Computers & Education 83 (2015) 57–63

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Gamification in assessment: Do points affect test performance?

Yigal Attali [a, *], Meirav Arieli-Attali [b, 1]

[a] Educational Testing Service, Rosedale Rd. MS-10-R, Princeton, NJ 08541, United States
[b] Educational Testing Service, Rosedale Rd. MS-16-R, Princeton, NJ 08541, United States

ARTICLE INFO

Article history:
Received 24 October 2014
Received in revised form
16 December 2014
Accepted 18 December 2014
Available online 27 December 2014

Keywords:
Gamification
Assessment
Performance
Engagement

ABSTRACT

Gamification, applying game mechanics to nongame contexts, has recently become a hot topic across a wide range of industries, and has been presented as a potential disruptive force in education. It is based on the premise that it can promote motivation and engagement and thus contribute to the learning process. However, research examining this assumption is scarce. In a set of studies we examined the effects of points, a basic element of gamification, on performance in a computerized assessment of mastery and fluency of basic mathematics concepts. The first study, with adult participants, found no effect of the point manipulation on accuracy of responses, although the speed of responses increased. In a second study, with 6–8 grade middle school participants, we found the same results for the two aspects of performance. In addition, middle school participants' reactions to the test revealed higher likeability ratings for the test under the points condition, but only in the first of the two sessions, and perceived effort during the test was higher in the points condition, but only for eighth grade students.

1. Introduction

Increasingly, web-based and mobile applications look to gamification, the use of game design elements (e.g., points, leaderboards, and badges) in nongame contexts to promote user engagement (Deterding, Sicart, Nacke, O'Hara, & Dixon, 2011). Gamification relies on the argument that many traditional activities (including school activities and traditional learning) are not inherently interesting, that games, especially computer-games, are "fun," and therefore introducing game-like features into these otherwise dull activities would make them more attractive (McGonigal, 2011; Zichermann & Linder, 2010).

A typical example of embedding educational activities within a game-like environment is the DimensionU™ series of math games. In one of these games (TowerStorm), a virtual character (avatar) immersed within a 3-D environment "retrieves" a multiple-choice question from the "fountain of knowledge," answers the question to collect a colored ball (each answer option has a different color), runs to a tower that generates stacked colored rings, and shoots the ball at a ring on the tower. Feedback about the correctness of the answer is provided when the ball hits the tower. If the answer was correct, the character will earn points.

Although it is intuitively clear that games are a strong motivating factor for students (Gee, 2003; Shaffer, 2006), there remains a dearth of research in which the effectiveness of the gaming environment has been directly compared with a traditional computerized application (Jackson & McNamara, 2013). For example, Kebritchi, Hirumi, and Bai (2010) investigated the effects of using the DimensionU™ series of mathematics games on high school students by assigning a treatment group to play the games for 30 min each week for 18 weeks. The treatment group showed higher gains in scores on a standardized mathematics achievement test. Unfortunately, the control group was not assigned to perform the same educational activities in a nongame environment, making it difficult to disentangle the gaming effect from the practice effect (Ericsson, Krampe, & Tesch-Römer, 1993).

In contrast, Papastergiou's (2009) control (nongame-like) group of high school students was exposed to the same content (2 h of instruction on computer memory) as the treatment (game-like) group. Nevertheless, the treatment group showed higher gains in performance from a pretest to the same posttest measuring knowledge of the content covered in instruction. In another recent study, Jackson,

* Corresponding author. Tel.: +1 609 734 1747.
E-mail addresses: yattali@ets.org (Y. Attali), mattali@ets.org (M. Arieli-Attali).
[1] Tel.: +1 609 734 5256.

http://dx.doi.org/10.1016/j.compedu.2014.12.012
0360-1315/

Dempsey, and McNamara (2012) compared a traditional tutoring system environment for reading strategy training with a counterpart game-based system which used the same task. Results indicated that performance of participants (college students) in the game-like condition was significantly lower than in the traditional environment, although participants in the game-like environment rated it as more engaging.

The purpose of this study was to test the effects of one particular game design element, by experimentally manipulating whether or not students were accumulating points while they were completing a mathematics assessment.

From a theoretical perspective, points provide feedback to the student. Providing feedback regarding task performance is one of the most frequently applied psychological interventions (Kluger & DeNisi, 1996). To have a positive effect on learning, feedback needs to provide information related to the task or process of learning (Sadler, 1989). It can do so through a number of different cognitive processes, including restructuring understandings, confirming to students that they are correct or incorrect, and/or indicating alternative strategies to understand particular information. Alternatively, it can operate through affective processes, such as increased effort, motivation, or engagement (Hattie & Timperley, 2007). However, despite a huge literature (Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Mory, 2004), the specific mechanisms relating feedback to performance are still not well understood. Historical reviews and meta-analyses on the subject describe the findings as "inconsistent," "contradictory," and "highly variable" (Azevedo & Bernard, 1995; Kluger & DeNisi, 1996). Moreover, in a comprehensive meta-analysis, Kluger and DeNisi found that although feedback interventions improve performance on average, they reduce performance in more than one third of the cases.

In and of themselves, points and other game-like elements provide information about success in the task. However, the gamification argument emphasizes the motivational aspect of game design elements over the possible cognitive or informational aspects. There are a number of motivational theoretical constructs that have been shown to mediate the effect of feedback and could be relevant to the points manipulation. In particular, the concept of locus of attention is of interest in this respect (Butler, 1987). Properties of feedback can direct attention to the self or to the task, and attention to self has been shown to attenuate or even reverse the effects of feedback (Butler, 1987) because it interferes with task performance. It is not clear whether points would be perceived by students as information about task performance (and then attention is more likely to be directed to the task) or as a form of prize for good performance (and then attention may be directed to the self).

The difference between intrinsic and extrinsic motivators is another theoretical distinction that may shed light on the effect of game-like features. In other words, points and badges can be seen as extrinsic rewards for performing the task. However, although early research has demonstrated the power of extrinsic rewards in controlling behavior (Skinner, 1953), later research has shown that for intrinsically motivated activities, tangible rewards may undermine this intrinsic motivation (Deci, Koestner, & Ryan, 1999). That is, rewarding students for performing well in an educational setting may be counterproductive in the long run. However, even in the short run, research on changing incentives (either by increasing personal stakes or providing external rewards) as a way to increase student performance is inconclusive.

On the one hand, past research supports the assumption that tests with no personal consequences, that is, low-stakes tests, are associated with a decrease in motivation and performance. Wise and DeMars (2005) reviewed 12 studies that included 25 comparisons between the performance of motivated and less-motivated groups of test takers. In most comparisons, the performance of the motivated group was significantly higher than that of the less-motivated group, and the average effect size was around .6. The manipulations used to motivate test takers varied. In most cases, the students were told the scores would count towards course or school grades. In a recent study, Liu, Bridgeman, and Adler (2012) found that increasing the personal stakes of a test by telling test takers that their scores could be released to faculty in their college or potential employers significantly increased test performance, with an effect size of .41.

On the other hand, studies that used monetary incentives to motivate students in tests have generally found weak effects on performance. O'Neil, Sugrue, and Baker (1996) offered 8th and 12th grade students taking the National Assessment of Educational Progress (NAEP) 1 dollar for each item they answered correctly, and their performance was compared to control groups that received standard NAEP instructions. A significant effect was found only for a subsample of 8th graders (those who correctly remembered their treatment condition), and only on the easy items. More recently, this study was replicated (O'Neil, Abedi, Miyoshi, & Mastergeorge, 2005) with 10 dollars per item, but again no effect on scores was found. Baumert and Demmrich (2001) offered to one group of students a flat payment if they answered more items than expected based on their school grade, but failed to find an effect on scores. Finally, Braun, Kirsch, and Yamamoto (2011) administered a NAEP assessment to 12th grade students and found weak effects (effect size of .08−.25) for monetary incentives (either a fixed amount of 20 dollars or variable according to performance) compared to a control group.

Taking into account the scarcity of research on the effectiveness of gamification and the inconclusive results in related lines of research, in the studies described below we sought to evaluate the effects of one particular gamification feature, points, on different aspects of performance in the context of an educational assessment. The assessment that was used in the two studies focuses on mastery and fluency of basic mathematical concepts and is itself part of the *CBAL*™ (Cognitively Based Assessment *of*, *for*, and *as* Learning) research initiative[2] to develop assessments that maximize the potential for positive effects on teaching and learning (Arieli-Attali & Cayton-Hodges, 2014; Bennett, 2011). The two studies reported in this paper were part of the development and piloting of this assessment. At the heart of the mathematics curriculum in the early years of elementary and middle school are concepts of the number system and operations with numbers. Research has shown that proficiency with numbers and numerical operations is an important foundation for further education in mathematics (National Mathematics Advisory Panel, 2008). Constraints on cognitive capacity provide one explanation for this connection. The more automatically a procedure is executed, the less mental effort is required, and this enables complex tasks to be carried out more efficiently (Case, 1985). Inefficient estimation and mental computations often lead to declarative and procedural errors, which carry on and impede subsequent problem solving (Cumming & Elkins, 1999). There are also important conceptual continuities between whole and rational number concepts and concepts of algebra (Empson, Levi, & Carpenter, 2011).

Since performance on an assessment of basic mathematical concepts could be determined by both accuracy and speed of the response, the instructions and point systems that were used in these studies reflected these two aspects with different emphases. All participants in

[2] See also CBAL website http://www.ets.org/research/topics/cbal/initiative.

these studies (adults in study 1 and middle school students in study 2) were encouraged to answer the questions as quickly and accurately as possible. However, some participants received points for accurate and speedy responses. The points accumulated were prominently displayed in the top-right portion of the screen with a flip counter. Each time a participant in the points condition answered a question correctly, the system would provide notice of how many points were gained and the point counter would increase appropriately in an animated way. Analyses were carried out to determine the possible effect of points on performance (both accuracy and speed), as well as meta-cognitive reactions to the test.

2. Study 1

As part of initial piloting of the assessment, adult participants completed a single test session under one of two points conditions or a control condition. The two points conditions emphasized either accuracy (by providing more points for accurate than speedy responses) or speed (see below for specific instructions). Analyses focused on the possible effect of the points manipulation on performance, both accuracy and speed. Sex was also considered in the analyses of the points manipulation. In many studies girls and women are reported to display less interest in digital games, have less game-related knowledge, and play less frequently and for shorter durations than do boys and young men (e.g., Hartmann & Klimmt, 2006). Moreover, when girls do play, they often prefer different games and for different reasons – females are more likely to list social reasons for playing and less likely to list achievement and immersion as important for their play (Yee, Ducheneaut, & Nelson, 2012). Therefore, it is important to take sex into account in an investigation of the effects of gamification.

2.1. Method

2.1.1. Participants

Participants for this study were recruited from Amazon.com's Mechanical Turk (MTurk) crowdsourcing marketplace, which allows researchers to post experiments to be completed by Amazon.com users in return for monetary compensation. This platform has seen a growing interest among researchers as a way to recruit subjects for social-science experiments (Buhrmester, Kwang, & Gosling, 2011). Participants, 1218 in all, were paid 2 dollars and completed the assignment in 25 min on average. All participants were US residents and their first language was English, their age varied from 18 to 74 years ($M = 32$, $SD = 11$), 46% were women, and most had at least some post-secondary education (13% were high school graduates, 30% some college, 10% associate degree, 38% bachelor degree, and 9% graduate degree).

2.1.2. Materials

Test items were based on item "models" (Bejar, 1993), schemas of problems with parameters that can be instantiated with specific values. For example, the model "$X + Y = ?$", where X and Y can be whole numbers in the range 1–10, has two parameters that can be instantiated to display an actual exercise. The assessment used in this study used 50 item models with 8 instances from each model, with a total of 400 items. Instances were generated to cover a representative range of model parameters. Some examples of items used are:

The fraction 4/7 is equivalent to 8/___
1/3 is ___ times larger than 1/6
Write the fraction 6/30 in simplest terms
Round 0.8644 to nearest tenth
The number 24 is divisible by which of the following numbers? Select all that applies: 2,3,6,9
The greatest common factor of 9 and 15 is ___

2.1.3. Design

The complete set of 400 items was divided into four nonoverlapping tests, each composed of four instances from 25 item models. Each participant was randomly assigned to one of the tests. Participants were also randomly assigned to one of three *points* feedback conditions. In the 1–10 points condition, participants were told they will receive 1–10 points for each correct answer, depending on the speed of their response. In this condition, speed was as important as accuracy – accurate but slow responses would not result in significant accumulation of points. In the 10 + 5 points condition, participants were told they will receive 10 points for each correct answer and extra points (0–5) based on the speed of their response. In this condition, accuracy is more important than speed – the contribution of speedy responses is at most half of what accurate responses can contribute. In addition, speediness is framed as a "bonus" over accuracy. For these two conditions, the points accumulated were prominently displayed in the top-right portion of the screen with a flip counter. An example of the feedback for a correct answer is presented in Fig. 1. The participant had 110 points before that response and received 14 points (10 for a correct answer and 4 extra points for speediness) for the response shown. A third group, a control, was not awarded points for correct answers.

2.1.4. Procedures

Participants were told they will be asked to answer simple mathematics problems. They were given instructions on how to record their responses (click on options or type numbers). All participants were encouraged to answer as quickly and accurately as possible. In addition, specific instructions about the points system were provided as appropriate.

For each participant, the order of the sections as well as the order of items within the section was randomized. Therefore, each item could appear in any position in the test, but instances from the same model could only appear once in every section.

The items appeared on the screen one after the other. Participants were instructed to use the Enter key to submit their answers, or click a "Submit" button. Following response submission, immediate feedback on the correctness of the response together with the correct response, was displayed (as shown in Fig. 1). Participants were then required to click a button to continue to the next item. Following each

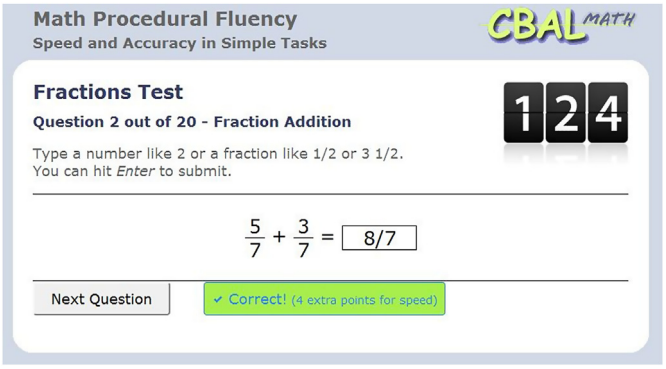

Fig. 1. Example of interface and feedback in 10 + 5 points condition.

section of the test, a message appeared advising participants they could take a break if they wanted to. Following the completion of the test, participants were asked to indicate their age, sex, and level of education.

2.2. Results

Table 1 presents psychometric properties of response accuracy and response time, or slowness. In all response time analyses the common transformation of the natural log of time in seconds to answer was used as this transformed measure has a more normal-like distribution than the raw time and is more appropriate for parametric tests. Overall, the items were easy and answered quickly. The average percentage correct score was 77% and average mean log response time was 2.06 (corresponding to 7.8 s). The items in test 1 were somewhat less difficult than those in test 2, and took less time to answer. Cronbach's alpha coefficients of internal consistency are reported for the 100 items as well as for the 25 item model scores (each score the sum of four item instances) in order to account for possible local dependency among the four instances of the same item model (by comparing the 100-item estimates to those for the 25-itemestimates; Sireci, Thissen, & Wainer, 1991). The internal consistency of response accuracy was high, with Cronbach's alpha coefficients of .93–.94 for the 100 item scores and .90–.91 for the 25 item model scores. The internal consistency of response time was even higher, with Cronbach's alpha coefficients of .98 for the 100 item scores and .97 for the 25 item model scores.

A series of ANOVAs was performed to examine the possible effects of the point manipulation and the increased emphasis on speed (from no points, to 10 + 5, to 1–10 points) on test performance and response time and the relationship between test performance and response time. In addition to the point conditions, sex and level of education (high school, partial college education, and college graduate) were included as independent variables. All interactions between independent variables were included in these analyses. Age was not included in these analyses because it did not show any relation with response accuracy ($r = .03$, $p > .05$).

A 3 (points condition) × 2 (sex) × 3 (education level) ANOVA was performed on test scores (the number of correct answers, out of 100 questions). A significant sex effect was found for test scores as the dependent variable, $F(1, 1200) = 33.33$, $p < .01$, with higher performance for male ($M = 79.1$, $SE = 0.6$) than for female participants ($M = 73.7$, $SE = 0.6$). A significant education effect was also found, $F(1, 1200) = 81.73$, $p < .01$, with decreasing performance from college graduates ($M = 80.5$, $SE = 0.5$) to partial college education ($M = 75.7$, $SE = 0.6$) and high school education ($M = 65.1$, $SE = 1.5$). Mean test performance was similar for the no points ($M = 77.5$, $SE = 0.7$), 10 + 5 points ($M = 76.4$, $SE = 0.6$), and 1–10 points ($M = 75.3$, $SE = 0.6$) conditions, and this effect was not significant, $F(2, 1200) = .54$, $p = .58$. None of the interactions between main effects were significant, $ps > .13$.

A 3 (points condition) × 2 (sex) × 3 (education level) ANOVA was performed on the average (natural) log of response times (in seconds). A significant sex effect was found on the log response time data, $F(1, 1200) = 28.83$, $p < .01$, with lower response times for male ($M = 1.99$, $SE = .01$, corresponding to 7.3 s) than female participants ($M = 2.14$, $SE = .01$, corresponding to 8.5 s). A significant education effect was also found, $F(1, 1200) = 10.17$, $p < .01$, with decreasing speed from college graduates ($M = 2.01$, $SE = 0.1$) to partial college education ($M = 2.10$, $SE = 0.2$) and high school education ($M = 2.14$, $SE = 0.3$). A significant points effect was also found, $F(2, 1200) = 6.12$, $p < .01$. Post-hoc Tukey tests showed that mean response time was higher for the no points ($M = 2.13$, $SE = .02$, corresponding to 8.4 s) than for either the 10 + 5 points ($M = 2.04$, $SE = .02$, corresponding to 7.7 s) or 1–10 points ($M = 2.03$, $SE = .02$, corresponding to 7.6 s) conditions. None of the interactions between main effects were significant, $ps > .11$.

The relationship between test performance and response time was similar in the no points ($r = -.31$), 10 + 5 points ($r = -.31$), and 1–10 points ($r = -.40$) conditions; the differences between these correlations were not significant ($ps > .14$).

Table 1
Psychometric properties of response accuracy and slowness.

Test	*N*	Percentage correct				Average response time (log sec.)			
		M	*SD*	α items	α models	*M*	*SD*	α items	α models
1a	310	.817	.128	.931	.897	1.963	.345	.980	.971
1b	300	.786	.142	.938	.906	2.031	.346	.979	.970
2a	303	.729	.150	.942	.909	2.179	.355	.977	.966
2b	305	.731	.139	.935	.905	2.077	.354	.977	.967
Total	1218	.766	.145			2.062	.358		

Note. Tests 1a and 1b share the same item models with different instances, and the same for 2a and 2b.

3. Study 2

In study 1, the points manipulation had no effect on accuracy scores and a small effect ($d = .28$) on the speed of responses. It is possible that children, who are more immersed in computer games in their ordinary lives, will respond more powerfully to the gamification of the assessment. Therefore, as part of piloting this assessment, a second study was conducted to examine the effects of the points manipulation with middle school students in grades 6–8. To better gauge participant reactions to the points manipulation, we also asked participants to indicate how much they liked the test and how much effort they exerted during the test. In addition, participants completed two separate test sessions, which enabled us to consider the possible wearing off of any initial gamification effects. Finally, since no differences were found between the two points conditions in the first study, only the 10 + 5 points and control conditions were used in the second study in order to have larger sample sizes and increased statistical power in each condition.

3.1. Method

3.1.1. Participants

For this study, a New Jersey middle school was recruited. The school was paid 10 dollars per participating student. All 6–8 grade (in almost equal numbers per grade) students in the school participated, 693 in all. 50% were girls, 48% were Asian, 31% were White, 12% were Hispanic, and 8% were Black. The average mathematics ability of the students, as indicated by their state assessment scores, was relatively high: 258, 249, and 234 for grades 6–8, respectively. The corresponding standardized differences from the average NJ student were .51, .54, and .43 (calculated from data in New Jersey Department of Education, 2014, p. 67).

3.1.2. Materials

The test items used in the first study were also used in this study.

3.1.3. Design

The complete set of 400 items was divided into four nonoverlapping sets, each composed of two instances from all 50 item models. Each set was further divided into two subsets with 25 item models (and two instances) in each. Each participant was randomly assigned to one of the sets and was further randomly assigned to take one of the subsets in the first test session and the other in the second test session (therefore, whereas in study 1 participants answered four instances from 25 models, in study 2 participants answered two instances from 50 models in order to have responses for a wider range of item models). Participants were also randomly assigned to one of two points feedback conditions, either the 10 + 5 points condition or a control condition in which points were not awarded for correct answers.

3.1.4. Procedures

Procedures similar to those of the first study were used in this study, except that teachers were asked to schedule two test sessions for each class, at their convenience. Each test session lasted one class period (the median time to complete each session, including instructions and posttest questions, was 20 min) and was conducted in a computer lab. The two sessions were completed within one day of each other by most students (70%) and within five days by almost all students (except a few make-up sessions).

Following the completion of each session, participants were asked to indicate how much they liked the test, compared to other mathematics tests (with the answer options *much more*, *somewhat more*, *about the same*, *somewhat less*, *much less*), how much effort did they put into answering the questions in the test (with the answer options *a great deal*, *quite a lot*, *somewhat*, *a little*, *none*), and for the students in the points condition, whether they liked getting points for their answers (with the answer options *yes*, *not sure*, *no*).

3.2. Results

Table 2 presents psychometric properties of response accuracy and response time, or slowness (natural log of time in seconds to answer). Overall, the items were more difficult for and answered more slowly by the middle school students than by the adults in Study 1. The average percentage correct score was 68% and average mean log response time was 2.58 (corresponding to 13 s). The internal consistency of response accuracy was very high, with Cronbach's alpha coefficients around .97 for the 100 item scores and slightly lower for the 50 item model scores (each score the sum of two item instances). The internal consistency of response time was similar, with Cronbach's alpha coefficients of .96 for the 100 item scores and slightly lower for the 50 item model scores.

The predictive validity coefficients of the accuracy scores (number of correct responses across the two sessions) with respect to the students' state assessment scores (obtained a year before the study) were high, .84, significantly higher than the correlation between students' mathematics grades for the entire year and state assessment scores, .68 ($p < .01$ for the test of correlated correlations). The predictive validity coefficients of the slowness scores (average of log seconds) with respect to the students' state assessment scores were lower, −.38, and significantly lower than mathematics grades, $p < .01$.

Table 2
Psychometric properties of response accuracy and slowness.

Test	*N*	Percentage correct				Average response time (log sec.)			
		M	*SD*	α items	α models	*M*	*SD*	α items	α models
1	172	0.680	0.203	0.967	0.960	2.601	0.349	0.962	0.957
2	178	0.668	0.200	0.966	0.958	2.561	0.359	0.963	0.957
3	173	0.678	0.199	0.966	0.958	2.586	0.357	0.963	0.957
4	170	0.710	0.186	0.962	0.955	2.574	0.350	0.962	0.958
Total	693	0.684	0.197			2.580	0.353		

Note. All tests share the same 50 models with two different instances for each test.

A series of ANOVAs was performed to examine the possible effects of the point manipulation on test performance and response time, the relationship between test performance and response time, as well as the degree to which students liked the tests and felt they were making an effort in answering the questions of the test. All interactions between independent variables were included in these analyses.

A 2 (points condition) × 2 (sex) × 3 (grade level) × 2 (session number) mixed effects ANOVA was performed on test scores (percentage of correct answers, out of 50 questions for each session), with session number as within-subjects effect. The only significant effect with test scores as the dependent variable was a session number effect, $F(1, 681) = 6.03$, $p = .01$, with slightly higher performance in the first session ($M = .69$, $SE = .01$) than in the second session ($M = .68$, $SE = .01$).

The same mixed effects ANOVA was performed on the average (natural) log of response times (in seconds). A significant points effect was found for the log response time data, $F(1, 681) = 4.52$, $p = .03$, with lower (faster) response times in the points condition ($M = 2.56$, $SE = .02$, corresponding to 12.9 s) than the no-points condition ($M = 2.61$, $SE = .02$, corresponding to 13.6 s). No other interactions with the points condition were significant, $F < 2.4$. Other significant effects included a session number effect, $F(1, 681) = 258.84$, $p < .01$, with lower response times in the second session ($M = 2.48$, $SE = .01$) than the first session ($M = 2.68$, $SE = .01$); a sex effect, $F(1, 681) = 15.96$, $p < .01$, with lower response times for boys ($M = 2.53$, $SE = .01$) than girls ($M = 2.63$, $SE = .01$); and a grade effect, $F(1, 681) = 23.57$, $p < .01$, where post-hoc Tukey tests indicated that response times for grade 8 ($M = 2.45$, $SE = .02$) were lower than either grade 7 ($M = 2.62$, $SE = .02$) or grade 6 ($M = 2.66$, $SE = .02$).

The relationship between test performance and response time across both sessions was weaker in the no points ($r = -.27$) than the points condition ($r = -.47$) and the difference was significant ($p < .01$).

Overall, participants liked the test more than other mathematics tests, with 35% *much more*, 29% *somewhat more*, 20% *about the same*, 7% *somewhat less*, and 9% *much less*. The same mixed effects ANOVA was performed on degree to which students liked the test (responses coded as 1–5). The only significant effect was a session number with points interaction effect, $F(1, 681) = 5.80$, $p = .02$. Post-hoc Tukey tests showed that in the first session participants in the point condition ($M = 3.83$, $SE = .07$) liked the test more than in the no-points condition ($M = 3.59$, $SE = .07$), $p = .01$. However, no significant difference in the second session was found ($M = 3.75$, $SE = .07$ for point condition, versus $M = 3.73$, $SE = .07$ for control condition). The correlations between the degree to which students liked the test and their accuracy scores for the no-points ($r = .27$) and points ($r = .30$) conditions were similar, $p = .71$. The correlations between the degree to which students liked the test and their slowness scores for the no-points ($r = -.01$) and points ($r = -.11$) conditions were also similar, $p = .19$.

Overall, participants indicated a considerable amount of effort during the test, with 20% *a great deal*, 37% *quite a lot*, 30% *somewhat*, 10% *a little*, and 4% *none*. The same mixed effects ANOVA was performed on student self-reported effort during the test (responses coded as 1–5). A significant points effect was found, $F(1, 681) = 14.89$, $p < .01$, with higher effort in the points condition ($M = 3.71$, $SE = .04$) than the no-points condition ($M = 3.44$, $SE = .04$). A significant grade effect was also found, $F(1, 681) = 15.37$, $p < .01$, with lower reported effort for grade 8 ($M = 3.31$, $SE = .05$) than grade 7 ($M = 3.62$, $SE = .05$) and higher still for grade 6 ($M = 3.77$, $SE = .04$). However, these main effects were qualified by a points by grade significant interaction, $F(1, 681) = 3.61$, $p = .03$. Post-hoc Tukey tests showed that only in grade 8 was perceived effort under the points condition ($M = 3.58$, $SE = .09$) significantly higher than under the no-points condition ($M = 3.05$, $SE = .09$).

Finally, a large majority of participants in the points condition (80%) liked getting points for their performance (15% were not sure and 5% did not like it), and the percentage of participants that liked getting points did not differ by sex, session, or grade (all $p > .08$).

4. Summary and conclusions

In the two studies described in this paper we examined the effect of introducing a gamification element into a mathematics assessment in the form of points awarded for accurate and speedy responses. In summary, the point manipulation showed only minor effects on various measures for both adults (study 1) and teenage students (study 2). The point manipulation had no effect on the main performance outcome, response accuracy, in either population. The point manipulation had a significant effect on the second performance outcome, response speed, in both populations. Speed of response decreased, but the effect sizes were small for adults ($d = .27$) and even smaller for children ($d = .14$). In addition, no interactions were found between the point manipulation and sex.

There is some evidence that the relationship between accuracy and speed was somewhat stronger in the points condition and with an emphasis on speed. In the adult population, the differences in correlations between speed and accuracy scores across conditions were not significant. For children, the difference in correlations was significant (−.27 and −.47 for the no-points and points conditions, respectively). These effects, the overall increase in speed and the changed relation between speed and accuracy, indicate that participants responded to the experimental manipulation, but not by increasing accuracy of performance. It is interesting to note that the stronger speed-accuracy relation in the student sample is likely to make the speed score *less* useful as a performance measure, because a stronger speed-accuracy correlation makes the speed score more redundant as a predictor of other measures.

Middle school participants' reactions to the test revealed higher likeability ratings for the test under the points condition, but only in the first of the two sessions, possibly as a result of the dissipation of the points novelty. In addition, perceived effort during the test was higher in the points condition, but only for grade 8 students. The grade 8 students also reported less effort than their peers in grades 7 and 6. This result raises the possibility that the points manipulation has a beneficial effect on students with low motivation who exert less effort. However, it is not clear that the lower reported effort of grade 8 students actually manifested itself in less optimal performance. Although there were no significant differences across grade levels in accuracy performance, this may have been caused by a multitude of reasons, including the fact that instruction in higher grades shifts away from basic arithmetic concepts and towards algebra. Finally, most students (80%) across all grade levels liked the points manipulation.

In summary, in the two studies we found similar results in terms of the effect of points on performance: no effects were found on accuracy, whereas speed of response increased in the points condition. For the middle school students, only minor points effects on the likeability of the test and the perceived effort during the test were found, although most students liked getting points during the assessment.

In addition to the limited nature of gamification features manipulated in this study, several other limitations should be mentioned, including the educational activity, domain and type of questions, and sample of participants in this study. The focus of this study was on a short-term test taking activity. It is possible that the potential motivational benefits of gamification would manifest themselves over longer periods of time (Jackson & McNamara, 2013). In addition, this study was focused on one particular domain and type of questions (short questions measuring mastery and fluency in core mathematics topics). It is possible that reactions in other domains (e.g., nonanalytical

content areas) or even other types of questions (e.g., questions that require more complex problem solving) would be different. Finally, some limitations on the sample of participants should be noted. In study 1, adult participants were volunteers and thus may have had a more positive attitude towards mathematics than other adults. Only one school participated in study 2, and its students showed, as a group, higher than average state assessment scores, which may have also contributed to more positive attitudes towards mathematics. It is possible that gamification effects would be stronger for students with less favorable attitudes towards the subject area. The two samples also completed the activity in different contexts (adults individually on their home computer and students in a classroom environment). It would be interesting, for example, to replicate the student results in a less formal environment.

These limitations notwithstanding, the relative lack of effects for the point manipulation indicates that the manipulation provided limited incentives beyond that already provided by other characteristics of the test. One possibility is that the test itself already provides powerful incentives that mask any additional effects of points. For example, it is possible that for many students the experience of taking this test was not altogether negative (e.g., they liked the assessment more than other mathematics assessments), in contrast to the underlying assumption of the gamification argument. Success in mathematics may have intrinsic value for many students. In addition, students may have been more motivated to perform well because they received immediate feedback on their responses.

Another explanation for these minor effects is that only one feature was implemented and manipulated in this study. Although points, as a measure of overall success throughout the assessment, represent a major aspect of possible feedback and game-like attribute, other aspects were not examined in this study. Other gamification aspects that could be implemented in an assessment context include introduction of elements of competition and embedding the assessment in an attractive narrative. Although it is important to study the effects of each of these aspects separately, it may be the case that none of them will prove effective in isolation and only a combination of different aspects would be powerful enough to make an important contribution to student performance.

In any case, these results demonstrate the complexity of designing educational assessment and instruction systems. Design features that are assumed to have a beneficial effect do not always show the expected effects. Whereas providing immediate feedback on performance has been shown repeatedly to be helpful for students in terms of performance and motivation, additional “bells-and-whistles” may not always do the trick.

References

Arieli-Attali, M., & Cayton-Hodges, G. A. (2014). *Expanding the CBAL competency model for mathematics assessments and developing a Rational Number learning progression (Research Report 14-08)*. Princeton, NJ: Educational Testing Service.

Azevedo, R., & Bernard, R. M. (1995). A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research, 13*(2), 111–127.

Baumert, J., & Demmrich, A. (2001). Testing motivation in the assessment of student skills: the effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16*, 441–462.

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323–359). Hillsdale, NJ: Erlbaum.

Bennett, R. (2011). *CBAL: Results from piloting innovative K-12 assessments (Research report 11–23)*. Princeton, NJ: Educational Testing Service.

Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade NAEP reading assessment. *Teachers College Record, 113*, 2309–2344.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*, 3–5.

Butler, R. (1987). Task-involving and ego-involving properties of evaluation: effects of different feedback conditions on motivational perceptions, interest, and performance. *Journal of Educational Psychology, 79*, 474–482.

Case, R. (1985). *Intellectual development: Birth to adulthood*. New York, NY: Academic Press.

Cumming, J., & Elkins, J. (1999). Lack of automaticity in the basic addition facts as a characteristic of arithmetic learning problems and instructional needs. *Mathematical Cognition, 5*, 149–180.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin, 125*, 627–668.

Deterding, S., Sicart, M., Nacke, L., O'Hara, K., & Dixon, D. (2011, May). Gamification: using game-design elements in non-gaming contexts. In *PART 2-Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems* (pp. 2425–2428). Association for Computing Machinery.

Empson, S. B., Levi, L., & Carpenter, T. P. (2011). The algebraic nature of fractions: developing relational thinking in elementary school. In J. Cai, & E. Knuth (Eds.), *Early algebraization: A global dialogue from multiple perspectives* (pp. 5–24). New York, NY: Springer.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*, 363–406.

Gee, J. P. (2003). *What video games have to teach us about learning and literacy*. New York, NY: Palgrave Macmillan.

Hartmann, T., & Klimmt, C. (2006). Gender and computer games: exploring females' dislikes. *Journal of Computer – Mediated Communication, 11*, 910–931.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112.

Jackson, G. T., Dempsey, K. B., & McNamara, D. S. (2012). Game-based practice in a reading strategy tutoring system: showdown in iSTART–ME. In H. Reinders (Ed.), *Digital games in language learning and teaching* (pp. 115–138). Basingstoke, England: Palgrave Macmillan.

Jackson, G. T., & McNamara, D. S. (2013). Motivation and performance in a game-based intelligent tutoring system. *Journal of Educational Psychology, 105*, 1036–1049.

Kebritchi, M., Hirumi, A., & Bai, H. (2010). The effects of modern mathematics computer games on mathematics achievement and class motivation. *Computers & Education, 55*, 427–443.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254–284.

Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: motivation matters. *Educational Researcher, 41*, 352–362.

McGonigal, J. (2011). *Reality is broken: Why games make us better and how they can change the world*. New York, NY: Penguin.

Mory, E. H. (2004). Feedback research revisited. In D. Jonassen (Ed.), *Handbook of research on educational communications and technology* (pp. 745–783). Mahwah: Erlbaum.

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. Washington, DC: U.S. Department of Education.

New Jersey Department of Education. (2014). *NJASK 2013 Technical Report: Grades 3–8*. Retrieved from state.nj.us/education/assessment/ms/5–8/.

O'Neil, H. F., Jr., Abedi, J., Miyoshi, J., & Mastergeorge, A. (2005). Monetary incentives for low-stakes tests. *Educational Assessment, 10*, 185–208.

O'Neil, H. F., Jr., Sugrue, B., & Baker, E. (1996). Effects of motivational interventions on the NAEP mathematics performance. *Educational Assessment, 3*, 135–157.

Papastergiou, M. (2009). Digital game-based learning in high school computer science education: Impact on educational effectiveness and student motivation. *Computers & Education, 52*, 1–12.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119–144.

Shaffer, D. W. (2006). *How computer games help children learn*. New York, NY: Palgrave Macmillan.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237–247.

Skinner, B. F. (1953). *Science and human behavior*. Oxford, England: Macmillan.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: problems and potential solutions. *Educational Assessment, 10*, 1–17.

Yee, N., Ducheneaut, N., & Nelson, L. (2012, May). Online gaming motivations scale: development and validation. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems* (pp. 2803–2806). Association for Computing Machinery.

Zichermann, G., & Linder, J. (2010). *Game-based marketing: Inspire customer loyalty through rewards, challenges, and contests*. Hoboken, NJ: Wiley.