**Original Paper**

# No effect of gamification on attrition from a web-based longitudinal cognitive testing study

Jim Lumsden[1,2] *, Andy Skinner[1,2], David Coyle[3], Natalia Lawrence[4], Marcus Munafò[1,2]

[1] MRC Integrative Epidemiology Unit (IEU) at the University of Bristol, Bristol, United Kingdom

[2] UK Centre for Tobacco and Alcohol Studies, School of Experimental Psychology, University of Bristol, Bristol, United Kingdom

[3] School of Computer Science, University College Dublin, Dublin, Ireland

[4] School of Psychology, College of Life and Environmental Sciences, University of Exeter, Exeter, United Kingdom

* Corresponding author

*E-mail:* jim.lumsden@bristol.ac.uk

*Telephone:* +44(0)117 92 88450

*Address:* Jim Lumsden, 12a Priory Rd, School of Experimental Psychology, University of Bristol, BS8 1TU, UK

# No effect of gamification on attrition from a web-based longitudinal cognitive testing study

## Abstract

**Background:** The prospect of assessing cognition longitudinally and remotely is attractive to researchers, health practitioners and pharmaceutical companies alike. However, such repeated-testing regimes place a considerable burden on participants, and with cognitive tasks typically being regarded as effortful and unengaging, these studies may experience high levels of participant attrition. One potential solution is to gamify these tasks to make them more engaging: increasing participant willingness to take part and reducing attrition. However, such an approach must balance task validity with the introduction of entertaining gamelike elements.

**Objectives:** We set out to investigate the effects of gamelike features on participant attrition using a between-subjects, longitudinal online testing study.

**Methods:** We used three variants of a common cognitive task, the stop signal task, with a single gamelike feature in each: one variant where points were rewarded for performing optimally, another where the task was given a graphical theme, and a third variant which was a standard stop signal task and served as a control condition. Participants completed four compulsory test sessions over four consecutive days before entering a six-day voluntary testing period where they faced a daily decision to either drop out or continue taking part. Participants were paid for each session they completed.

**Results:** 482 participants signed up to take part in the study, with 265 completing the requisite four consecutive test sessions. We saw no evidence for an effect of gamification on attrition. A log-rank test showed no evidence of a difference in dropout rates between task variants ($X^2$ (2, $N$ = 265) = 3.022, $p$ = .22) and a one-way ANOVA of the mean number of sessions completed per participant in

each variant also showed no evidence for a difference ($F$ [2,262] = 1.534, $p$ = .21, *partial* $\eta^2$ = 0.012.

**Conclusions:** Our findings raise doubts about the ability of gamification to reduce attrition from longitudinal cognitive testing studies.

*Keywords:* gamification, gamelike, online, stop signal task, attrition, engagement

## Introduction

The prospect of assessing cognition remotely and longitudinally is attractive to researchers, health practitioners and pharmaceutical companies alike. However, until relatively recently, assessments of cognitive functioning were typically done in a laboratory or clinical setting, making multiple testing sessions expensive and a burden to both researchers and participants. Recently, the use of online platforms for crowd-sourcing participants, such as such as MTurk [1] and Prolific Academic [2], combined with the growing number of platforms for delivering online cognitive assessments, such as Testable [3] and Gorilla [4], have given researchers the ability to gather data on large numbers of people within very short time spans [5–8]. These new technologies have allowed psychological experiments and interventions to be delivered online, easily and inexpensively [9–11].

However, one issue for online studies (and particularly longitudinal studies) is that they must compete against the wealth of entertainment and distraction available on the internet in order to attract and retain their participants. This is made more difficult by the fact that dropping out of an online study is easier than doing so in the laboratory: a participant need only close their browser window [12]. Many authors have reported difficulties sustaining participant numbers for the duration of their online studies [13,14], and reviews of adherence to intervention trials have documented dropout rates of around 50% [15,16], considerably higher than in lab studies where dropout rates are around 13% [17]. The gradual reduction in the number of participants who continue to provide study data over time is known as attrition [16,18]. High levels of attrition may cause studies to suffer from smaller than intended sample sizes, incomplete data sets, wasted participant compensation and potentially biased results [19–21].

Attrition is often characterized as a 'lack of participant engagement' [22,23], but the definition of 'engagement' is unclear. One potential definition conceptualizes engagement in twofold sense [24], both referring to participants' subjective experience of taking part in a study (i.e. their enjoyment of the procedure) and participants' behavior when interacting with the study (i.e. how often they return to the study website, or how quickly they drop out). Under this definition, attrition is a sub-component of engagement: an objective behavioral measure which could be assumed to relate to the concept of engagement as a whole.

In recent years, gamification has been heralded as a potential mechanism for increasing participant engagement with online studies and interventions [25–27]. The premise is that by adding gamelike (points, graphics, levels, competition etc.) features to an otherwise mundane task, we might be able to create a more enjoyable and compelling experience for the user [28–30]. By utilizing games' ability to engage individuals, it may be possible to make the testing experience less burdensome, thereby reducing attrition. In previous studies, self-report questionnaires of participant enjoyment have found that gamelike experiments are typically rated as more enjoyable than their non-gamelike counterparts [25,31–35]. There are also some examples of gamification increasing *objective* measures of engagement, such as number of optional trials completed [34] or the number of optional testing blocks chosen [36].

Two recent systematic reviews looked at the effect of gamification on engagement with 'online programs' (mostly e-learning) [37] and web-based mental health interventions [38]. Drawing on the data from 15 studies comparing engagement with gamified programs to non-gamified programs, Looyestyn and colleagues found medium to large effects of gamification on objective measures of engagement such as time-spent using the program, number of website visits and volume of contributions. In contrast, Brown and colleagues assessed the impact of gamification on adherence

to 61 online mental-health interventions and found that not only was gamification applied fairly lightly (most studies used only one gamelike element), there was also little evidence for its efficacy [38]. These conflicting findings could be the result of the reviews' different scopes, the lack of studies in Brown's review which specifically assessed the impact of gamification on adherence, or the very minimal gamification found to have been applied in the reviewed mental-health interventions.

The reluctance of researchers to liberally apply gamification to precisely designed mental-health interventions is understandable; any small change might impact the intervention's efficacy. Within our own field of gamified cognitive assessment, efforts to increase participant engagement and reduce attrition must be implemented carefully to avoid introducing additional cognitive load and affecting the cognitive constructs under test, thus invalidating the task. Although some studies have reported a positive effect of game mechanics on participant performance [39–41], others have found evidence that gamelike tests *do not* improve performance, and may in fact worsen it [31–33,42–44]. For example, Katz et al., [42] found that adding a point-scoring system to a working-memory training task negatively impacted the task's ability to train cognition. These contrasted findings are likely due to the diverse range of cognitive tasks being used and the variety of gamification approaches applied to them, hence highlighting the need for research which systematically manipulates gamification approaches within a single type of task [25].

We recently conducted a study exploring the impact of two simple game mechanics (Points and Theme) on the data collected by, and subjective participant ratings of, a response inhibition task [43]. The Points variant rewarded participants with points in accordance with their performance on the task while the Theme variant utilized a variety of narratively themed stimuli and task backgrounds. A Non-Game variant was included as a control condition. This was comparable to a

clinical version of the task, with some minor graphical changes to ensure suitability for online use. We found that Points were rated highest of the three variants on a subjective questionnaire of enjoyment and engagement, and did not negatively affect participant performance on the test. However we found that the narratively themed task was less liked and negatively affected participant performance. We also saw ceiling effects on participant accuracy in all three task variants due to the ease of the response inhibition task we used.

In the present study, we aimed to investigate whether simple gamification could reduce participant attrition from an online longitudinal cognitive testing study. Building on our previous study, we used three variants of a response inhibition task, but we switched to using the Stop Signal Task (SST) in order to increase task difficulty and avoid ceiling effects. We used the same gamelike features (Non-Game, Points, and Theme) as in the previous study. We aimed to assess the effect of gamification on attrition using a longitudinal design whereby participants signed-up to four compulsory test sessions over four consecutive days before entering a six-day voluntary period where they could continue to take part once per day if they desired. Participants were told they would receive £4 for completing all compulsory sessions and an additional 50p for each optional session they completed.

We hypothesized that the non-game variant would have the highest attrition rate, losing participants quickly once the fourth session was complete. We expected the Points variant initially to maintain high numbers, before falling rapidly around day 6-7. Finally, we expected the Theme variant to lose participants steadily at first before stabilizing to a low attrition rate, eventually retaining a higher number of participants than either the Non-Game or Points variants. For more information on why we predicted these hypotheses, please see Supplementary Methods.

# Methods

## Design and Overview

We used a between-subjects, repeated measures experimental design that took place online over four to ten days. The independent variable was SST variant (Non-Game, Points, Theme). The dependent variables of interest were participant attrition, scores on a questionnaire of enjoyment and engagement, two pilot objective measures of engagement and Stop Signal Reaction Times (SSRTs), We pre-registered the study on the Open Science Framework [45].

## Participants and Procedure

Participants were recruited from the user base of Prolific Academic [2], which handles the process of checking inclusion criteria, displaying study information and participant reimbursement. We required participants to be older than 18 and to have English as a first language, but had no further criteria. Once registered, participants were directed to the 'Mindgames Platform' where they entered their Prolific ID and received a unique link which they used to access the study thereafter. They were then randomly assigned to a single task variant for the duration of the study and completed an online consent form before testing commenced.

Participants were required to complete one ten-minute session per day for the first four days of the study in order to receive £4 as compensation for their time. If participants dropped out of the study before completing four sessions, and did not contact us with a reason (technical difficulties, etc.) then they did not receive any compensation. This was made clear on the information sheet which participants read before they signed up to the study, and on the study website itself. For the first four sessions participants were sent daily reminders via the Prolific Academic messaging system. On the fourth day participants were informed there would be no more reminders, and that they were

free to either drop out, or continue to take part in the study each day thereafter for up to six days, with each additional session earning them 50p, for a total of between £4 and £7.

The appropriate compensation for the optional sessions was determined by way of a pilot study using the Non-game variant only. We randomly allocated participants to one of three levels of compensation: 50p, £1 or £2 per optional session completed (the base compensation was still £4), and found that the average number of sessions completed per participant was 7.1, 8.4 and 9.4 respectively. Given that we anticipated the Non-Game variant to be the least motivating of the three variants, that we wanted to avoid ceiling effects, and that we wanted to minimize the motivational influence of the compensation, we opted for a reward of 50p per optional session.

Ethics approval was obtained from the Faculty of Science Research Ethics Committee at the University of Bristol (40361) and the study was conducted according to the revised Declaration of Helsinki [46].

## Materials

### *The Mindgames platform*

Aside from participant recruitment, daily reminders and reimbursement, all other elements of the study were hosted on a custom website [47]**.** The website was a single page web app written in JavaScript, with a JSON based Firebase database [48] and PixiJS [49] as the 2D renderer. The site opened to a main-menu screen from which the participant could view the number of sessions they had completed and the amount of money they'd earned so far (see Figure 1). Participants had access to a 'history' screen which allowed them to view their previous progress and monitor their results over time. Clicking the start button displayed a series of instruction screens followed by the SST task and a short questionnaire. The session ended on the history screen, and the main menu's 'start' button became inactive until midnight that night. Each session took approximately 10

minutes to complete. On the first day of taking part, participants also completed a short demographic questionnaire which collected data on age, sex, ethnicity, level of education and the number of hours spent playing video games each week.

**Figure 1: Menu screens of the three task variants.** (A) Non-Game variant, (B) Points variant, (C) Theme Variant


***Stop signal task: Non-Game variant***

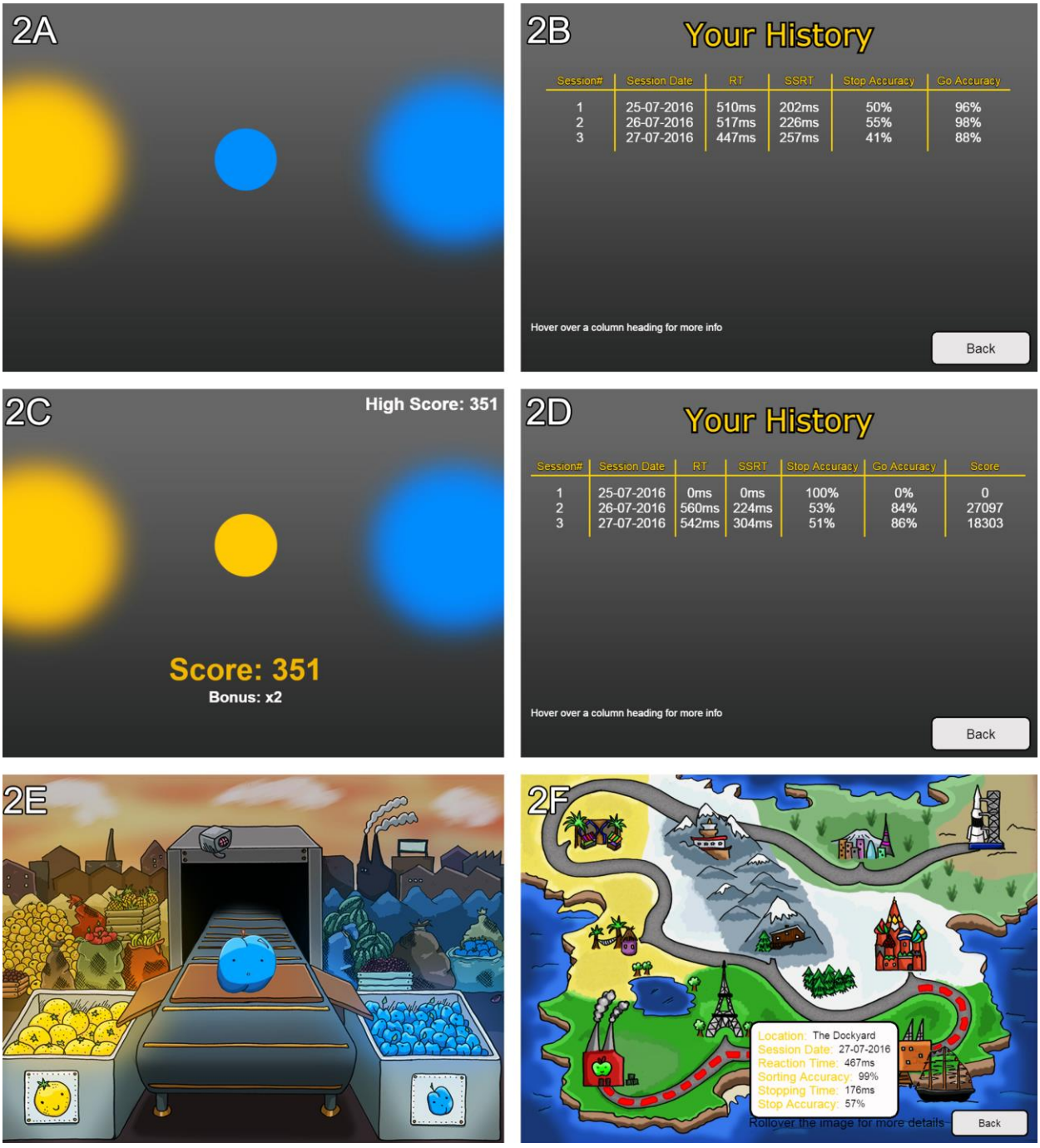
The SST measures response inhibition (see [50] and [51]), a key feature of executive control [52]. It tests the participant's 'action restraint' by presenting a series of stimuli to which the participant must respond as quickly as possible, but are occasionally required to withhold a response. These 'stop trials' are indicated by a visual warning presented a brief delay after stimulus presentation. The primary outcome measure of the SST is the SSRT which is the number of milliseconds of warning a participant needs for them to be able to successfully inhibit their planned response [52].

In this study we decided to use the SST, as opposed to the Go-NoGo task from our previous study [43]. We did this because we found many participants to be performing at ceiling in the Go-NoGo task, which limited our ability to detect differences between the task variants. The SST is more challenging than the Go-NoGo task because it dynamically adjusts the task's difficulty to match the inhibitory control of the user, therefore reducing the likelihood of a participant performing at ceiling.

We based our SST on the widely used CANTAB SST [53,54] albeit with a visual rather than auditory stop signal and some graphic upgrades to make the task more suitable for online use. Each trial began with a fixation cross that was displayed in the middle of the screen, with two colored zones on the left and right of the fixation cross (see Figure 2A). 500 ms later a colored circle appeared over the fixation cross and participants had to respond as rapidly as possible by pressing either the left or right arrow key to indicate which colored zone matched the color of the circle. On 25% of trials, white brackets appeared around the circle after it was shown: when this occurred the subject had to withhold their response and wait until the next trial began (each trial was displayed for 900 ms). If the participant responded before the stop signal was displayed, then the trial was recorded as failed, but white brackets were not displayed. Between each trial there was a random inter-trial interval of between 500-1000 ms. The delay between the circle onset and the bracket onset is called the Stop Signal Delay (SSD), and was varied according to a four-staircase tracking algorithm , designed to sample across the SSD/Inhibition-Probability space (see Supplementary Information) [55,56]. The task consisted of five blocks of 48 trials each, with a 10 second break between each block. If the participant minimized the browser window or changed tabs then the task was paused (due to the default way in which timers in JavaScript operate). However, if the browser window was not in focus but was still visible (on a 2nd monitor for example) then the task was not paused.

In the Non-Game variant, the participant's history was presented as a list of previous sessions, with median reaction times and estimated SSRTs (see Figure 2B). Hovering over a column displayed a brief explanation of the variable (e.g. "The Reaction Time column shows the average time, in milliseconds, which it took you to respond to the circles appearing in each session")

**Figure 2: Screenshots of the Stop Signal Task variants and their associated history Screens.** (A/B) Non-Game variant, (C/D) Points variant, (E/F) Theme variant




***Stop signal task: Points variant***

The Points variant was similar to the Non-Game variant but with the addition of a points mechanic and the task being framed as a game. Points are a common feature of gamified tasks [25], and are classed as "1st Step" gamification [57]. In our task, the participant's points score was displayed at the bottom of the screen throughout (see Figure 2C). The scoring system was very similar to that used in our previous study [43], which in turn was based on that used by Miranda et al., [33]. The

scoring system also incorporates the findings of Guitart-Masip and colleagues [58] who found that subjects were much more successful in learning active (go) choices when rewarded for them, and passive choices (stop) when punished. On each successful non-stop-

trial the participant earned points equal to Bonus*(800-reaction time)/5, and the number of points gained was displayed briefly in the inter-trial interval. This Bonus was a multiplier (x2, x3, x4...), which increased by 1 every 3 trials but decreased by 3 when the participant failed a stop trial. The bonus was not lost on stop trials to which the participant responded before the stop signal was displayed (to all appearances, the trial was not a stop trial). On a successful inhibition to a stop signal the bonus was not lost, but no points were awarded (as there was no reaction time on which to base the score for that trial). Scores were maintained over blocks, but not over sessions. The scoring system was outlined to the participants in the instructions for the task.

The participant's history was presented as a list of median reaction times, SSRTs and scores from each testing session (see Figure 2D). Additionally, the participant's highest score was saved as a High Score, and was displayed in the top right-hand corner throughout every testing session.

***Stop signal task: Theme variant***

The Theme variant was similar to the Non-Game variant but with the addition of a graphical theme and a sense of progression. The task was framed as a game and featured themed graphics and stimuli, with the yellow and blue stimuli being replaced by images of objects, though still predominantly blue or yellow (see Figure 2E). The task was presented on a series of different graphical backgrounds (see Supplementary Figure 1), but with some shared elements: a conveyor belt on which objects appeared and two bins to the left and right into which these objects were sorted. The stop signal was explained as an automatic "fault detector" which scanned objects as they sat on the conveyor.

The participant's history was presented as a map (see Figure 2F), and previous sessions' summary data was displayed when the user hovered over the corresponding icon. Each level on the map had a unique name and thematic instruction text, with the intention of creating an overarching goal, perceptual curiosity and fostering a sense of participant progression [59–61].

### *Enjoyment and Engagement Questionnaire*

The Enjoyment and Engagement Questionnaire was designed to collect subjective ratings of the task and was delivered after every session for all three variants. Session 1,4,7 and 10 delivered the full ten-item questionnaire while the remaining sessions delivered a shorter five item questionnaire. These items were answered using a continuous visual-analogue scale, presented as a horizontal line 500 pixels long, with a label at either end and no subdivisions. Participants marked a point between these two labels using their mouse.

The following questions were based on those used in previous studies [31,33,43], and were presented in a random order: (1) How enjoyable did you find that? (2) How frustrating did you find that? (3) How difficult was it to concentrate for the duration of that? (4) How well do you think you performed on that? (5) How mentally stimulating did you find that to be? (6) How boring did you find that? (7) How much effort did you put in throughout that? (8) How repetitive did you find that? (9) How willing would you be to do that again tomorrow? (10) How willing would you be to recommend the study to a friend? Questions 3,4,6,7 and 10 appeared only in on the long version of the questionnaire.

## Dependent Variable Calculation

### *Attrition*

Attrition was measured in two ways: Firstly, we calculated the mean number of sessions completed per participant (sessions which were started but not finished were excluded from this calculation).

Secondly, we calculated the percentage of participants that completed at least one session, two sessions, etc.

***Subjective Measures of Engagement***

Subjective engagement with the task was measured by calculating a mean score from the 10-item Enjoyment and Engagement Questionnaire. Questions 2,3,6 and 8 were reverse-scored in this calculation. This measure was calculated for each participant's first and fourth sessions, and we also created a 'combined score' by taking the mean of the participant's scores from sessions 1 and 4.

***Objective Measures of Engagement***

We piloted two measures that could potentially serve as objective proxies for engagement: we counted the number of times that participants hid the browser window or moved focus to another window while completing the SST, hypothesizing that unengaged participants would be more likely to briefly visit other websites while testing. We combined the counts of both these events into a single measure: loss-of-focus events. We then created an overall measure of loss-of-focus for each participant by calculating the mean number of loss-of-focus events from their first four sessions.

We also investigated coefficients of variation, which quantify reaction time intra-individual variability with respect to mean reaction time , as there is some evidence that changes in motivation can be reflected in reaction time variation [62,63]. Coefficients of variation were calculated by dividing the standard deviation of non-stop trial reaction times by the mean non-stop trial reaction time. Similarly, we created an overall measure of reaction time variation for each participant by calculating the mean coefficient of variation from their first four sessions.

***Stop Signal Reaction Times***

We calculated SSRTs for each session separately, excluding sessions where the assumptions of the race model did not hold. The race model is a commonly used model of inhibitory control and aims

to describe the relationship between stop and go processes [64]. The race model is used to derive the SSRT and so if the assumptions underlying the race model are broken, then the resultant SSRTs are not good representations of the data [50,64]. To that end, we excluded sessions where the median non-stop-trial reaction time was longer than the median failed-stop trial reaction time, where SSDs were not positively correlated with their corresponding median failed-stop reaction times, and where stop-trial accuracy was not negative correlated with SSD.

For the sessions which did meet the assumptions of the race model, SSRTs were calculated by modelling an inhibition function [65], and using it to estimate the SSD at which the participant's probability of inhibiting to a stop signal was 50% [56], we then use this SSD to calculate the SSRT for that session [50,51]**.** We also created a combined measure of SSRT for each participant by taking the mean SSRT of their first four sessions.

## Statistical Analysis

The data that form the basis of our results are available from the University of Bristol Research Data Repository [66].

### *Attrition*

Differences in attrition curves were assessed visually using the Kaplan Meier method to estimate survival functions, a Log-Rank test, and a one-way ANOVA of 'number of sessions completed'.

### *Subjective Measures of Engagement*

We assessed differences in subjective ratings both visually, using bar-charts, and using a repeated-measures ANOVA of total score with session number as the time factor and task variant as the between-subjects factor. Where there was evidence of a difference between task variants, we used post-hoc t-tests to investigate further.

### *Objective Measures of Engagement*

We assessed differences in coefficient of variation and website loss of focus events between task variants using one-way ANOVAs with data combined across the first four sessions. Where there was evidence of a difference between task variants, we used post-hoc t-tests to investigate further.

### *Stop Signal Reaction Times*

We used boxplots and a one-way ANOVA with task variant as the between-subjects factor to investigate the effects of gamification on SSRT.

## Bayesian Analyses

The three task variants were designed with the aim of minimizing differences in primary task reaction time and non-stop trial accuracy. Therefore, given that Frequentist statistics are not ideal for testing equivalences [67,68], we used Bayesian t-tests to assess the evidence for equality of means where Frequentist methods failed to find a difference [69,70]. A Bayesian t-test produces a Bayes Factor (BF), which compares the evidence for two hypotheses. If the evidence favors one hypothesis over the other then the BF will reflect that, but if the evidence is equal for both hypotheses then the BF will imply that the data are insensitive [70–72], see Table 1. We used the Bayesian t-test procedure in JASP [73], with a Cauchy prior width of 0.707. Setting the Cauchy prior width to 0.707 means that in our analysis one hypothesis is "the effect size is zero" and the other is "the effect size is between -0.707 and 0.707". Although both hypotheses are centered on an effect size of 0, the former makes a stronger claim than the latter. As such, effect sizes which are not close to 0 are better represented by the latter hypothesis. A prior width of 0.707 was selected for our analysis because it represents the expectation of a medium-large effect, thus weighting the BF against small effects and reducing the likelihood of a false positive.

*Table 1: Interpreting Bayes factors (adapted from* [71]*)*

| **Hypothesis 0:** | **Strength of Evidence** | **Hypothesis 1:** |
|---|---|---|

| The difference between means is 0 | | The difference between means is between 0 – X |
|---|---|---|
| .33 ≤ BF ≤ 1 | No support either way | 1 ≤ BF ≤ 3 |
| .1 ≤ BF ≤ .33 | Positive | 3 ≤ BF ≤ 10 |
| .01 ≤ BF ≤ .1 | Strong | 10 ≤ BF ≤ 100 |
| BF<.01 | Decisive | BF>100 |

## Sample Size Determination

At the time of study design, to the best of our knowledge, no other studies had investigated the impact of gamification on attrition from a cognitive testing program and therefore we had no previous effect size on which to base a sample size determination. Instead, we hypothesized attrition curves (see Supplementary Methods) for each variant, and calculated the anticipated effect size ($\phi = 0.231$) resulting from a Kaplan-Meier method/Log-rank test (i.e. a chi-square test) on those attrition curves. To detect this difference with alpha = 0.05 and 95% power a sample size of 290 was required. We set this to 291 to allow for equal group sizes.

## Results

### Characteristics of Participants

Participants were recruited in two waves: one starting October 2016 and another starting in January 2017. In both waves the intended sample size was met within three days of the study being posted on Prolific Academic. A total of 482 participants signed up to take part in the study, with 419 (86.9%) of those completing at least one session. A total of 265 (54.9%) participants completed four sessions over four consecutive days as was required by the study criteria (henceforth called *conforming participants*). We excluded five participants from the analysis because their reaction times or blue/yellow accuracy scores were more than four interquartile ranges away from the

group median. We excluded data from sessions that were started but not completed, and we removed trials from the analysis where participants responded in less than 150ms.

The analysis below presents data from 260 participants: less than our intended sample size of 291. This because 32 participants failed to complete the required four sessions in four days, but instead managed to complete four sessions within five days. During the study, we intended on including these *loosely conforming* participants in the analysis, and so stopped recruitment once our intended sample size was achieved. However, for simplicity and adherence to the protocol, we have now decided to present only strictly conforming participant's data below. Analysis of the non-conforming and loosely conforming participants' attrition is presented in Supplementary Information.

Excluding outliers, 260 conforming participants took part: 91 in the Non-Game variant, 86 in the Points variant and 83 in the Theme. The number of hours spent playing video games was comparable between the groups, and participants typically had a high level of education (See Table 2). The most common browser used to complete the experiment was Google Chrome (71%), with others including Firefox (19%), Netscape (5%), Safari (4%), Opera (0.5%) and Internet Explorer (0.5%).

*Table 2: Conforming participant demographic information, shown separate by task variant.*

| | *Non-Game* | *Points* | *Theme* |
|---|---|---|---|
| *Mean age (SD)* | 36 (12) | 35 (12) | 34 (11) |
| *% Male* | 47% | 57% | 51% |
| *Mean video game hours per week (SD)* | 6 (12) | 8 (16) | 8 (14) |
| *Median level of education* | Bachelor's Degree | Bachelor's Degree | Bachelor's Degree |
| *Mode ethnicity (percentage)* | Caucasian (88%) | Caucasian (86%) | Caucasian (90%) |

## Attrition

Figure 3 shows the attrition of conforming participants, while Table 3 shows the mean number of sessions completed per participant in each variant. A Log-Rank test showed no evidence of a difference between the distributions ($X^2$ (2, $N$ = 265) = 3.022, $p$ = .22) and a one-way ANOVA of the number of sessions completed also found no clear evidence of a difference between task variants ($F$ [2,262] = 1.534, $p$ = .21, *partial* $\eta^2$ = 0.012). Given the similarity between Non-Game and Points in mean number of sessions completed, we used a Bayesian T-test to assess their equality and found substantial evidence that they were equal (BF = 0.16), but there was no evidence of equality between the Theme and the Points variant (BF = 0.49) or the Non-Game variant (BF = 0.43).

*Figure 3: Percentage of conforming participants plotted against the number of sessions they completed, shown separately by task variant.*

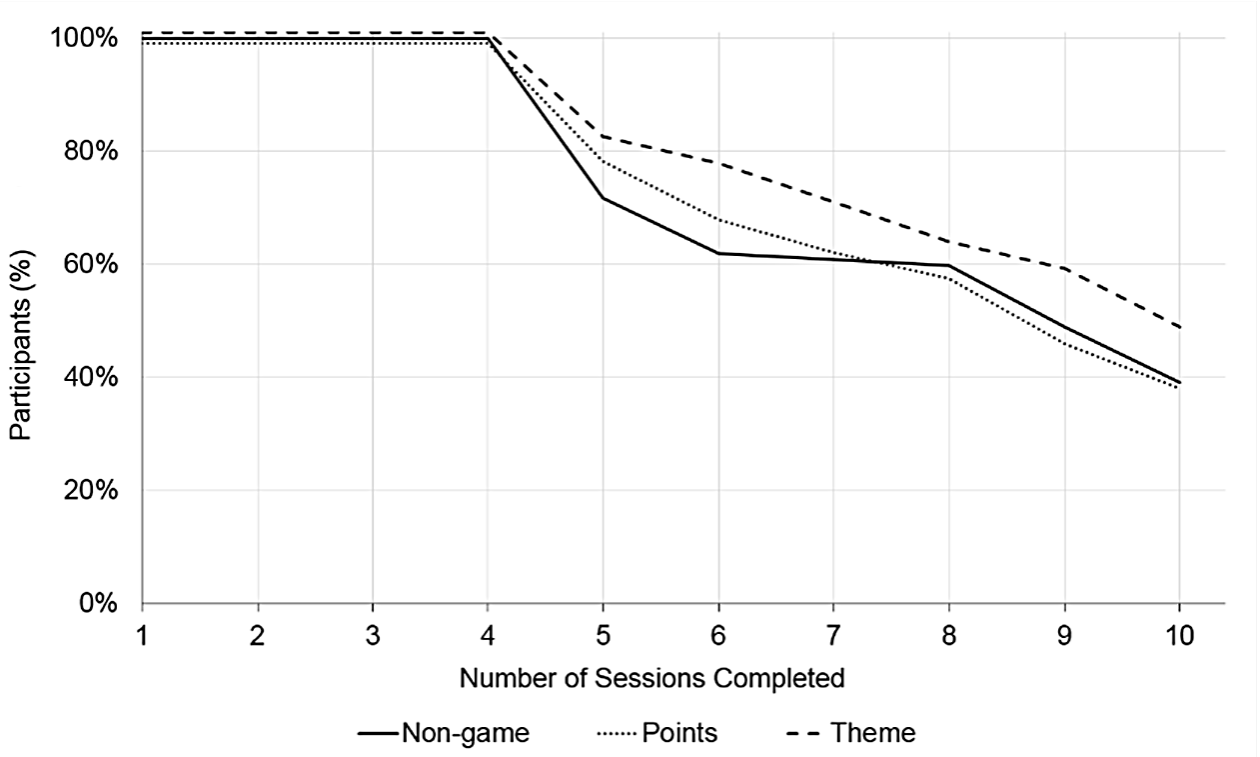


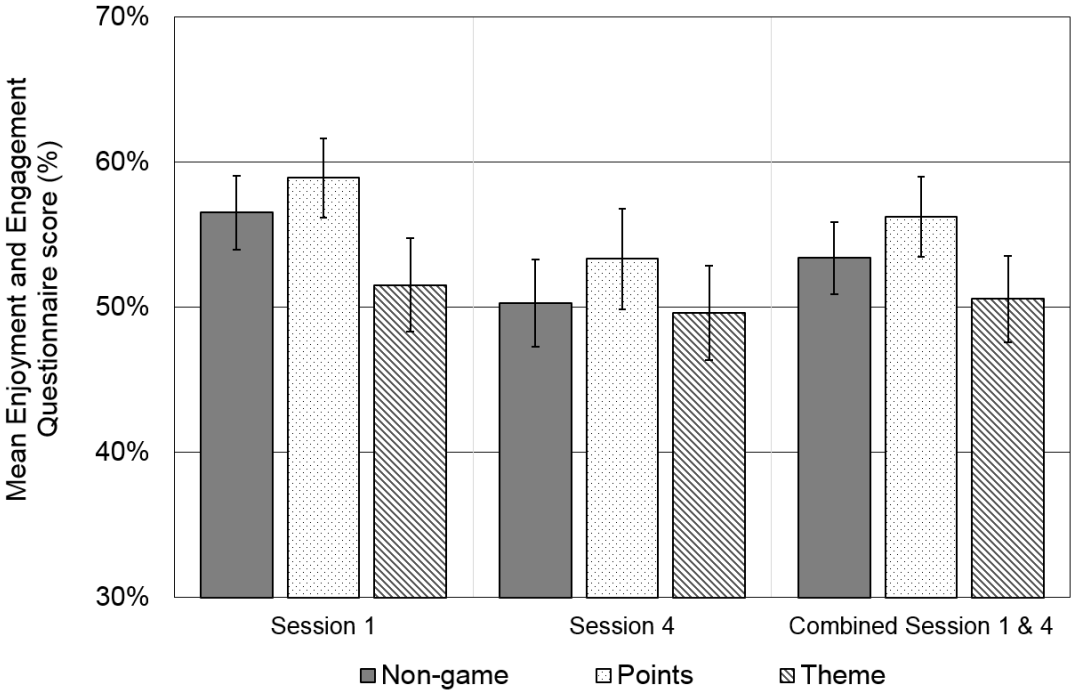

*Table 3: Mean number of sessions completed per participant, shown separately by task variant.* Conforming participants are those who completed their first four sessions within four days as required. 'All participants' includes all who signed up, regardless of their number of sessions completed.

| | *All participants (95% CI)* | *Conforming participants (95% CI)* |
|---|---|---|
| *Non-Game* | 4.9 (4.4 to 5.5) | 7.4 (6.8 to 8.0) |
| *Points* | 5.1 (4.5 to 5.6) | 7.5 (7.0 to 8.0) |
| *Theme* | 5.3 (4.7 to 5.9) | 8.0 (7.5 to 8.6) |

## Subjective Measures of Engagement

We used a repeated-measures ANOVA of mean score from the Enjoyment and Engagement questionnaire with session number (1,4) as the within-subjects factor, and task variant as the between. We used only the two full-length questionnaires completed on the 1st and the 4th session, and completed by all participants (for short-form questionnaire results see Supplementary Information). We saw evidence for small effects of both task variant ($F$ [2,261] = 3.805, $p$ = .02, *partial* $\eta^2$ =0.028) and time ($F$ [1,261] = 35.693, $p$ < 0.001, *partial* $\eta^2$ = 0.120), and weak evidence of an interaction ($F$ [2,261] = 3.014, $p$ = .05, *partial* $\eta^2$ = 0.023). We saw ratings of all task variants decrease between the first (M = 56, 95% CI 54 to 57) and fourth session (M = 51, 95% CI 49 to 53), but it appears the Non-Game and Points variants were the main drivers of the interaction effect: dropping by 6% (95% CI 4% to 8%) between Session 1 and Session 4, whereas ratings of the Theme task decreased only by 2% (95% CI -1% to 5%). Post-hoc t-tests on the combined scores showed no evidence for differences between Non-Game and Points, nor Non-Game and Theme ($ps$ > .15), but did show Points and Theme to be different (mean difference = 5.7%, 95% CI 1.6 to 9.7, t(171) = 2.749, $p$ = .007, $d$ = 0.42). Figure 4 shows the mean scores from each task variant at the two time points, and a combined score taking the averages of both sessions. A breakdown of ratings by individual questions is presented in the Supplementary Information.

*Figure 4: Overall scores from the subjective enjoyment and engagement questionnaire.* Mean responses of visual-analogue scale scores from questionnaires delivered on sessions 1 and 4, and the averaged scores from Sessions 1 and 4, shown separately by task variant and time point. Error bars represent 95% confidence intervals




As an unplanned exploratory analysis, we were interested in whether a participant's rating one day predicted their return to the study on the following day. We ran a logistic regression with "returned following day" as the binary dependent variable and the previous day's score on the subjective questionnaire as the predictor variable. However, we saw no evidence that subjective questionnaire scores predicted return the following day, ($\beta$ = 0.008, SE = 0.005, Wald(1) = 2.166, $p$ = .14, OR = 1.001, 95% CI 0.997 to 1.019)

## Objective Measures of Engagement

We analyzed reaction time coefficient of variation and website loss of focus events from the four compulsory sessions combined (see Table 4). A one-way ANOVA of coefficient of variation showed strong evidence for a medium effect of task variant ($F$ [2,260] = 3.131, $p$ = *.045*, *partial* $\eta^2$ = 0.024) on participants' reaction time variability, with lower coefficients indicating there was less variability. Post-hoc t-tests showed strong evidence of a difference between the Points and Theme variants (mean difference = 1.5%, 95% CI 0.2 to 2.7, t(170) = 2.349, $p$ = .02, $d$ = 0.36), but no clear evidence for other differences were found (*ps* > .06).

Loss-of-focus events were rare in all task variants, with each participant switching away from the task less than once per session on average. Regardless, we assessed differences in loss-of-focus events between the three task variants using a one-way ANOVA but found no evidence for any differences ($F$ [2,260] = 1.137, $p$ = .32, *partial* $\eta^2$ = 0.008).

*Table 4: Mean objective measures of participant engagement from the first four sessions, shown separately by task variant.*

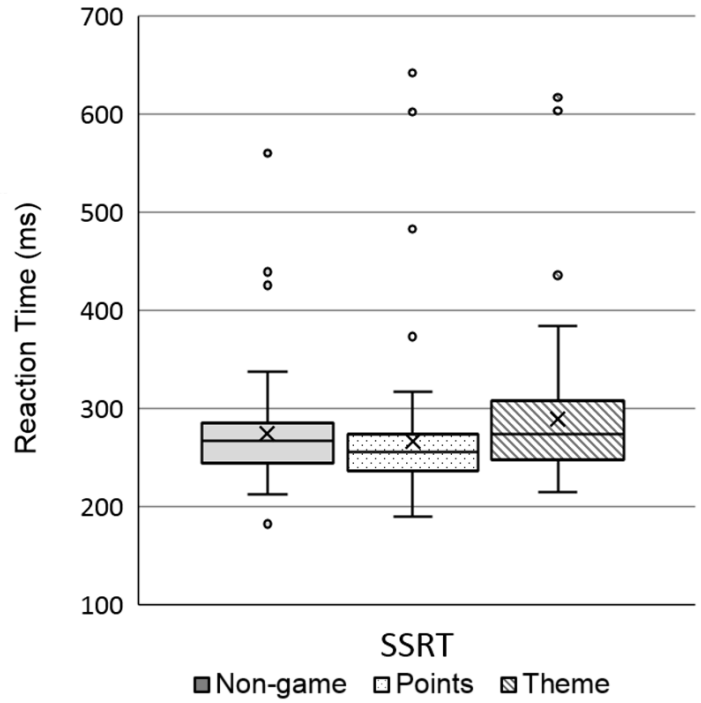
| | *Coefficient of variation (95% CI)* | *Loss-of-focus events (95% CI)* |
|---|---|---|
| *Non-Game* | 18.7% (17.9 to 19.6) | 0.85 (0.50 to 1.19) |
| *Points* | 19.0% (18.1 to 19.8) | 0.82 (0.43 to 1.20) |
| *Theme* | 17.5% (16.7 to 18.4) | 1.21 (0.75 to 1.67) |
| *Overall* | 18.4% (17.9 to 18.9) | 0.95 (0.72 to 1.18) |

## Stop signal reaction times

We checked the data from each session against the assumptions of the race model. Of the 1050 sessions assessed, we excluded 161: 75 from the Non-Game variant, 37 from Points and 49 from Theme. 3 participants failed to meet the assumptions of the race model in all four compulsory sessions, resulting in their exclusion from this analysis. We then analyzed each participant's mean SSRT, with boxplots shown in Figure 5.

A one-way ANOVA showed weak evidence for a small effect of task variant on SSRT ($F$ [2,255] = 2.954, $p$ = .05, *partial* $\eta^2$ = 0.022) with post-hoc t-tests showing a difference between the Theme variant (M = 289, SD = 67) and Points variant (M = 266, SD = 66) (mean difference = 23, 95% CI 5 to 42, t(169) = 2.386, $p$ = .05, $d$ = 0.35). There was no evidence for other differences ($p$s > .24). Bayesian t-tests showed no evidence of equality between the SSRTs of the Non-Game and Theme variants (BF = 0.59), but found substantial evidence for equality between the Non-Game (M = 274, SD = 55) and the Points variants (BF = 0.22).

*Figure 5: Boxplots of mean Stop Signal Reaction Time.* Data combined per participant over the first four sessions and shown separately by task variant




For brevity, not all the analyses planned in the study protocol have been presented– for more detailed methods and analyses please see the Supplementary Information.

## Discussion

Contrary to our hypotheses, we saw no clear evidence of an effect of task variant on participant attrition. This was further strengthened when we included data from loosely conforming participants (see Supplementary Information), which showed strong evidence that the mean number of sessions completed was equal in all task variants. To the best of our knowledge, this is the first empirical study examining the effects of gamification on participant attrition within a cognitive testing context, and our results raise doubts about the efficacy of gamelike tasks for reducing participant dropout.

Despite there being no difference in usage between the variants, we did find an effect of task variant on the subjective ratings of the tasks, with the Points variant having the highest 'combined sessions' mean, followed by the Non-game variant and the Theme. One possible explanation for these findings relates to Self-determination theory; a popular theory of motivation which centers around the concept of psychological needs and need-satisfaction. Self-determination theory posits that human beings have three needs: competence, autonomy and relatedness, and that we find activities to be intrinsically motivating if they help us to fulfill these needs [74]. In the case of our gamelike variants, the Points variant would seem to address competency needs by providing constant feedback on their performance which reinforces the player's success [75], but we do not consider the Theme variant or the Non-Game variant to adequately meet any of the three needs. Since the Points variant was the only variant to address any of these needs, this may explain why it was rated as the most enjoyable in both this study and our previous study [43].

The Theme variant was rated as the worst of the three tasks, which was surprising as it maintained the highest percentage of participants until day ten. One potential explanation is that the task was framed as a game and looked like a game, but offered no actual gameplay. Secondly, the map screen and changing graphical backgrounds may have hinted at player autonomy and exploration as is typical in other games, but ultimately the player experience was railroaded. These two factors may have undermined autonomy and violated participant expectations, resulting in a dissatisfying experience [76,77]. Despite this, it is possible that the clear end goal on the map and novelty of changing backgrounds could explain the maintenance of participants in the Theme variant, while still not being a very satisfying or enjoyable experience.

One additional factor to consider, in the light of self-determination theory, is that paying participants in attrition studies such as this may be counterproductive to measuring true

engagement. There is evidence that providing extrinsic rewards for otherwise motivating tasks may undermine participant autonomy, therefore affecting the task's ability to meet our psychological needs [78,79]. In this study, it is not possible to determine whether intrinsic motivation to take part was affected by the incentive of 50p per additional session. This is further complicated by the potentially unrepresentative nature of a Prolific Academic sample: all of whom have voluntarily signed up to take part in science experiments online, but can choose studies based on the amount of monetary compensation awarded in exchange for their data. Given these issues, one potentially informative avenue for future research in this area would be to explore the effects of these same gamification mechanisms on attrition, but without providing financial incentives.

Money can be a powerful motivator; for example, [80] showed that offering a £10 Amazon voucher to each participant in a longitudinal study resulted in a 9% increased response rate. In our case, it may simply be that money was the most important factor for taking part, and that the similar attrition rates were driven by the identical incentives.

We also found no evidence that participant ratings of engagement and enjoyment could predict the number of optional sessions they would complete. This, combined with the disconnect between the Theme variant ratings and Theme variant usage, serves to highlight the split in different types of engagement that has recently begun to be conceptualized in the literature [24]. In short, the word engagement has been used in the past to refer to both *engagement as subjective experience* and *engagement as usage*, and this study is further evidence that the two concepts are not as closely related as one might assume. Evidence from the video game literature has found that game enjoyment does not relate strongly to game usage, and that game usage can be driven by many other factors including boredom, loneliness and need for escapism [81,82]. This highlights the need for future studies of engagement which collect both subjective and objective measures.

Our two pilot objective measures of engagement: reaction time variation (coefficients of variation) and loss-of-focus events were difficult to interpret. We saw no evidence that losses of focus differed between the task variants, and this is likely because such events were rare (less than one loss of focus per session on average). This is a positive finding, as it shows that participants are willing to properly engage with online cognitive tasks, concentrating for the duration. With respect to coefficients of variation, the pattern of results is directly in contrast with our subjective measures of engagement: the Points variant had the most variable response times but the highest subjective rating, while the Theme variant had the lowest variability and the lowest rating. This is either evidence contrary to the idea that reaction time variability is related to motivation [62,63], or signals that our subjective ratings are not good measures of motivation. Regardless, further research is necessary to understand whether these objective measures provide are related to the broader concept of engagement.

When assessing cognitive data, we found evidence that SSRTs were equivalent between the Points variant and the Non-Game variant. Although the Points variant introduced additional elements to the task which may have increased cognitive load, it is possible that the highly salient feedback and motivational effect of points served to increase participant performance, as has been found in a number of previous studies [41,83–85].

## Limitations and Conclusions

Firstly, we consider the fact that we did not achieve our intended sample size an important limitation of this study. However, we maintain that the results of our supplementary analyses including the loosely conforming participants are quite conclusive, and strengthen our finding that there was no effect of gamification on attrition. Nevertheless, we accept that a balanced group analysis would be preferable. Secondly, we acknowledge that our sample, recruited from Prolific

Academic, with high levels of education, may not be representative of the wider population. Thirdly, we acknowledge that the design of study used is not suitable to validate our gamelike variants as measures of response inhibition, as that would require a within-subjects design in order to test predictive validity [57,86]. Fourthly, the gamelike features we implemented were very lightweight, and certainly wouldn't constitute a full game. Indeed, neither of our games were likely enjoyable enough that a participant would consider doing them for their own sake. Though this was necessary in order to try to reduce the impact of gamification on the cognitive data, it likely reduced any effects of gamification we might've seen. Fifthly, the time course of our study, which took place over days, may not be informative about attrition in studies that take place over longer periods of weeks or months. Sixthly, as mentioned previously, there are issues relating to motivation and incentives, as in reality participants completing cognitive assessments will be presented with requests to complete a study over a fixed period for a fixed fee, and not with the option to continue for additional recompense. Finally, both incentives and reminders have been well established as effective methods of increasing engagement, and we used both in our study [87]. Although all three task variants had the same incentives and the same program of reminders (which stopped on day four), it is possible that these baseline engagement strategies acted as confounders, potentially muddying the effect of gamification on attrition.

In conclusion, the Theme variant had negative effects on the cognitive data and showed no clear evidence of reducing attrition. It was also rated as the least enjoyable and was the task switched away from most often. This suggests that themed gamelike tasks, at least those that use graphics alone, are non-optimal for use in cognitive assessment studies. In contrast, and replicating our previous finding [43], subjective ratings showed the Points variant to be well received. We found

SSRTs from the Points and Non-game variants to be equal, showing that points can be an effective way of increasing participant enjoyment of a cognitive task while still collecting valid data.

Despite differences in subjective ratings between the task variants, we saw no effect of gamification on participant attrition over the six-day optional testing period. Gamification has been promoted as a potential solution to engagement problems in both psychology and digital health care for several years, but we found no effect of gamification on *engagement as usage* in this case. The term gamification may have existed for a decade, but the formalization of gamification's implementation and effectiveness is only just beginning, and there is clearly further work to be done to understand how we can translate differences in subjective ratings to differences in usage.

## Acknowledgements

The authors gratefully acknowledge the artistic contribution of Melissa Groves who provided graphical resources for the themed task variant. The authors are members of the United Kingdom Centre for Tobacco and Alcohol Studies, a UKCRC Public Health Research: Centre of Excellence. Funding from British Heart Foundation, Cancer Research UK, Economic and Social Research Council, Medical Research Council, and the National Institute for Health Research, under the auspices of the UK Clinical Research Collaboration, is gratefully acknowledged. This work was supported by the Medical Research Council (MC_UU_12013/6 and MC_UU_12013/7), and a PhD studentship to JL funded by the Economic and Social Research Council and Cambridge Cognition Limited. The funders had no role in review design, data extraction and analysis, decision to publish, or preparation of the manuscript. JL and MM designed the study. JL programmed the task software and collected the data. JL analyzed the data. AS, NL and MM contributed to development of the manuscript.

## Conflicts of Interest

Dr Coyle is a Director of Handaxe CIC, a not-for-profit company that develops technology, including computer games, to support mental health interventions for children and adolescents.

## Abbreviations

BF: Bayes Factor

SSD: Stop Signal Delay

SSRT: Stop Signal Reaction Time

SST: Stop Signal Task

## References


1. Amazon MTurk (www.mturk.com) [Internet]. [cited 2017 Jul 11]. Available from: http://www.webcitation.org/6rsMtCfuJ

2. Prolific Academic (http://prolific.ac) [Internet]. [cited 2017 Jul 11]. Available from: http://www.webcitation.org/6rsN1syuF

3. Testable (www.testable.org) [Internet]. [cited 2017 Jul 11]. Available from: http://www.webcitation.org/6rsNDoV2R

4. Gorilla (https://gorilla.sc) [Internet]. [cited 2017 Jul 11]. Available from: http://www.webcitation.org/6rsNEk4Uj

5. Peer E, Brandimarte L, Samat S, Acquisti A. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. J Exp Soc Psychol 2017 May;70:153–163. [doi: 10.1016/j.jesp.2017.01.006] PMID: 28045305

6. Woods AT, Velasco C, Levitan CA, Wan X, Spence C. Conducting perception research over the internet: a tutorial review. PeerJ 2015;3:e1058. [doi: 10.7717/peerj.1058] PMID: 26244107

7. Crump MJC, McDonnell JV, Gureckis TM. Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. PLoS ONE 2013 Mar 13;8(3):e57410. [doi: 10.1371/journal.pone.0057410] PMID: 23516406

8. Griffiths F, Lindenmeyer A, Powell J, Lowe P, Thorogood M. Why Are Health Care Interventions Delivered Over the Internet? A Systematic Review of the Published Literature. J Med Internet Res [Internet] 2006 Jun 23;8(2). PMID: 16867965

9. Buhrmester M, Kwang T, Gosling SD. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? Perspect Psychol Sci J Assoc Psychol Sci 2011 Jan;6(1):3–5. PMID:26162106

10. Shapiro DN, Chandler J, Mueller PA. Using Mechanical Turk to Study Clinical Populations. Clin Psychol Sci 2013 Apr 1;1(2):213–220. [doi: 10.1177/2167702612469015]

11. Bennett GG, Glasgow RE. The delivery of public health interventions via the Internet: actualizing their potential. Annu Rev Public Health 2009;30:273–292. PMID:19296777

12. Birnbaum MH. Human research and data collection via the Internet. Annu Rev Psychol 2004;55:803–832. [doi: 10.1146/annurev.psych.55.090902.141601]

13. Etter J-F. Comparing the Efficacy of Two Internet-Based, Computer-Tailored Smoking Cessation Programs: A Randomized Trial. J Med Internet Res 2005;7(1):e2. [doi: 10.2196/jmir.7.1.e2] PMID: 15829474

14. Farvolden P, Denisoff E, Selby P, Bagby RM, Rudy L. Usage and Longitudinal Effectiveness of a Web-Based Self-Help Cognitive Behavioral Therapy Program for Panic Disorder. J Med Internet Res 2005;7(1):e7. [doi: 10.2196/jmir.7.1.e7] PMID: 15829479

15. Wangberg SC, Bergmo TS, Johnsen J-AK. Adherence in Internet-based interventions. Patient Prefer Adherence 2008 Feb 2;2:57–65. PMID:19920945

16. Kelders SM, Kok RN, Ossebaard HC, Van Gemert-Pijnen JE. Persuasive System Design Does Matter: A Systematic Review of Adherence to Web-Based Interventions. J Med Internet Res [Internet] 2012 Nov 14;14(6). PMID:23151820

17. Lodwick RK. Crossover designs: issues in construction, use, and communication [Internet] [Thesis]. Queen Mary University of London; 2016 [cited 2017 Sep 15]. Available from: http://qmro.qmul.ac.uk/xmlui/handle/123456789/12860

18. Dumville JC, Torgerson DJ, Hewitt CE. Reporting attrition in randomised controlled trials. BMJ 2006 Apr 22;332(7547):969–971. PMID:16627519

19. Zhou H, Fishbach A. The Pitfall of Experimenting on the Web: How Unattended Selective Attrition Leads to Surprising (Yet False) Research Conclusions. J Pers Soc Psychol 2016 Jun 13; [doi: 10.1037/pspa0000056] PMID: 27295328

20. Christensen H, Mackinnon A. The Law of Attrition Revisited. J Med Internet Res [Internet] 2006 Sep 29;8(3). PMID:17032636

21. Eysenbach G. The Law of Attrition. J Med Internet Res 2005 Mar 31;7(1):e11. [doi: 10.2196/jmir.7.1.e11] PMID: 15829473

22. Saul JE, Amato MS, Cha S, Graham AL. Engagement and attrition in Internet smoking cessation interventions: Insights from a cross-sectional survey of “one-hit-wonders.” Internet Interv 2016 Sep 1;5(Supplement C):23–29. [doi: 10.1016/j.invent.2016.07.001]

23. Guertler D, Vandelanotte C, Kirwan M, Duncan MJ. Engagement and Nonusage Attrition With a Free Physical Activity Promotion Program: The Case of 10,000 Steps Australia. J Med Internet Res 2015;17(7):e176. [doi: 10.2196/jmir.4339] PMID: 26180040

24. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2016 Dec 13;1–14. [doi: 10.1007/s13142-016-0453-1] PMID: 27966189

25. Lumsden J, Edwards EA, Lawrence NS, Coyle D, Munafò MR. Gamification of Cognitive Assessment and Cognitive Training: A Systematic Review of Applications and Efficacy. JMIR Serious Games 2016 Jul 15;4(2):e11. [doi: 10.2196/games.5888] PMID: 27421244

26. Primack BA, Carroll MV, McNamara M, Klem ML, King B, Rich M, Chan CW, Nayak S. Role of video games in improving health-related outcomes: a systematic review. Am J Prev Med 2012 Jun;42:630–638. [doi: 10.1016/j.amepre.2012.02.023] PMID: 22608382

27. Cugelman B. Gamification: What It Is and Why It Matters to Digital Health Behavior Change Developers. JMIR Serious Games 2013 Dec 12;1(1):e3. [doi: 10.2196/games.3139] PMID: 25658754

28. Deterding S, Sicart M, Nacke L, O'Hara K, Dixon D. Gamification. Using Game-design Elements in Non-gaming Contexts. CHI 11 Ext Abstr Hum Factors Comput Syst New York, NY, USA: ACM; 2011. p. 2425–2428. [doi: 10.1145/1979742.1979575]

29. Rigby S, Ryan RM. Glued to games: How video games draw us in and hold us spellbound. 2011;186. ISBN-10: 0313362246

30. King D, Greaves F, Exeter C, Darzi A. "Gamification": Influencing health behaviours with games. J R Soc Med 2013 Mar 1;106(3):76–78. [doi: 10.1177/0141076813480996] PMID: 23481424

31. Hawkins GE, Rae B, Nesbitt KV, Brown SD. Gamelike features might not improve data. Behav Res Methods 2013;45(2):301–318. [doi: 10.3758/s13428-012-0264-3] PMID: 23055169

32. McPherson, Burns. Gs Invaders: Assessing a computer game-like test of processing speed. Behav Res Methods 2007; PMID: 18183904

33. Miranda AT, Palmer EM. Intrinsic motivation and attentional capture from gamelike features in a visual search task. Behav Res Methods 2014 Mar;46(1):159–172. PMID:23835649

34. Prins PJM, Dovis S, Ponsioen A, ten Brink E, van der Oord S. Does computerized working memory training with game elements enhance motivation and training efficacy in children with ADHD? Cyberpsychology Behav Soc Netw 2011 Mar;14(3):115–122. [doi: 10.1089/cyber.2009.0206] PMID: 20649448

35. Tong T, Chignell M. Developing a Serious Game for Cognitive Assessment: Choosing Settings and Measuring Performance. Proc Second Int Symp Chin CHI 2014. p. 70–79. [doi: 10.1145/2592235.2592246]

36. Dörrenbächer S, Müller PM, Tröger J, Kray J. Dissociable effects of game elements on motivation and cognition in a task-switching training in middle childhood. Cognition 2014;5:1275. [doi: 10.3389/fpsyg.2014.01275] PMID: 25431564

37. Looyestyn J, Kernot J, Boshoff K, Ryan J, Edney S, Maher C. Does gamification increase engagement with online programs? A systematic review. PloS One 2017;12(3):e0173403. PMID:28362821

38. Brown M, O'Neill N, Woerden H van, Eslambolchilar P, Jones M, John A. Gamification and Adherence to Web-Based Mental Health Interventions: A Systematic Review. JMIR Ment Health 2016;3(3):e39. [doi: 10.2196/mental.5710] PMID: 27558893

39. Delisle J, Braun CMJ. A Context for Normalizing Impulsiveness at Work for Adults with Attention Deficit/Hyperactivity Disorder (Combined Type). Arch Clin Neuropsychol 2011;26:602–613. PMID: 21653627

40. Dovis, Oord, Wiers, Prins. Can Motivation Normalize Working Memory and Task Persistence in Children with Attention-Deficit/Hyperactivity Disorder? The Effects of Money and Computer-Gaming. J Abnorm Child Psychol 2011;40(5):669–681. [doi: 10.1007/s10802-011-9601-8] PMID: 22187093

41. Ninaus M, Pereira G, Stefitz R, Prada R, Paiva A, Neuper C, Wood G. Game elements improve performance in a working memory training task. Int J Serious Games [Internet] 2015 Feb 10 [cited 2016 Apr 27];2(1). [doi: 10.17083/ijsg.v2i1.60] PMID: 20649448

42. Katz B, Jaeggi S, Buschkuehl M, Stegman A, Shah P. Differential effect of motivational features on training improvements in school-based cognitive training. Front Hum Neurosci 2014;8:242. [doi: 10.3389/fnhum.2014.00242] PMID: 24795603

43. Lumsden J, Skinner A, Woods A, Lawrence N, Munafò M. The effects of gamelike features and test location on cognitive test performance and participant enjoyment. PeerJ 2016; PMID: 27441120

44. McPherson, Burns. Assessing the validity of computer-game-like tests of processing speed and working memory. Behav Res Methods 2008; PMID: 19001388

45. Effects of game mechanics on participant attrition, data quality and engagement in an online longitudinal cognitive assessment program [Internet]. [cited 2017 Jul 11]. Available from: https://osf.io/ysaqe/ DOI 10.17605/OSF.IO/YSAQE

46. WMA Declaration of Helsinki - Ethical Principles for Medical Research Involving Human Subjects [Internet]. 2013 [cited 2016 Jun 20]. Available from: http://www.wma.net/en/30publications/10policies/b3/index.html

47. Lumsden J. Mindgames (https://mindgames.firebaseapp.com) [Internet]. [cited 2017 Jul 11]. Available from: http://www.webcitation.org/6rsNag7Fy

48. Firebase (https://firebase.google.com) [Internet]. [cited 2017 Jul 11]. Available from: http://www.webcitation.org/6rsNEk4Uj

49. Pixi.JS (www.pixijs.com) [Internet]. [cited 2017 Jul 11]. Available from: http://www.webcitation.org/6rsNHAhV4

50. Logan GD. On the ability to inhibit thought and action: A users' guide to the stop signal paradigm. In: Dagenbach D, Carr TH, editors. Inhib Process Atten Mem Lang San Diego, CA, US: Academic Press; 1994. p. 189–239.

51. Logan GD, Cowan WB. On the ability to inhibit thought and action: A theory of an act of control. Psychol Rev 1984;91(3):295–327. [doi: 10.1037/0033-295X.91.3.295]

52. Verbruggen F, Logan GD. Response inhibition in the stop-signal paradigm. Trends Cogn Sci 2008;12(11):418–424. [doi: 10.1016/j.tics.2008.07.005] PMID: 18799345

53. Cantab Research Suite [Internet]. cambridgecognition.com. 2014 [cited 2014 Oct 20]. Available from: http://www.cambridgecognition.com/academic/products

54. CANTAB Stop Signal Task (http://www.cambridgecognition.com/cantab/cognitive-tests/executive-function/stop-signal-task-sst/) [Internet]. [cited 2017 Sep 13]. Available from: http://www.webcitation.org/6tRv9eeZj

55. Logan GD, Schachar RJ, Tannock R. Impulsivity and Inhibitory Control. Psychol Sci 1997;8(1):60–64. [doi: 10.1111/j.1467-9280.1997.tb00545.x]

56. Band GPH, van der Molen MW, Logan GD. Horse-race model simulations of the stop-signal procedure. Acta Psychol (Amst) 2003;112(2):105–142. [doi: 10.1016/S0001-6918(02)00079-3] PMID: 12521663

57. Boendermaker WJ, Prins PJM, Wiers RW. Cognitive Bias Modification for adolescents with substance use problems – Can serious games help? J Behav Ther Exp Psychiatry 2015 Dec;49, Part A:13–20. [doi: 10.1016/j.jbtep.2015.03.008] PMID: 25843611

58. Guitart-Masip M, Huys QJM, Fuentemilla L, Dayan P, Duzel E, Dolan RJ. Go and no-go learning in reward and punishment: interactions between affect and effect. NeuroImage 2012 Aug 1;62(1):154–166. [doi: 10.1016/j.neuroimage.2012.04.024] PMID: 22548809

59. Malone TW. What Makes Things Fun to Learn? Heuristics for Designing Instructional Computer Games. Proc 3rd ACM SIGSMALL Symp First SIGPC Symp Small Syst [Internet] New York, NY, USA: ACM; 1980 [cited 2016 May 17]. p. 162–169. [doi: 10.1145/800088.802839]

60. Malone TW. Toward a theory of intrinsically motivating instruction. Cogn Sci 1981 Oct;5(4):333–369. [doi: 10.1016/S0364-0213(81)80017-1]

61. Schell J. The Art of Game Design [Internet]. CRC Press; 2008 [cited 2015 Sep 11]. ISBN:978-0-12-369496-6

62. Andreou P, Neale BM, Chen W, Christiansen H, Gabriels I, Heise A, Meidad S, Muller UC, Uebel H, Banaschewski T, Manor I, Oades R, Roeyers H, Rothenberger A, Sham P, Steinhausen H-C, Asherson P, Kuntsi J. Reaction time performance in ADHD: improvement under fast-incentive condition and familial effects. Psychol Med 2007 Dec;37(12):1703–1715. [doi: 10.1017/S0033291707000815] PMID: 17537284

63. Garrett DD, MacDonald SWS, Craik FIM. Intraindividual reaction time variability is malleable: feedback- and education-related reductions in variability with age. Front Hum Neurosci 2012;6:101. [doi: 10.3389/fnhum.2012.00101] PMID: 22557954

64. Verbruggen F, Logan GD. Models of Response Inhibition in the Stop-Signal and Stop-Change Paradigms. Neurosci Biobehav Rev 2009;33(5):647–661. [doi: 10.1016/j.neubiorev.2008.08.014] PMID: 18822313

65. Hsu JK, Thibodeau R, Wong SJ, Zukiwsky D, Cecile S, Walton DM. A Wii bit of fun: The effects of adding Nintendo Wii® Bowling to a standard exercise regimen for residents of long-term care with upper extremity dysfunction. Physiother Theory Pract [Internet] 2011;27 PMID: 20698793

66. Lumsden J, Skinner A, Coyle D, Lawrence N, Munafò M. Dataset: No effect of gamification on attrition from a web-based longitudinal cognitive testing study [Internet]. data.bris, 2016. Available from: https://databris-ui.ilrt.bris.ac.uk/data/dataset/xxxxxxxxxxxxxxxxxxxxxxxxx

67. Berger JO, Sellke T. Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence. J Am Stat Assoc 1987 Mar 1;82(397):112–122. [doi: 10.1080/01621459.1987.10478397]

68. Blackwelder WC. “Proving the null hypothesis” in clinical trials. Control Clin Trials 1982 Dec;3(4):345–353. [doi: 10.1016/0197-2456(82)90024-1] PMID: 7160191

69. Wetzels R, Raaijmakers JGW, Jakab E, Wagenmakers E-J. How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t test. Psychon Bull Rev 2009 Aug;16(4):752–760. [doi: 10.3758/PBR.16.4.752] PMID: 19648463

70. Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G. Bayesian t tests for accepting and rejecting the null hypothesis. Psychon Bull Rev 2009 Apr;16(2):225–237. [doi: 10.3758/PBR.16.2.225] PMID: 19293088

71. Raftery AE. Bayesian Model Selection in Social Research. Sociol Methodol 1995;25:111–163. [doi: 10.2307/271063]

72. Jeffreys H. Theory of Probability. 3rd ed. Oxford: Clarendon Press; 1961. ISBN: 9780198503682

73. JASP Team T. JASP (Version 0.8.1.2)[Computer software] [Internet]. 2017. Available from: https://jasp-stats.org/

74. Deci, Ryan RM. Handbook of self-determination research. University Rochester Press; 2002. ISBN: 9781580461566

75. Sailer M, Hense JU, Mayr SK, Mandl H. How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. Comput Hum Behav 2017 Apr 1;69:371–380. [doi: 10.1016/j.chb.2016.12.033]

76. Boendermaker WJ, Maceiras SS, Boffo M, Wiers RW. Attentional Bias Modification With Serious Game Elements: Evaluating the Shots Game. JMIR Serious Games 2016;4(2):e20. [doi: 10.2196/games.6464] PMID: 27923780

77. Ryan RM, Rigby CS, Przybylski A. The Motivational Pull of Video Games: A Self-Determination Theory Approach. Motiv Emot 2006 Dec 1;30(4):344–360. [doi: 10.1007/s11031-006-9051-8]

78. Deci EL, Ryan RM, Koestner R. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychol Bull 1999;125(6):627–668. PMID: 10589297

79. Deterding S. Situated motivational affordances of game elements: A conceptual model. Gamification Using Game Des Elem Non-Gaming Contexts Workshop CHI 2011.

80. Khadjesari Z, Murray E, Kalaitzaki E, White IR, McCambridge J, Thompson SG, Wallace P, Godfrey C. Impact and Costs of Incentives to Reduce Attrition in Online Trials: Two

Randomized Controlled Trials. J Med Internet Res 2011 Mar 2;13(1):e26. [doi: 10.2196/jmir.1523] PMID: 21371988

81. Boyle EA, Connolly TM, Hainey T, Boyle JM. Engagement in digital entertainment games: A systematic review. Comput Hum Behav 2012 May;28(3):771–780. [doi: 10.1016/j.chb.2011.11.020]

82. Lee D, LaRose R. A Socio-Cognitive Model of Video Game Usage. J Broadcast Electron Media 2007;51:632–650.

83. Attali Y, Arieli-Attali M. Gamification in assessment: Do points affect test performance? Comput Educ 2015 Apr;83:57–63. [doi: 10.1016/j.compedu.2014.12.012]

84. Leotti LA, Wager TD. Motivational influences on response inhibition measures. J Exp Psychol Hum Percept Perform 2010 Apr;36(2):430–447. [doi: 10.1037/a0016802] PMID: 20364928

85. Mekler ED, Brühlmann F, Tuch AN, Opwis K. Towards understanding the effects of individual gamification elements on intrinsic motivation and performance. Comput Hum Behav [Internet] 2015 [cited 2016 Mar 4]; [doi: 10.1016/j.chb.2015.08.048]

86. Kato PM. What do you mean when you say your serious game has been validated? Experimental vs. Test Validity [Internet]. 2013. Available from: http://www.webcitation.org/6gt9POLIu

87. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, Eccles MP, Cane J, Wood CE. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med Publ Soc Behav Med 2013 Aug;46(1):81–95. PMID: 23512568