had an ability level of '+3', we would expect her to get the answer correct. An unexpected behavior leaves a residual of -0.5 between the expected and observed behavior. If several test-takers with similar ability to B get the answer wrong, it suggests we are not measuring cognitive ability, but something else. By producing Chi-Square estimates of the sum of all of those residuals, unexpected behavior during the testing process can be uncovered, and the numerical estimation of the location and size of the problems is very

helpful in fixing issues. Something similar to this figure could be produced in the Bayes model, if it were based on interval data.

This leaves the question of how to anchor values, which is more within the sphere of the test design. Even though psychometric techniques such as Rasch modelling have been applied to a wide range of testing scenarios, from multiple-choice to human ratings of complex skills, they do not readily adapt to gaming data. Psychometric measures function best with dichotomous or partial scoring of conditionally independent variables [2]. This is not the kind of data produced in a dynamic hypertext online game. Most widely-used high stakes tests currently delivered online are heavily reliant on discrete questions presented in a linear form, reflecting their pen and paper equivalent tests. In a 2012 paper, Mislevy et al, reflected on their experience of scoring a simulation city management game. They observed that the total scores often reflected choices the learners had made, but they also represented multiple attempts, the use of clues or hints, wrong-turns, situationally dependent variables (such as different time limits) and co-dependent actions [5]. Gaming log files often contain additional para data, such as timestamps or details about iterations. This data may offer insight into the situationally-dependent states of the test, but there is no common agreement even on how these are to be conceptualized. This study looks at incorporating conceptualizations of three key factors into a Games scoring model: Dealing with missing data; Integrating a response time into the score; and Identifying which iteration to record. This is a case study of gaming data from a game that assesses primary and secondary maths skills.

1.2 The Data Set

The data set was provided by Blue Duck Education, from their Manga High games (URL: https://www.mangahigh.com). The original log file set contained over a billion data points for over a million learners around the world at the time of extraction. The games are in a single-user and multi-player environments, directed at the 8-16 age range, and are aligned to British Key Stages 2 and 3 mathematics outcomes, and US Common Core outcomes for that age group. Students select games and carry out simple tasks using mathematical knowledge and the game mechanics. The games have a rich, colorful interface, with sound and animation. Interactions average around 3 minutes. A convenience sample of the 80 players was randomly selected from the players with the top 90-95% activity. This group was chosen to provide the richest evidence of learning, while excluding possible non-targeted learners, such as the games developers themselves.

1.3 Missing data

The assumption that the same version of the assessment was delivered to all learners simply does not hold in games. When students are allowed to self-select them-

selves into certain pathways of gameplay, it generates very large amounts of missing data. Even after significant cleaning up of the data to include the most active players in the data set and to record a maximum of 2 iterations per player, out of total of 3,552 possible observations, 1,381 were missing.

In the initial analysis, in estimating the difficulty of the tasks using Rasch approaches to produce interval data. The missing data was treated as not administered, and so 'not presented' rather than 'incorrect'. It is an approach first suggested by Miselevy and Wu in 1996 [6] and expanded upon by Ludlow and O'leary [7]. The log-odd estimate of task difficulty was based on the maximum number of points for the games attempted. The original intention was to anchor or fix those values before any further analysis. However, an anomaly emerged in that some of the games had 'lite' versions. The 'lite' version used the same game mechanics, but the mathematics were easier. As can be seen in Table 1, the 'lite' games appear to be actually as hard or harder than the regular versions of two games, Sigma Prime and Sundae Times.

Table 1 Table comparing the initial difficulty estimates for 'lite' and 'regular' versions of the same game, based on data from the full sample of test-takers

Game Name	**Logit score of difficulty***	
	Lite	Regular
Bidmass Blaster (order of operations)	-0.6	+0.6
Sigma Prime (factorization with multiplication and division)	+0.2	+0.2
Pyramid Panic (geometry)	-0.2	+2.8
Sundae Times (times tables)	+0.6	+0.2

** Higher score represents greater challenge*

This is an early indication that learners' ability to self-select themselves into the version they feel is appropriate may be distorting the scores. It may not be that the game was easier, but that the leaners who played it had lower ability. It suggests that a more expansive interpretation of missing data may be needed, such as basing difficulty estimates on values obtained through subsets of learners with similar characteristics.

1.4 Response time

Response time has traditionally been fixed, albeit somewhat arbitrarily, by imposing a blanket time limit to complete the whole test. How learners manage their time has been seen as influential on the end result, but out-of-scope of the assessment model, or captured indirectly in the score. It may be harder to make such assumptions in game play, particularly where response time is seen as key to separating performance in games.

There has been some research on response time carried out in the field of assessment. Van der Linden, in a paper in 2009, pointed out that the correlation between time and ability is not necessarily linear [8]. What we are in fact measuring, is speed, which is a property of the task, the learner and the completion time. Tasks can be no less challenging cognitively, but simply require more stages in order to complete them, and, therefore, more time. It is highly plausible, for example, that the shortest play time with the correct outcome represents luck. Figure 1 shows the distribution of response times for the four band scores, 0, 1, 2 and 3. In the initial analysis, a large number of scores of '0' have playtimes of under 10 seconds. This is most likely to be browsing behavior, rather than a serious attempt at the task.

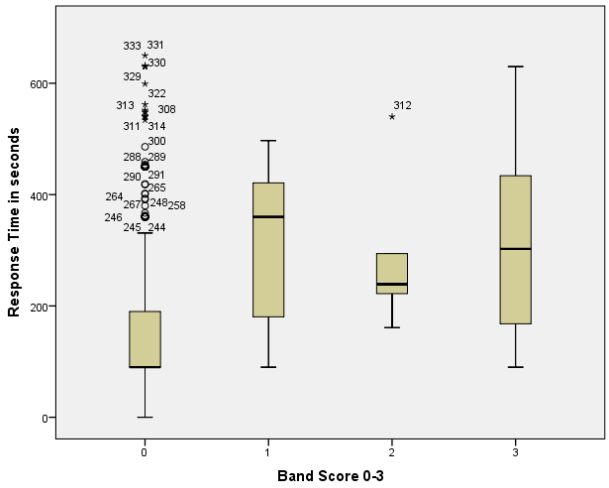

Figure 1 Mean distribution of response times for Sundae Times for the four bands, 0, 1, 2 and 3

There is also considerable difference in the patterns of response time for the other three scoring bands (1-3). This suggests response time of 300 seconds would be fast for Band 1, slow for Band 2 but about average for Band 3 in this particular game. It appears that speed is not just a property of the test-taker, time and task, but the scoring band as well. Given this variation, it may also be preferable to report speed on tasks with scores ≥1 as a separate dimension to performance in ability, rather than bring the two scores back together to create on overall score, as Van der Linden does at the end of his analysis [8].

1.5 Iterations

Very little has been written on the subject of iterations, but it can impact the result. Games encourage young leaners to repeat, and this case, one particular multi-player platform game, Jet Stream Riders, encouraged a very large number of iterations among the young players.

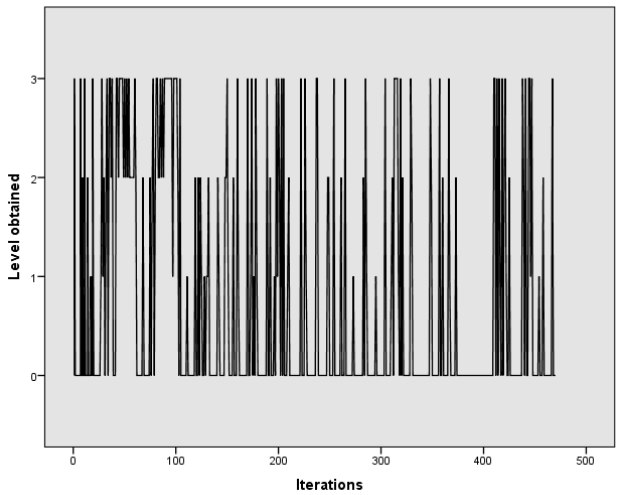

Figure 2 Line graph showing the results of gameplay for one test-taker in one game, Jet Stream Riders

Figure 2 shows one child's results from over 469 iterations of gameplay of the game Jet Stream Riders. All four scores (0-3) are present, but it does not obviously show an upward trend of improvement. Taking the same child's performance, Figure 3 shows how the scores were distributed on a normal curve.

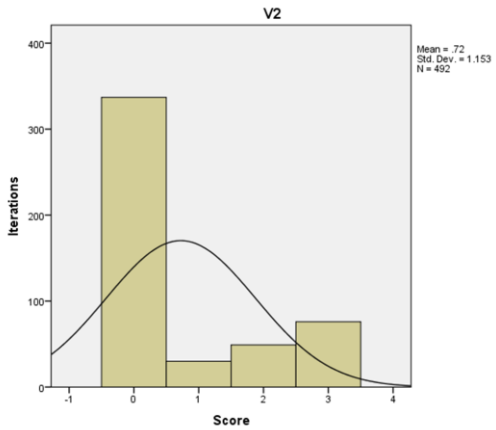

Figure 3 Average scores for the same child over the 469 iterations of Jet Stream Riders

If the median is taken, then the child scores ‘0’. If the mean is taken, the score lies at Cos .72, or within Band ‘1’. If the upper quartile limit is taken, in order to eliminate outlying behaviors, the score lies in Band ‘2’. Finally, if the highest score is taken, the score is in Band ‘3’. Taking just the highest score for all players of this particular game produced relatively stable Infit and Outfit scores of 1.16 and .95 respectively for this task. Whether this conceptualization is desirable may need to be decided based on other aims of the game scoring process, such as value judgments on whether perseverance is to be rewarded.

The results of this preliminary study show that there is a stability in this gaming data set, with overall Infit and Outfit scores of 1.04 and 0.88 respectively from the initial analysis. However, further investigation may well show that these statistics are sensitive to changes in the scoring design. Some problems may also need to be addressed, not by the mathematical model, but by changes in the rules, such as restricting the number of iterations, or not allowing browsing behaviors. There is considerable collaborative work to be done to bring the communities of games designers and assessors together.

This project is part of a PhD Thesis in Web Science, funded by a Digital Economy Network grant, within the Web Science Doctoral Training College at the University of Southampton. It is an interdisciplinary study co-supervised by the Education Department and the Electronics and Computer Science Department.

References

[1] R. J. Mislevy *et al.*, "Psychometric considerations in game-based assessment," *GlassLab Report*, 2014 (page 13).

[2] R. G. Almond, R. J. Mislevy, L. S. Steinberg, D. Yan, and D. M. Williamson, *Bayesian networks in educational assessment*. Springer, 2015 (pages 162-166 and pages 14-16).

[3] F. M. Lord, *Applications of item response theory to practical testing problems*. Routledge, 1980.

[4] T. Bond and C. M. Fox, *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge, 2015.

[5] R. J. Mislevy, J. T. Behrens, K. E. Dicerbo, D. C. Frezzo, and P. West, "Three things game designers need to know about assessment," in *Assessment in game-based learning*: Springer, 2012, pp. 59-81.

[6] R. J. Mislevy and P. K. Wu, "Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing," *ETS Research Report Series*, vol. 1996, no. 2, 1996.

[7] L. H. Ludlow and M. O’leary, "Scoring omitted and not-reached items: Practical data analysis implications," *Educational and Psychological Measurement*, vol. 59, no. 4, pp. 615-630, 1999 (pages 616-617).

[8] W. J. Van Der Linden, "Conceptual issues in response-time modeling," *Journal of Educational Measurement*, vol. 46, no. 3, pp. 247-272, 2009.