The current issue and full text archive of this journal is available on Emerald Insight at:
https://www.emerald.com/insight/0268-3946.htm



# Gamifying a situational judgment test with immersion and control game elements

## Effects on applicant reactions and construct validity

Received 11 October 2018
Revised 22 May 2019
12 June 2019
Accepted 18 December 2019

Richard N. Landers and Elena M. Auer
*Department of Psychology, University of Minnesota, Minneapolis, Minnesota, USA, and*
Joseph D. Abraham
*PSI Services LLC, Tulsa, Oklahoma, USA*

## Abstract

**Purpose** – Assessment gamification, which refers to the addition of game elements to existing assessments, is commonly implemented to improved applicant reactions to existing psychometric measures. This study aims to understand the effects of gamification on applicant reactions to and measurement quality of situational judgment tests.

**Design/methodology/approach** – In a 2 × 4 between-subjects experiment, this study randomly assigned 315 people to experience different versions of a gamified situational judgment test, crossing immersive game elements (text, audio, still pictures, video) with control game elements (high and low), measuring applicant reactions and assessing differences in convergent validity between conditions.

**Findings** – The use of immersive game elements improved perceptions of organizational technological sophistication, but no other reactions outcomes (test attitudes, procedural justice, organizational attractiveness). Convergent validity with cognitive ability was not affected by gamification.

**Originality/value** – This is the first study to experimentally examine applicant reactions and measurement quality to SJTs based upon the implementation of specific game elements. It demonstrates that small-scale efforts to gamify assessments are likely to lead to only small-scale gains. However, it also demonstrates that such modifications can be done without harming the measurement qualities of the test, making gamification a potentially useful marketing tool for assessment specialists. Thus, this study concludes that utility should be considered carefully and explicitly for any attempt to gamify assessment.

**Keywords** Selection, Research and development, Management skills

**Paper type** Research paper

## Introduction

Assessment gamification, which refers to the addition of game elements to existing assessments, is a relatively new but popular approach used to improve a variety of assessment outcomes through redesign inspired by analog and video game design (Armstrong *et al.*, 2016). Because gamification involves the integration of game design, which is typically studied by human–computer interaction researchers (see Salen and Zimmerman, 2004), with employee selection, which is typically studied by industrial-organizational psychology and human resources researchers, successful execution of gamification requires consulting and integrating knowledge

© Richard N. Landers, Elena M. Auer and Joseph D. Abraham. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial & non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at: http://creativecommons.org/licences/by/4.0/legalcode

*Funding statement:* PSI Services LLC provided participant payments for this study.

across disciplines. Thus, in the employee selection marketplace, gamification may be driven by people with backgrounds in either area or interdisciplinary teams of both, and this has led to dramatically different approaches being labeled "gamification" (cf. Chamorro-Premuzic *et al.*, 2016).

Armstrong *et al.* (2016) attempted to avoid the construct proliferation common in such situations by distinguishing between assessment gamification and game-based assessment, putting both under the broader heading of "game-thinking." Whereas assessment gamification involves the modification of an existing assessment by adding game elements, a game-based assessment requires full-fledged game development. For practitioners, these are dramatically different development processes. Game development involves the combination of many game elements simultaneously to create a standalone experience that people can "play." Typically, an initial prototype game is developed containing the *core gameplay loop*, the cyclically repeated behaviors in which players engage to progress from start to end (Salen and Zimmerman, 2004). That prototype is then refined iteratively until a final satisfactory gameplay experience has been developed (Macklin and Sharp, 2016) simultaneously with satisfactory psychometric characteristics. By contrast, gamification involves identifying and establishing linkages between specific targeted outcomes, psychological state changes likely to lead to those outcomes and game elements, defined as artifacts or social elements commonly used in games like narrative or leaderboards (Deterding *et al.*, 2011), likely to elicit those state changes (Landers *et al.*, 2020). Identified game elements are then added to an existing assessment and evaluated to test if the desired outcome, such as increased organizational attractiveness, was achieved without sacrificing any of the psychometric properties of the existing test to do so.

In the present study, we focus upon gamification of situational judgment tests (SJTs). Among common selection methods, SJTs are perhaps the most amenable to gamification because they rely upon one game element already: narrative. SJT questions typically consist of situational prompts and potential behavioral responses. More game elements can be added easily to a "basic" SJT; for example, a running narrative could be crafted across questions with alternative narrative pathways such that the person taking the SJT feels they progress in a storyline across the test. To do this successfully requires an understanding of how best to design and build a narrative to unfold with maximum affective impact when the test-taker is the central character in the narrative. Thus, gamification with narrative requires an expertise in narrative design, a subarea of game design. Any type of gamification requires a similar expertise within a relevant subarea of game design to maximize likely effectiveness; however, even with this expertise, there is no assurance of success without empirical evaluation and iterative improvement.

Given modern game-thinking, considering the interdisciplinary landscape of this research, and because one of the key advantages to assessment gamification in comparison to game-based assessment is reduced cost, we focus here upon evaluating the addition of game elements from two game element categories that are easy to grasp and implement within SJTs: immersion and control. First, immersion refers to game elements intended to draw a person into an experience. Immersion as a target of assessment gamification can be pursued in different ways, as long as the overall design goal is to increase test-taker perceptions of the assessment as immersive. One approach already commonly used to redesign SJTs for immersion is the use of audio or video prompts instead of text; instead of reading about a situation requiring their judgment, the test-taker sees and hears that situation. Second, control refers to the degree to which the test-taker can affect the environment they are in. In SJTs, this typically involves control over question order or presentation style. Thus, in this empirical study, we experimentally assigned test-takers to experience immersion and control elements to observe the effects of these changes on both

applicant reactions and the psychometric properties of an SJT developed by Industrial and Organizational (I-O) psychology practitioners and currently operational across the USA.

## Gamifying situational judgment tests for employee selection

SJTs, a measurement method in which applicants are asked how they would behave given a series of scenario prompts (Campion *et al.*, 2014), have a long history in I-O psychology and represent a popular method for employee selection with significant research support. SJTs are low-fidelity simulations that provide scenario prompts, typically paired with multiple choice response options, and respondents may be asked to choose one or to rank the effectiveness of each option (Lievens and De Soete, 2012; Weekley and Ployhart, 2013). These tests were originally designed as measures of judgment in work situations, but recent research has suggested that SJTs typically measure a more complex combination of capabilities (Lievens and Motowidlo, 2016; Whetzel and McDaniel, 2009).

Assessment gamification, which can be applied to SJTs, is a design process used to add game elements to existing measures or processes to meet specific system-level goals. Although SJTs are typically built using narrative game elements, this does not necessarily make them "gamified." An SJT is only "gamified" if it has been redesigned to add game elements beyond its original form; if it was designed with those elements initially, it was instead "gamefully designed" (Deterding, 2015), and the specific design skills required differ between these approaches. The gamification theory, which is depicted in Figure 1, illustrates how design choices and desired outcomes are related. Game elements are able to influence targeted outcomes through intermediary changes in psychological behaviors or states such as engagement, motivation and enjoyment. Gamification redesigns all share the use of game element(s); however, there are infinite specific ways to design game elements. Thus, there is no single "correct" approach to assessment gamification.

Given this, a common design goal for assessment gamification is the improvement of applicant reactions in an existing instrument without sacrificing measurement quality. For example, adding narrative or storyline game elements to a personality test has the potential to increase engagement and improve reactions, yet the addition of information for the assessee to read and recall throughout the assessment may also increase the cognitive load of the personality test, ultimately contaminating measurement. Sometimes, such tradeoffs are quite complex; for example, if an assessment is gamified in a way that allows applicants to progress through the test by making narrative decisions leading to different questions, not all

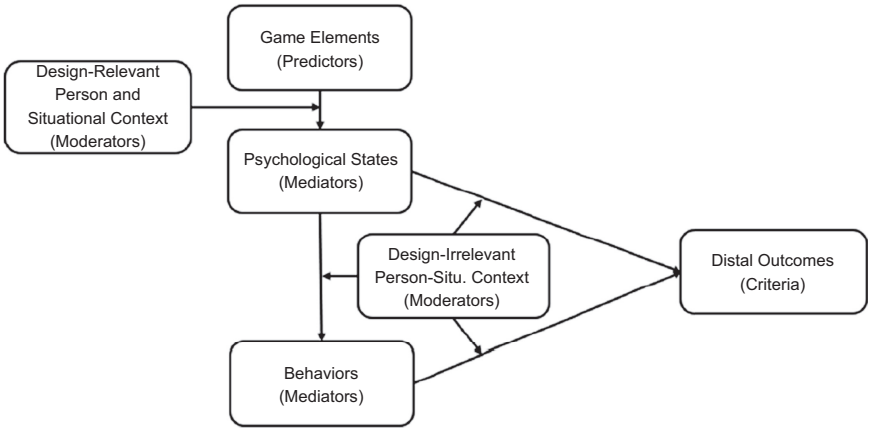


**Figure 1.**
Gamification theory
(from Landers
*et al.*, 2018)

applicants will take the same assessment, which alters key measurement characteristics of the test.

## Redesigning an SJT to improve applicant reactions

Considering such tradeoffs between applicant reaction gains and measurement quality losses, we identified two specific game element categories as particularly promising for the gamification of SJTs that became the focus of the present study: immersion and control. Bauer and Truxillo (2006) describe how SJTs can affect procedural justice rules and served as the basis for our identification of these element categories and their likely effects.

First, immersion game elements are intended to draw a person into an experience. In the context of an SJT, one way to do this is by increasing the fidelity of the stimulus (i.e. low immersion is a written scenario and high immersion is a video scenario). Previous literature explored the immersive nature of an SJT stimuli almost exclusively in the context of comparing video SJTs to text-based SJTs and generally found more positive applicant reactions toward video SJTs (Kanning *et al.*, 2006; Lievens and Sackett, 2006; Chan and Schmitt, 1997). Although some research has explored specific scenario stimuli, such as animation type and the use of live video versus animation (e.g. Bruk-Lee *et al.*, 2013), little research has conceptualized SJT media along the full continuum of choices practitioners currently typically face, a key research gap. As in prior research, we reasoned that procedural justice perceptions should improve through increased immersion, because it has been theorized that SJTs perceived as more job-related will appear fairer (Bauer and Truxillo, 2006), and because previous studies have found support for videos improving justice-related perceptions when compared to text. Furthermore, organizational attitudes should improve by increasing immersion, because initial applicant interactions with an organization (i.e. taking an assessment) tend to "spill over" into perceptions of the organization (Smither *et al.*, 1993).

*H1.* Integration of immersive elements in an SJT improves (a) applicant reactions to selection, specifically test attitudes and procedural justice perceptions, as well as (b) distal organizational attitudes, specifically organizational attractiveness and perceptions of organizational technological sophistication.

Second, control game elements are intended to increase feelings of autonomy. In assessment, this is realized as the degree to which the test-taker can affect aspects of their assessment environment. In high-control assessments, test-takers can control multiple aspects of the selection experience, whereas in low-control assessments, they cannot. Research in the training context has explored the idea of control more thoroughly than in selection; for example, learner control elements can increase motivation and learning (Garris *et al.*, 2002; Landers and Reddock, 2017) by increasing a user's sense of autonomy and competence (Deci and Ryan, 1985) and buy-in (Behrend and Thompson, 2012). Giving applicants greater control and therefore greater flexibility over how they use their time was predicted to improve reactions. Specifically, increased control should lead to improved motivational test attitudes by way of increases in expectancy, instrumentality and autonomy. Overall, procedural justice perceptions are expected to increase because adding control gives test-takers less structured and routinized assessment, increasing their input into the selection process and their perceived opportunity to perform well on the assessment (Bauer and Truxillo, 2006). Like immersion, we expected improved reactions to the test to "spill over" to organizational attitudes to improve perceptions of organizational technological sophistication.

*H2.* Integration of control elements in an SJT improves (a) applicant reactions to selection, specifically test attitudes and procedural justice perceptions, as well as (b)

distal organizational attitudes, specifically organizational attractiveness and perceptions of organizational technological sophistication.

### Potential effects of immersion and control on psychometric properties

As described earlier, an improvement in applicant reactions may not be worthwhile if it decreases psychometric quality. Accurate assessment of individual differences is a primary goal in designing selection systems, not to mention a legal requirement for systems in many nations (Myors *et al.*, 2008). We examined construct validity from two perspectives. First, we tested for mean differences in the test score itself across gamification types. Second, the test we examined had been validated by its publisher against cognitive ability and conscientiousness according to its test manual. Thus, we sought to determine if convergent validity coefficients differed by gamification design.

*RQ1.* Is performance on the SJT affected by the addition of immersion and control game elements?

*RQ2.* Is the convergent validity of the SJT with general cognitive ability and conscientiousness affected by the addition of immersion and control game elements?

## Method

### *Participants*

An *a priori* power analysis was conducted of the most statistically demanding analysis planned, which revealed 212 participants would be required for 80 percent power to detect small moderation effects ($\Delta R^2 > 0.05$). Ultimately, 315 participants in the USA were recruited from Mechanical Turk. After cleaning for careless responding on the basis of Mahalanobis distances (Meade and Craig, 2012), the final effective sample size was 240. Participants received US$4 for completing the 30–60 min study. To mimic a high-stakes environment, participants were told that the top ten scorers would be awarded an additional US$20, a sampling approach chosen to balance internal and external validity concerns (cf. Landers and Behrend, 2015). Specifically, using a non-applicant sample allowed for the experimental manipulation of game elements in the SJT, but mitigated some of the demographic skew of a college student sample. For example, gender was nearly evenly distributed (43.8 percent female), most participants (86.6 percent) in this study were employed and 44.58 percent of participants were between 30 and 39 years old.

### *Design*

Using a 2 × 4 between-subjects experiment, participants were randomly assigned among eight conditions. This resulted in final cell counts for low control of 28, 34, 36 and 32 participants in the text, audio, still and video conditions, respectively, and 34, 27, 29 and 20 for the high-control conditions in the same pattern.

*Control.* To maintain experimental power across many conditions while simultaneously maximizing the difference in applicant experiences between conditions, several distinct technologies varied together between low- and high-control conditions. In the high-control condition, participants were able to play, pause, rewind or skip ahead in the audio, video and still frame media conditions and choose the order in which they experienced the scenarios by selecting from an interactive menu of game-like icons (Figure 2). In the low-control condition, the SJT scenarios were completed in a fixed order and could not be repeated or rewound.


**Figure 2.**
Scenario selection menu in a high-control condition

*Immersion.* The four immersion conditions were designed to realistically reflect common levels of immersion typically found in SJTs. In the text condition, participants read a transcript containing narration and dialog of each SJT scenario. In the audio condition, participants listened to this content as an audio recording. In the still condition, participants listened to the same audio while viewing a still image depicting a drawing of the characters in each scenario. In the video condition, participants viewed a fully animated video. The content of the situational prompts was held constant.

*Interaction.* Although all "high-control" conditions contained high-control elements, the specific technical implementation varied slightly between immersion conditions. Specifically, there was no playback to be controlled when viewing text, so playback controls were not present in that experimental cell. However, the interactive menu was still present. Thus, "high" control was operationalized in terms of design context; in other words, "high" control was as high as achievable given technical limitations.

*Materials*

*Situational judgment test.* The SJT administered was an assessment product developed, validated and currently used by a large consulting and test development firm to evaluate candidates applying for customer service roles. Specifically, the SJT was designed to target self-management skills (e.g. professionalism, dependability) and interpersonal skills (e.g. social skills, empathy, responsiveness, service orientation). The test consisted of a series of nine customer-service-related scenarios with five questions per scenario, a total of 45. In the validation study conducted by the firm, scores on this SJT were found to correlate with

cognitive ability ($r = 0.25$), conscientiousness ($r = 0.46$), service orientation ($r = 0.46$) and annual salary growth ($r = 0.28$).

*Applicant reactions.* Motivational test attitudes were measured using the lack of concentration (LOC; our items), belief in test (BIT; four items) and test ease (TE; four items) factors of the test attitude survey (Arvey *et al.*, 1990) using a seven-point Likert-type agreement scale. Ultimately, two items were dropped from the TE scale because they did not co-vary strongly with any of the other three items on the scale.

*Procedural justice.* Perceptions of procedural justice were measured using the job relatedness–predictive (JR-P; two items), chance to perform (CP; four items), consistency of administration (CA; three items) and job relatedness–content (JR-C; two items) factors of the selection procedural justice scale (SPSJ; Bauer *et al.*, 2001) on a five-point Likert-type agreement scale. A composite score of the subscale means was used.

*Distal organizational attitudes.* Perceptions of organizational attractiveness were measured using the five-item general attractiveness subscale of the organizational attractiveness scale (Highhouse *et al.*, 2003), and perceptions of technological sophistication were measured using the three-item perceptions of organizational technological sophistication scale (Bauer *et al.*, 2004), both using a five-point Likert-type agreement scale.

*Convergent validation measures.* Conscientiousness was measured using the nine-item conscientiousness subscale of the big five inventory (John and Srivastava, 1999). Cognitive ability was measured with a composite of standardized subscale means using the 20-item numerical reasoning (EAS-6) and 30-item verbal reasoning (EAS-7) titles from the employee aptitude survey series (Ruch *et al.*, 2001). One item was dropped from EAS 6 due to high difficulty (0.96).

### *Procedure*

After recruitment on Mechanical Turk, participants were asked to imagine that they were applying for a customer service representative position at an office supply company, and a job description was supplied. They were also told that the top ten scorers on the test in this study would earn a US$20 bonus. Next, participants were randomly assigned to a condition and required to complete all nine SJT scenarios. After completing the SJT, participants were asked to complete the reaction measures. Lastly, participants completed the conscientiousness and cognitive ability measures.

## Results

Each of the continuous measures was assessed with a confirmatory factor analysis (CFA) with estimation by maximum likelihood. The SPSJ was modeled hierarchically ($\chi^2(39) = 52.06$, SRMR = 0.03, CFI > 0.99, RMSEA = 0.04), with JR-P and CP forming a structure factor, and this structure factor being modeled at the same level as JR-C and CA with no other freed paths, per the description of the measure's structure as outlined above. Both LOC ($\chi^2(1) = 4.18$, SRMR = 0.01, CFI > 0.99, RMSEA = 0.12) and BIT ($\chi^2(1) = 0.28$, SRMR = 0.03, CFI > 0.99, RMSEA = 0.04) were each modeled as a single-level CFA, in each case freeing error covariance between each scale's two negatively worded items. OA ($\chi^2(2) = 9.14$, SRMR = 0.01, CFI > 0.99, RMSEA = 0.12) was also modeled as a single-level CFA; however, upon examination of its factor structure, its single negatively worded item caused substantial misfit and was subsequently dropped from both the CFA and all future analyses. Conscientiousness ($\chi^2(23) = 45.16$, SRMR = 0.04, CFI = 0.98, RMSEA = 0.06) was modeled as a single-level CFA; however, a latent factor was also defined and loadings were freed on the four negatively worded items. Fit was deemed to be within tolerances for valid measurement.

Next, cognitive ability was modeled as two correlated latent EAS6 and EAS7 factors, utilizing a diagonally weighted least squares estimator ($\chi^2(1126) = 2998.16$, SRMR = 0.17,

CFI = 0.91, RMSEA = 0.08). This measure displayed substantially poorer fit than any of the continuous measures. However, because the cognitive ability composite was only being used to estimate convergent validity, it was important to ensure that the score used here was as similar in formulation as possible to the one in the test manual, which prevented any *post hoc* measure modification.

A table containing means, standard deviations and a correlation matrix appears in Table I. Compared to the test manual, the correlations between the SJT and general validation constructs were similar. It correlated similarly with cognitive ability ($r = 0.23$ vs. $r_{manual} = 0.25$) and with gender ($r = 0.22$ vs. $r_{manual} = 0.14$). The correlation with conscientiousness was much lower ($r = 0.14$ vs. $r_{manual} = 0.46$). Although the discrepancy here is large in an absolute sense, it may be explained by differing conscientiousness measure content; although we utilized the same cognitive ability measure as examined in the study reported in the test manual (i.e. a composite of the standardized EAS6 and EAS7 scores), we utilized a different conscientiousness measure. Specifically, the test manual validated the SJT score against conscientiousness using the general personality survey (GPS; Abraham and Morrison, 2010), which was also developed by the creators of the SJT, whereas we validated against the BFI, which is a research measure that correlates with the GPS at $r = 0.41$ ($r = 0.46$ corrected for unreliability). Thus, the two measures may weight facets of conscientiousness differently or otherwise produce different measurement profiles, which implies that a lower correlation between the SJT and BFI than between the SJT and GPS can likely be attributed at least in part to this relatively low convergent validity. Because we had *a priori* assumed that the GPS and BFI measured the same construct but did not find evidence to support this assumption, the trustworthiness of this difference as a diagnostic of validity differences attributable to gamification was reduced. Thus, we ultimately discarded it in our evaluation of convergence. This left the cognitive ability and gender convergent validities as the focal comparisons, which both supported the validity of the SJT as used in this study; we concluded that the SJT is likely measuring the same constructs as when the test is used in authentic selection contexts.

*Applicant reactions*

To test Hypotheses 1 and 2, it was necessary to first test for the presence of interactive effects of control and immersion in combination. A series of hierarchical multiple regressions were conducted testing incremental prediction using models containing interactive effects beyond main effects alone. We chose to conduct these as individual tests due to their individual hypothesized effects. These analyses appear in Table II. In no cases did the addition of interactive effects increase the predictive power of the models, so we restricted our interpretation to main effects models.

To test Hypotheses 1 and 2, the main effects models were in these tables were examined for statistically and practically significant model effects. In only one case did a model achieve statistical significance, which was for the organizational technological sophistication perceptions outcome. Practical significance appeared low, with only 4 percent of the variance in sophistication explained by condition main effects. Thus, Hypotheses 1a, 2a and 2b were not supported, whereas Hypothesis 1b had mixed support.

*Construct validity*

To address Research Question 1, we ran a similar hierarchical regression analysis as used to test the hypotheses by hierarchically regressing SJT performance on experimental conditions. This analysis also appears in Table II. On again, the interactive effect was not statistically significant, focusing our attention upon the main effects model. The main effects

**Table I.**
Descriptive statistics and correlation matrix for all study variables

| | | *M* | SD | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Control (yes) | 0.46 | 0.50 | | | | | | | | | | | | | | | |
| 2 | Immersion (audio) | 0.27 | 0.45 | −0.01 | | | | | | | | | | | | | | |
| 3 | Immersion (still) | 0.25 | 0.44 | −0.02 | −0.36 | | | | | | | | | | | | | |
| 4 | Immersion (video) | 0.22 | 0.41 | −0.08 | −0.32 | −0.31 | | | | | | | | | | | | |
| 5 | Immersion (text) | 0.26 | 0.44 | 0.11 | −0.36 | −0.34 | −0.31 | | | | | | | | | | | |
| 6 | Procedural justice | 4.39 | 0.46 | −0.07 | −0.04 | 0.12 | −0.02 | −0.06 | (0.69) | | | | | | | | | |
| 7 | LOC | 6.28 | 1.06 | 0.01 | 0.04 | −0.02 | 0.03 | −0.06 | 0.27 | (0.93) | | | | | | | | |
| 8 | Belief in test | 5.41 | 1.32 | −0.03 | −0.08 | 0.05 | 0.08 | −0.04 | 0.64 | 0.33 | (0.92) | | | | | | | |
| 9 | TE | 4.56 | 1.44 | −0.09 | 0.00 | −0.11 | 0.07 | 0.05 | 0.12 | 0.00 | 0.11 | (0.94) | | | | | | |
| 10 | Organizational attractiveness | 3.88 | 0.98 | −0.03 | 0.01 | 0.07 | 0.02 | −0.10 | 0.46 | 0.35 | 0.56 | −0.02 | (0.96) | | | | | |
| 11 | Organizational technological sophistication | 3.43 | 1.12 | −0.09 | −0.05 | 0.07 | 0.13 | −0.15 | 0.53 | 0.24 | 0.73 | 0.10 | 0.59 | (0.95) | | | | |
| 12 | Conscientiousness | 4.15 | 0.70 | −0.03 | −0.01 | 0.06 | −0.06 | 0.01 | 0.15 | 0.39 | 0.16 | −0.04 | 0.22 | 0.06 | (0.90) | | | |
| 13 | Cognitive ability | 0.00 | 0.85 | 0.00 | −0.08 | −0.05 | −0.04 | 0.17 | −0.03 | −0.03 | −0.14 | 0.04 | −0.24 | −0.23 | −0.14 | (0.61) | | |
| 14 | SJT score | 86.30 | 6.75 | 0.06 | −0.10 | 0.06 | 0.03 | 0.01 | 0.13 | 0.23 | −0.03 | −0.03 | 0.05 | −0.04 | 0.14 | 0.23 | | |
| 15 | Gender (female) | 0.45 | 0.50 | −0.03 | −0.02 | 0.01 | 0.06 | −0.04 | 0.12 | 0.03 | 0.04 | 0.04 | 0.05 | 0.05 | 0.02 | 0.00 | 0.22 | |
| 16 | Age | 36.63 | 9.86 | −0.06 | 0.09 | −0.08 | 0.00 | −0.01 | −0.12 | 0.09 | −0.13 | 0.07 | −0.09 | −0.07 | 0.11 | 0.08 | 0.27 | 0.18 |

**Note(s)**: All correlations ≥ |0.13| are statistically significant. Coefficient alpha reliability appears on the diagonal. In the case of procedural justice and cognitive ability, reliability is of the subscale means. Condition variables 1–5 in this table and all subsequent tables represent dummy codes used in analyses. Gender: 1 = Female, 0 = Male

**Table II.**
Hierarchical multiple regressions of study variables on experimental conditions

| | Procedural justice perceptions | | LOC | | Belief in tests | | TE | | Organizational attractiveness | | Technological sophistication | | SJT performance | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *B* | SE | *B* | SE | *B* | SE | *B* | SE | *B* | SE | *B* | SE | *B* | SE |
| *Main effects model* | | | | | | | | | | | | | | |
| (Intercept) | 4.37 | | 6.16 | | 5.35 | | 4.84 | | 3.74 | | 3.25 | | −0.05 | |
| Control (yes) | −0.06 | 0.06 | 0.03 | 0.14 | −0.07 | 0.17 | −0.26 | 0.19 | −0.04 | 0.13 | −0.17 | 0.14 | 0.12 | 0.13 |
| Immersion (audio) | 0.14 | 0.08 | 0.06 | 0.19 | 0.21 | 0.24 | −0.43 | 0.26 | 0.29 | 0.18 | 0.38 | 0.20 | 0.09 | 0.18 |
| Immersion (still) | 0.01 | 0.08 | 0.18 | 0.19 | −0.07 | 0.24 | −0.15 | 0.25 | 0.19 | 0.18 | 0.18 | 0.20 | −0.17 | 0.18 |
| Immersion (video) | 0.02 | 0.08 | 0.17 | 0.20 | 0.28 | 0.25 | 0.03 | 0.27 | 0.21 | 0.19 | 0.53* | 0.21 | 0.06 | 0.19 |
| *Interactive model* | | | | | | | | | | | | | | |
| (Intercept) | 4.40 | | 6.07 | | 5.33 | | 4.98 | | 3.69 | | 3.29 | | 0.14 | |
| Control (yes) | −0.12 | 0.12 | 0.20 | 0.27 | −0.03 | 0.34 | −0.52 | 0.37 | 0.03 | 0.25 | −0.24 | 0.29 | −0.22 | 0.26 |
| Immersion (audio) | 0.16 | 0.12 | 0.27 | 0.27 | 0.34 | 0.34 | −0.59 | 0.37 | 0.58* | 0.25 | 0.51 | 0.29 | −0.12 | 0.26 |
| Immersion (still) | −0.07 | 0.12 | 0.25 | 0.27 | −0.18 | 0.34 | −0.55 | 0.37 | 0.05 | 0.25 | −0.02 | 0.28 | −0.41 | 0.26 |
| Immersion (video) | −0.05 | 0.12 | 0.23 | 0.28 | 0.35 | 0.34 | 0.08 | 0.37 | 0.23 | 0.25 | 0.45 | 0.29 | −0.20 | 0.26 |
| Control × Immersion (audio) | −0.07 | 0.17 | −0.42 | 0.39 | −0.29 | 0.48 | 0.33 | 0.51 | −0.64 | 0.35 | −0.32 | 0.40 | 0.41 | 0.36 |
| Control × Immersion (still) | 0.18 | 0.17 | −0.14 | 0.38 | 0.26 | 0.48 | 0.84 | 0.51 | 0.33 | 0.35 | 0.42 | 0.39 | 0.46 | 0.36 |
| Control × Immersion (video) | 0.14 | 0.17 | −0.07 | 0.41 | −0.18 | 0.50 | −0.24 | 0.55 | −0.00 | 0.37 | 0.18 | 0.42 | 0.53 | 0.38 |
| *Model summaries* | | | | | | | | | | | | | | |
| Main effects $R^2$ | 0.02 | | 0.01 | | 0.01 | | 0.02 | | 0.02 | | 0.04* | | 0.01 | |
| Interactive $R^2$ | 0.03 | | 0.01 | | 0.02 | | 0.04 | | 0.05 | | 0.05 | | 0.02 | |
| Change in $R^2$ | 0.01 | | 0.00 | | 0.01 | | 0.02 | | 0.04 | | 0.01 | | 0.01 | |

**Note(s)**: *$p < 0.05$

model was also not statistically significant. Thus, it does not appear that gamification changed the difficulty of the test.

To address Research Question 2, we conducted two additional hierarchical multiple regression analyses, testing for moderation of the relationship between these variables and SJT performance by experimental design. Thus, we regressed the validation construct onto SJT performance and all media conditions in one model, then looked for incremental validity associated with the addition of condition by SJT performance interactions. These analyses appear in Table III. In neither case was the relationship between SJT performance and the validation construct moderated by experimental condition yet SJT performance itself did correlate with the covariates as expected; thus, it appears that the construct validity of the test was not dramatically affected by the addition of these game elements.

## Discussion

Overall, our results were mixed regarding our hypothesis tests. Although the effect of increased immersion on perceptions of organizational technological sophistication was statistically significant, the magnitude of the effect was relatively small ($R^2 = 0.04$). Other effects were nearer to zero (ranging $R^2 = 0.01–0.02$). Given this discrepancy, it is useful to explore why this particular effect appeared. The most self-evident interpretation is that the use of high-immersion elements leads to a general perception of greater technological sophistication. However, this might also be explainable via a novelty effect; specifically, if an animated video is relatively unusual in the marketplace of selection measures to the population we sampled, this novelty may have driven the effect instead of the game elements themselves. This distinction is important because it implies for future researchers and current

| | Conscientiousness | | | | Cognitive ability | | | |
|---|---|---|---|---|---|---|---|---|
| | Model 1 | | Model 2 | | Model 1 | | Model 2 | |
| | *B* | SE | *B* | SE | *B* | SE | *B* | SE |
| *Main effects predictors* | | | | | | | | |
| (Intercept) | 4.22 | | 4.14 | | 0.23 | | 0.26 | |
| Control (yes) | 0.10* | 0.05 | 0.39* | 0.16 | 0.20* | 0.05 | −0.01 | 0.16 |
| Immersion (audio) | −0.02 | 0.18 | 0.02 | 0.18 | 0.00 | 0.21 | −0.04 | 0.19 |
| Immersion (still) | 0.16 | 0.18 | 0.20 | 0.18 | −0.27 | 0.21 | −0.30 | 0.21 |
| Immersion (video) | −0.09 | 0.18 | −0.04 | 0.18 | −0.31 | 0.21 | −0.34 | 0.21 |
| Control × Immersion (audio) | −0.05 | 0.18 | −0.03 | 0.18 | −0.27 | 0.21 | −0.31 | 0.21 |
| Control × Immersion (still) | −0.26 | 0.25 | −0.31 | 0.25 | −0.12 | 0.30 | 0.05 | 0.22 |
| Control × Immersion (video) | 0.17 | 0.25 | 0.13 | 0.25 | 0.00 | 0.30 | 0.29 | 0.23 |
| *Moderation predictors* | | | | | | | | |
| SJT Score × Control | | | −0.34 | 0.20 | | | 0.07 | 0.21 |
| SJT Score × Immersion (audio) | | | −0.29 | 0.21 | | | −0.10 | 0.25 |
| SJT Score × Immersion (still) | | | −0.28 | 0.18 | | | 0.04 | 0.30 |
| SJT Score × Immersion (video) | | | −0.51* | 0.21 | | | −0.11 | 0.32 |
| SJT score × Control × Immersion (audio) | | | 0.37 | 0.27 | | | −0.04 | 0.32 |
| SJT score × Control × Immersion (still) | | | 0.43 | 0.26 | | | 0.11 | 0.31 |
| SJT score × Control × Immersion (video) | | | 0.39 | 0.32 | | | 0.27 | 0.38 |
| *Model summaries* | | | | | | | | |
| Model $R^2$ | 0.04 | | 0.07 | | 0.08* | | 0.11* | |
| Change in $R^2$ | | | 0.04 | | | | 0.03 | |

**Note(s)**: *$p < 0.05$

**Table III.** Moderation analyses assessing effects of experimental conditions on convergent validity

practitioners that if an aura of technological sophistication is an important goal in a selection system, achieving that may be a moving target, requiring consistent innovation.

More in line with our expectations, the addition of game elements to the text version of this SJT did not dramatically alter its measurement characteristics. In our experience, this is a significant fear of practitioners; specifically, there is worry that external pressure to gamify their assessments will lead them to sacrifice measurement quality. We have demonstrated here that although this is a risk, it can be managed and avoided. Importantly, it should not be assumed that measurement quality is safe by default when these game element categories are used; we recommend local validation studies to verify this for each implementation.

The game elements used in the redesign described here targeted immersion but do not differ substantially in form or function from the elements examined in traditional SJT media effects studies. In this sense, the effects observed here for media are a replication of past work, although we did not successfully replicate those findings. This was surprising to our team. Our immediate interpretation was simply that the use of video on a webpage is less novel than it was when much of the core work on multimedia assessment occurred, over a decade ago (e.g. Whetzel and McDaniel, 2009). Even in Oostrom *et al.*'s (2015) recent review of SJT research, all cited research on reaction to media was conducted between 1997 and 2006. This change in test-taker expectations as technology has matured may have similarly weakened the effect of our control manipulation; it simply may not have seemed very immersive or novel, considering the highly immersive experiences and high levels of control many people have on the internet on a daily basis today. In a sense, any directed behavior, such as being required to complete an online assessment, may now be viewed as limiting control. Manipulation checks assessing perceptions of control and immersion would have helped assess this, and we strongly recommend checks of this type in future research.

Assessment gamification is most usefully considered a redesign process, one concerned with adding new game-like characteristics to existing assessments, and not as any one change in technology, because desirable game elements will vary by context and by the characteristics of the assessment being gamified. We chose the elements that we reasoned, based upon prior literature, would have the greatest effects for this existing SJT, designed specific implementations of those elements and then tested their efficacy. We view documentation of that decision-making and design process to be the most critical contribution of this study to the assessment literature.

### *Limitations and future directions*

Immersion and control elements do not represent the full gamut of possibilities for gamifying selection in general or SJTs in particular, as specific representations of these game element categories vary by design and by purpose. Thus, the present study is not intended and should not be interpreted to imply that these two element categories, in general, have particular effects. Instead, we tested only prototypical implementations of those categories. Adopting a lens of *design thinking* distinguishing carefully between implementation characteristics is critical for psychologists to accurately understand the effects of technology implementations (Landers, 2019). Whereas a personality trait theoretically exists and affects outcomes, whether it is measured or not, a technology is designed and implemented with a specific purpose, so its specific design characteristics alter its effects. Future research should explore alternative implementation strategies for immersion and control elements within SJTs, as well as a broader array of game element categories. Immersion, for example, could be further targeted by stylizing the webpage surrounding the video to resemble a workplace in the same theme as the SJT prompts, using novel graphics and interactive Web components. Additionally, SJTs already incorporate narrative; larger gains in reactions might be found in contexts where no game elements are already present.

To address this technological landscape, we recommend a combination of both exploratory and confirmatory research in pursuit of developing theories of technological implementations in selection. Null effects in either context, assuming adequate power, are to be expected and considered carefully. Although it is traditionally more difficult to publish null results in managerial psychology (Kepes and McDaniel, 2013), most technologies are untested in rigorous peer-reviewed research, and we suspect that many researchers are even unwilling to study to technologies because of the greater risk of "failure." The only way to build a literature about such technologies, which is the only way to build a complete science of managerial psychology as it is actually practiced, is to abandon this counterproductive approach. Oversimplification of complex technological landscapes will only serve to minimize the impact, importance and relevance of this research. If the only published studies on the implementation of new technologies are the "successful" ones, the quality of advice and guidance provided both for future research and for practice are essentially zero. One must know how to design unsuccessful technological systems to design successful ones.

In testing this system, it is notable that we utilized Mechanical Turk as our source of research participants. Although we articulated a specific case for this, balancing internal and external validity concerns, and although we used incentives to simulate a selection context, the use of such a sample does still bring with it somewhat different motivational characteristics and outcome variables than an authentic selection context (Landers and Behrend, 2015). Additionally, we did not include a manipulation check to assess the motivational effect of the incentive level we chose, nor did we assess manipulation checks that the immersion and control elements actually elicited perceptions of immersion and control. Future research should endeavor to examine game elements quasi-experimentally in applicant contexts where applied to examine these questions with different priorities to determine if results across approaches converge on a single set of conclusions and should carefully consider and target meaningful secondary outcome variables, like game element perceptions.

## Conclusion

In summary, it appears that gamification with high immersion elements may be an expensive way to achieve a relatively small gain in applicant reactions for SJTs, most specifically of perceptions of organizational technological sophistication. Although the addition of control elements is quite inexpensive and did not affect measurement, it also was not associated with practically significant gains in reactions. Thus, we recommend organizations carefully consider their goals when gamifying SJTs and conduct a utility analysis to determine if likely gains are justified given design costs. We have demonstrated here that relatively small changes to the assessment experience are likely to also lead to small gains in reactions. If a dramatic improvement in reactions is needed, the most likely way to achieve that appears to be through fundamental and transformative change in the system itself. As such, assessment gamification of SJTs may be best considered in the marketplace a "style" of assessment; for example, organizational leadership may want to choose how "fun" they want their selection system to be as an organizational artifact, more of a culture choice than a design component with specific utility. This may have marketing advantages, if nothing else. Having said that, a real-world gamefully designed *system* might be able to achieve more positive reactions much more readily than gamifying a system that already exists. Gamification designers are constrained by the existing properties of in-place assessments, whereas system designers can start from scratch, giving much greater flexibility to implement game elements in a practical and efficient way; this possibility is left for future research.

## References

Abraham, J.D. and Morrison, J.D. (2010), *Viewpoint general personality survey (GPS)TM*, Technical manual version 8, PSI Services LLC, Burbank, California.

Armstrong, M.B., Ferrell, J., Collmus, A.B. and Landers, R.N. (2016), "Correcting misconceptions about gamification of assessment: more than SJTs and badges", *Industrial and Organizational Psychology*, Vol. 9, pp. 671-677.

Arvey, R.D., Strickland, W., Drauden, G. and Martin, C. (1990), "Motivational components of test taking", *Personnel Psychology*, Vol. 43, pp. 695-716.

Bauer, T.N. and Truxillo, D.M. (2006), "Applicant reactions to situational judgment tests: research and related practical issues", *Situational Judgment Tests: Theory, Measurement, and Application*, pp. 233-249.

Bauer, T.N., Truxillo, D.M., Paronto, M.E., Weekley, J.A. and Campion, M.A. (2004), "Applicant reactions to different selection technology: face-to-face, interactive voice response, and computer-assisted telephone screening interviews", *International Journal of Selection and Assessment*, Vol. 12, pp. 135-148.

Bauer, T.N., Truxillo, D.M., Sanchez, R.J., Craig, J.M., Ferrara, P. and Campion, M.A. (2001), "Applicant reactions to selection: development of the selection procedural justice scale (SPJS)", *Personnel Psychology*, Vol. 54, pp. 387-419.

Behrend, T.S. and Thompson, L.F. (2012), "Using animated agents in learner-controlled training: the effects of design control", *International Journal of Training and Development*, Vol. 16, pp. 263-283.

Bruk-Lee, V., Drew, E.N. and Hawkes, B. (2013), "Candidate reactions to simulations and media-rich assessments in personnel selection", in *Simulations for Personnel Selection*, Springer, New York, NY, pp. 43-60.

Campion, M.C., Ployhart, R.E. and MacKenzie, W.I., Jr (2014), "The state of research on situational judgment tests: a content analysis and directions for future research", *Human Performance*, Vol. 27 No. 4, pp. 283-310.

Chamorro-Premuzic, T., Winsborough, D., Sherman, R.A. and Hogan, R. (2016), "New talent signals: shiny new objects or a brave new world?", *Industrial and Organizational Psychology: Perspectives on Science and Practice*, Vol. 9, pp. 621-640.

Chan, D. and Schmitt, N. (1997), "Video-based versus paper-and-pencil method of assessment in situational judgment tests: subgroup differences in test performance and face validity perceptions", *Journal of applied psychology*, Vol. 82 No. 1, p. 143.

Deci, E.L. and Ryan, R.M. (1985), *Intrinsic motivation and self-determination in human behavior*, Plenum publishing, New York.

Deterding, S. (2015), "The lens of intrinsic skill atoms: a method for gameful design", *Human-Computer Interaction*, Vol. 30, pp. 294-335.

Deterding, S., Dixon, D., Khaled, R. and Nacke, L. (2011, September), "From game design elements to gamefulness: defining gamification", in *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, ACM, pp. 9-15.

Garris, R., Ahlers, R. and Driskell, J.E. (2002), "Games, motivation, and learning: a research and practice model", *Simulation and Gaming*, Vol. 33 No. 4, pp. 441-467.

Highhouse, S., Lievens, F. and Sinar, E.F. (2003), "Measuring attraction to organizations", *Educational and Psychological Measurement*, Vol. 63 No. 6, pp. 986-1001.

John, O.P. and Srivastava, S. (1999), "The Big Five trait taxonomy: history, measurement, and theoretical perspectives", *Handbook of Personality: Theory and Research*, No. 2, pp. 102-138.

Kanning, U.P., Grewe, K., Hollenberg, S. and Hadouch, M. (2006), "From the subjects' point of view", *European Journal of Psychological Assessment*, Vol. 22 No. 3, pp. 168-176.

Kepes, S. and McDaniel, M.A. (2013), "How trustworthy is the scientific literature in industrial and organizational psychology?", *Industrial and Organizational Psychology Perspectives*, Vol. 6, pp. 252-268.

Landers, R.N. (2019), "The existential threats to I-O psychology highlighted by rapid technological change", in Landers, R.N. (Ed.), *Cambridge Handbook of Technology and Employee Behavior*, Cambridge University Press, New York, NY, pp. 3-21.

Landers, R.N., Auer, E.M., Collmus, A.B. and Armstrong, M.B. (2018), "Gamification science, its history and future: Definitions and a research agenda", *Simulation & Gaming*, Vol. 49 No. 3, pp. 315-337.

Landers, R.N. and Behrend, T.S. (2015), "An inconvenient truth: arbitrary distinctions between organizations, mechanical turk, and other convenience samples", *Industrial and Organizational Psychology*, Vol. 8, pp. 142-164.

Landers, R.N. and Reddock, C.M. (2017), "A meta-analytic investigation of objective learner control in web-based instruction", *Journal of Business and Psychology*, Vol. 32, pp. 455-478.

Landers, R.N., Tondello, G.F., Kappen, D.L., Collmus, A.B., Mekler, E.D. and Nacke, L.E. (2020), "Defining gameful experience as a psychological state caused by gameplay: replacing the term 'gamefulness' with three distinct constructs", *International Journal of Human-Computer Studies*, in press.

Lievens, F. and De Soete, B. (2012), "Simulations", in Schmitt, N. (Ed.), *Handbook of Assessment and Selection*, Oxford University Press, pp. 383-410.

Lievens, F. and Motowidlo, S.J. (2016), "Situational judgment tests: from measures of situational judgment to measures of general domain knowledge", *Industrial and Organizational Psychology*, Vol. 9 No. 1, pp. 3-22.

Lievens, F. and Sackett, P.R. (2006), "Video-based versus written situational judgment tests: a comparison in terms of predictive validity", *Journal of Applied Psychology*, Vol. 91 No. 5, p. 1181.

Macklin, C. and Sharp, J. (2016), *Games, Design and Play: A Detailed Approach to Iterative Game Design*, Pearson Education, London.

Meade, A.W. and Craig, S.B. (2012), "Identifying careless responses in survey data", *Psychological Methods*, Vol. 17, pp. 437-455.

Myors, B., Lievens, F., Schollaert, E., Van Hoye, G., Cronshaw, S.F., Mladinic, A., . . . and Sackett, P.R. (2008), "International perspectives on the legal environment for selection", *Industrial and Organizational Psychology Perspectives*, Vol. 1, pp. 206-246.

Oostrom, J.K., De Soete, B. and Lievens, F. (2015), "Situational judgment testing: a review and some new developments", *In employee recruitment, selection and assessment*, Psychology Press, pp. 184-201.

Ruch, W.W., Stang, S.W., McKillip, R.H. and Dye, D.A. (2001), *Employee Aptitude Survey: Technical Report*, Psychological Services, Glendale, CA.

Salen, K. and Zimmerman, E. (2004), *Rules of Play: Game Design Fundamentals*, The MIT Press, Cambridge, MA.

Smither, J.W., Reilly, R.R., Millsap, R.E., Pearlman, K. and Stoffey, R.W. (1993), "Applicant reactions to selection procedures", *Personnel Psychology*, Vol. 46, pp. 49-76.

Weekley, J.A. and Ployhart, R.E. (2013), *Situational Judgment Tests: Theory, Measurement, and Application,* Psychology Press, Hove, UK.

Whetzel, D.L. and McDaniel, M.A. (2009), "Situational judgment tests: an overview of current research", *Human Resource Management Review*, Vol. 19, pp. 188-202.

**Corresponding author**
Richard N. Landers can be contacted at: rlanders@umn.edu

For instructions on how to order reprints of this article, please visit our website:
**www.emeraldgrouppublishing.com/licensing/reprints.htm**
Or contact us for further details: **permissions@emeraldinsight.com**