Grouches, Extraverts, and Jellyfish: Assessment validity and game mechanics in a gamified assessment

Laura Levy, Rob Solomon, Jeremy Johnson, Jeff Wilson, Amy Lambeth, Maribeth Gandy
Georgia Institute of Technology
85 5th Street NW
Atlanta, Georgia, USA 30308
+1-404-894-4728
[laura, rob, jeremy, jeff, amy, maribeth]@imtc.gatech.edu

Joann Moore, Jason Way, Ruitao Liu
ACT, Inc.
500 ACT Drive
Iowa City, Iowa, USA 52243
+1-319-337-1499
[joann.moore, jason.way, ruitao.liu]@act.org

ABSTRACT
Research into the use of both commercial and custom video games to assess individual differences, like personality, of players has revealed promising results. Virtual environments can allow researchers to analyze a variety of player behaviors and actions that correlate strongly with inherent personality traits. What is less understood is how an assessment game's mechanics might affect a player's inputs that determine the assessment's validity. In this study, we developed a custom game and logging framework for an online study assessing the reliability and validity of transferring a traditional personality questionnaire into a game environment. The game was played by 212 college-aged participants in one of three conditions. The conditions represented different levels of game mechanics; including enemies and point earning. Using results from a traditional personality assessment as our ground truth, we compared player responses and play behavior in the game. We found that responses between the traditional assessment and game-based assessment in all conditions were consistent, indicating that the game mechanics did not interfere or alter significantly a player's ability or decision to make personality-based responses. Additionally, we found several gameplay behaviors that can be used as predictors of individual differences.

Keywords
Assessment, individual differences, psychology, personality, gaming

Proceedings of 1st International Joint Conference of DiGRA and FDG

INTRODUCTION

Video games have expanded beyond their origins as pure entertainment and found themselves a promising means for a variety of purposed applications, like cognitive training (Basak et al. 2008; Nouchi et al. 2013), pain control therapy (Hoffman et al. 2008), and learning (Shute et al. 2009). Another area of promising research for games is in assessment.

There are many types of assessments that are employed to measure performance, knowledge, skills, beliefs and attitudes of individuals and groups. Traditional assessments are often pen-and-paper based (e.g., multiple choice) tests, though many are now taken digitally on a computer. Some of these assessments, like personality inventories, have been administered for decades or longer and there exists a wealth of knowledge into their reliability and validity. However, traditional assessments also pose some limitations. For example, traditional assessments can introduce unnecessary stress to the test-taker leading to responses that are inaccurate reflections of the person's knowledge or traits (Sarason, 1961; Zatz & Chassin, 1985). Relying on self-report questionnaires, common for personality inventories, can invite test-takers to over-exaggerate certain qualities of themselves to appear more desirable (Holtgraves, 2004; Paulhus, 1984). Additionally, traditional assessments are often limited by their very format in measuring certain domains (e.g., providing limited response options to measure a person's creativity; Kaufman et al. 2007).

The assessment research literature has long studied the impact of motivation on performance (Finn, 2015; Pintrich & DeGroot, 1990). Unmotivated or unengaged examinees may rush through the questions, answer randomly, or disengage before the test is complete; this lack of motivation particularly influences the outcome of "low stakes" tests (Wise & Kong, 2005). An analysis of 25 sets of comparisons found an average effect size of 0.59 standard deviations difference in group means between motivated and unmotivated examinees (Wise & DeMars, 2005). Games, on the other hand, present an interesting opportunity to benefit from the wealth of knowledge that has gone into traditional assessment design, while also potentially solving some of the limitations that exist for traditional tests. Recent games user research has indicated promising results that games can be employed as engaging and accurate means of assessing cognitive and non-cognitive measures of individuals (Levy et al. 2015; Spronck et al. 2012; Tekofsky et al. 2013).

As interest in assessment games increases, there is a growing need to understand the design of scientifically valid and accurate assessment games. Currently, what is lacking from this research body is a further understanding of the reliability and validity of game-acquired results on measured variables, as compared to the same variables measured in a traditional format. Additionally, research is needed into what effects, if any, a game and its mechanics might exert over a player's measured variables. Games present rich and engaging environments from which researchers can capture an abundance of player data. However, to make accurate assessment games we must first understand how the game might alter someone's ability or choices in response making that can be used to assess the player.

In this study, we examine the differences between results from a traditional assessment of personality and game-based results using a custom game that these authors developed, called *Bubble Trip*.

RELATED WORK

Most research into assessment games has centered on looking for relationships between results on traditional assessments and game play behavior. In particular, a lot of research has focused on looking at how gameplay behavior correlates with individual differences, like the personality of the player. Personality can be considered an outward expression of one's stable attributes. The most popular model for describing these attributes is the five-factor model (Wiggins, 1996). The five factors include Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Each factor is made of constituent sub-facets, or traits. For example, Extraversion is made up of other qualities including assertiveness, gregariousness, and excitement seeking behavior (Goldberg, 1999; Matthews et al. 2003).

An advantage of using a game to measure player personality is that a virtual world can provide a rich environment for a player to express a variety of different behaviors. Ideally, the unique types and chains of choices, responses, and behaviors a player makes should indicate something about themselves and their personality. Additionally, traditional personality inventories are most often administered as self-report via pen-and-paper. Test-takers wishing to represent themselves in a more positive light have no trouble taking advantage of the test's wording and choosing the answers that seem more desirable. An assessment game, however, has the potential to hide the assessment within the game play. This technique of "stealth assessment" (Shute, 2011; Shute et al. 2009) obfuscates what is being measured and prevents players from attempting to over-represent themselves.

One of the largest-scaled studies examining personality links with game-play behavior was conducted by Tekofsky et al. (2013), and compared results from the IPIP Big Five personality test and game play statistics from *Battlefield 3* (EA, 2011). Over 13,000 participants' game-play data was analyzed against their traditionally collected answers from the personality test. These authors found a number of correlations between play style and personality, concluding that player personality does manifest in the way a person plays a game. For example, they found the personality dimension of Conscientiousness (proclivity towards self-discipline and achievement outside of external expectations) to be a predictor for action speed within the game.

Promising results for linking personality dimensions with gameplay variables has been demonstrated for a variety of commercial games, including *Neverwinter Nights* (BioWare, 2002; van Lankveld et al. 2011) , *Fallout 3* (Bethesda Game Studios, 2008; Spronck et al. 2012), and *World of Warcraft* (Blizzard, 2014; Drachen et al. 2014). Thanks to the increasing partnerships between research groups and game companies, researchers have access to analyze more player data than ever before.

Research has proven promising in linking gameplay behavior to personality in non-commercial games, as well. The benefit to researchers creating their own games is that they can contrive specific situations and settings to try to elicit certain kinds of behaviors and gameplay that might be most helpful for assessing the player. Levy et al. (2015) used a custom-built game, *Food for Thought*, and its in-game logging capabilities to look for relationships between scores on a cognitive multi-tasking assessment, Big Five Inventory 44-item (BFI-44) personality test, and an assessment of college students' academic behaviors developed by ACT, Inc. (i.e., ACT Engage). The number of level retries initiated by a player was found to be the variable with the most associations with the other traditional assessments. For example, higher scores in the BFI-44 dimension of

Agreeableness were found to be associated with lower numbers of retries. Additionally, higher numbers of retries were correlated with higher scores on the Math component of the cognitive multi-tasking assessment, and lower scores on ACT Engage for the domains of Self-Confidence, General Determination, and Study Skills. This study presents promising results in linking game-play behavior with both cognitive and non-cognitive traits of an individual.

STUDY DESCRIPTION

Research has established that game-play behavior has the potential to assess certain dimensions of personality. However, what is currently less known in the literature is how the very format of a game meant for assessment might affect a player's responses. We hypothesize there are two main kinds of effects a game might potentially exert on a player and the ability to scientifically correlate their play behavior with measures of individual differences. First, the game mechanics themselves likely present some noise affecting a player's skill and ability to complete tasks within the game. Second, the game itself could affect the player's internal state, like their positive or negative affect, and thereby change behaviors that might be used to assess them on some dimension of their personality.

The motivation of this research is to investigate how a game and its mechanics might influence a person's responses to personality inventory questions presented within the game. Our research goal was to see what effects, if any, on question responses that game elements like score collecting, character navigation, and enemy avoidance might exist. To do so, we administered an assessment of personality in a traditional format and used those responses as a "ground truth" for that individual's personality. Then, participants answered these same questions within a game environment that included varying types of game mechanics. We examined the differences in these responses to gain some understanding of the game-based assessment's reliability and validity. We also performed analyses to see what personality traits might be predictors of game-play behavior and strategy.

METHODS

Materials

Personality Inventory

We used the 60-item version of the HEXACO Personality Inventory – Revised to measure player personality in its traditional format (Ashton & Lee, 2009). The HEXACO is based on the five-factor model of personality but includes slightly different descriptions of the five original domains, and introduces a sixth. The domains measured by this inventory are Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience.

The game

Bubble Trip is a game designed by a collaboration of computer scientists, game developers, graphic artists, and psychologists between an academic and corporate research organizations. *Bubble Trip* is the result of extensive prototyping and playtesting with the goal of creating an accurate and engaging assessment game. This game is designed to work simultaneously as an assessment tool and a playable game, allowing the player to choose which activity they wish to engage in without breaking the flow between either. In *Bubble Trip*, players control a fish in a single-screen marine environment that

can swim in all directions.

At the top of the game interface are a series of shells with iconography corresponding to a five point Likert scale questionnaire (strongly disagree, disagree, neutral, agree, strongly agree, see Figure 1). The HEXACO question text is displayed at the top of the screen with the Likert shell choices below. To answer a question, the player controls their fish to touch one of the shells, which opens the shell to reveal a sand dollar. This serves as a means of selecting the answer. To confirm the answer, the player must swim into the shell and collect this sand dollar. This served as confirmation of that answer. The interface was specifically designed to require two discrete actions by the player to complete a question in order to reduce accidental answers. The sand dollar was chosen as the inner collectable to be as neutral as possible in a game context versus a collectable that projects value such as a pearl, coin or gem. As previously mentioned, all design decisions were informed by a significant playtesting process with our target users.

Figure 1: *Bubble Trip* Gameplay. The counter in the upper left displays how many questions are remaining while the upper right counter shows how many bubbles have been collected.

Below this shell interface is a free area that the fish may freely swim in. Three conditions representing three different levels of game mechanics within *Bubble Trip* were presented to participants in this study. In the **full game condition**, bubbles spawn from the bottom of the screen randomly, floating towards the top. For every bubble that the player touches, they receive a point (reward). Jellyfish also periodically float horizontally across the screen. Touching a jellyfish (adversary) stuns the player momentarily to disrupt movement, but does not cause any reduction of score. Game goals were left to be as player-driven as possible in order to allow for different play-styles to emerge.

In the **environmental effects condition** there are no explicit rewards or adversaries, but the player may still engage with basic game mechanics. There are no jellyfish. Bubbles are present and can be collected, but the player does not receive points for collection. In

the **questions-only (control) condition**, neither bubbles nor jellyfish are present leaving the player to fully concentrate only on answering the questions, via the same fish controls.

During a game session, all actions initiated by the player are logged using a lightweight-logging framework, named Gloggr, written specifically for this project. Gloggr logs a number of variables including time durations for all stages of question answering (e.g., how long for a player to approach a shell, how long to select the answer, how long to confirm their answer), if a player changed their answer, as well as number of jellyfish collisions and bubbles collected. Gloggr also gave us the ability to log the 2D position of the player in the game space and to later create heat maps showing how different players utilized different areas of space within the game.

Participants

The data come from 212 mostly college-aged participants (55% female, 45% male) between the ages of 18 and 58 (85% of participants between 18 and 21 years) that completed both the traditional and game versions of the HEXACO. Most participants were white (53%), followed by Asian/Pacific Islander (21%) and Other/NR (15%). There were similar numbers of participants in each experimental condition (Table 1).

Table 1: Number of Participants

Condition	N	%
Full	70	33
Environmental	80	38
Control	62	29
Total	212	100

PROCEDURE

This study was completed entirely online and took no more than 30 minutes for a participant to complete. Links to the study website were distributed to students at four colleges. Participants first encountered an online consent form where they were informed of their rights should they engage in the study. Students could then opt to either, 1) take surveys and play the game with their data anonymously sent to researchers or 2) not take the surveys and play the game without their data sent to researchers.

Those participants that chose to participate in the study completed a demographics questionnaire and the HEXACO 60-item inventory through the online survey host, Qualtrics. Participants were then prompted to play the game and an instructions scene for *Bubble Trip* was displayed (see Figure 2). The website randomly assorted players into one of the three conditions; 1) the full game condition, 2) environmental effects condition, or 3) the questions-only condition.

Figure 2: The *Bubble Trip* instruction screen displayed to players in the study. Instruction cards for game elements were only displayed if those elements would be present in their selected condition.

After playing the game and answering the 60 HEXACO questions within it, participants concluded their time commitment in the study.

OBSERVATIONS AND RESULTS

Validity of game-based answers

Did examinees answer game questions seriously?

To investigate whether examinees may have been answering carelessly, we investigated how many times examinees provided the same response within each response category (e.g., did they answer "strongly agree" to everything), and also compared their average responses for items worded positively to items worded negatively that are reverse-scored: if they are not paying attention, they may answer "strongly agree" to two questions that contradict one another, such as:

> "When working on something, I don't pay much attention to small details."
> "People often call me a perfectionist."

Examinees were flagged if the difference between their forward-scored and reverse-scored items was extreme; if the absolute value of S*tudent's t* greater to or equal to 10, or if the standard deviation across items was less than 0.5, or if examinees answered more than 80% of items using the same response category.

Using these criteria, 17 students were flagged in the game, and 10 were flagged in the survey. There were no significant differences in the numbers of students flagged by experimental condition. These results indicate that overall, students appeared to take the survey questions seriously, in both the game and survey modes of administration.

Validity of Scores Obtained During Game Play

Table 2 contains mean scores for the online survey and each experimental condition. There were no significant differences in mean scores obtained by experimental condition or by administration mode.

Table 2: Mean Scores by Administration Mode and Condition

	Online Survey		Full Game		Environmental		Control	
Scale	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Honesty-Humility	3.3	0.6	3.4	0.6	3.3	0.6	3.3	0.7
Emotionality	3.3	0.6	3.4	0.6	3.2	0.7	3.4	0.6
Extraversion	3.2	0.7	3.2	0.7	3.2	0.7	3.2	0.7
Agreeableness	3.2	0.6	3.2	0.5	3.3	0.6	3.1	0.6
Conscientiousness	3.7	0.6	3.6	0.5	3.7	0.6	3.7	0.5
Openness	3.5	0.6	3.4	0.6	3.5	0.6	3.5	0.6

Cronbach's alpha is a measure of the internal consistency of a set of test items, and is often used as a measure of test reliability in the social sciences (Webb, Shavelson, & Haertel 2006). It ranges from 0 to 1, with higher values indicating higher consistency in responses to the items within a scale. Cronbach's alpha was reported on the HEXACO website for a sample of 1,126 college students , and for the purposes of this study is considered the gold standard for comparison. Table 3 shows the alphas reported from the website, compared to the survey version, and the game version, broken down by study condition. Cronbach's alphas obtained for both administration modes and all three game conditions were comparable to those obtained by the HEXACO researchers.

Table 3: Cronbach's Alphas for HEXACO Scales

Cronbach's Alpha	HEXACO Researchers	Online Survey	Full Game	Environmental	Control
Honesty-Humility	0.76	0.76	0.78	0.77	0.78
Emotionality	0.80	0.74	0.77	0.80	0.79
Extraversion	0.80	0.84	0.84	0.87	0.83
Agreeableness	0.77	0.74	0.67	0.79	0.78
Conscientiousness	0.76	0.77	0.78	0.82	0.75
Openness	0.78	0.75	0.75	0.76	0.79

Relationships between game-play behavior and the traditional assessment

Time Spent Playing the Game

On average, participants took 5.1 minutes to play the game (Table 4). There were small but non-significant differences in the time spent by experimental condition, such that

participants in the full experimental condition tended to take slightly more time than participants in the other two conditions.

Table 4: Game Play Time in Minutes

Condition	N	Mean	SD
Full Game	76	5.5	1.9
Environmental	86	5.1	1.6
Control	72	4.9	1.7
Total	234	5.1	1.7

Participants took significantly longer to answer the first items of the survey (Table 5), presumably because they were orienting themselves to the game mechanics and learning how to answer the items. The times leveled out after the first three items.

Table 5: Average Time in Seconds to Answer the First Items

	Full Game		Environmental		Control	
Item	Mean	SD	Mean	SD	Mean	SD
1	15.0	12.1	15.5	16.9	11.9	7.5
2	14.2	37.9	9.7	6.3	8.6	7.6
3	9.1	6.8	7.1	4.2	6.9	4.6
4	5.9	3.3	5.1	2.9	5.4	3.5
5	5.8	3.2	5.5	2.9	5.2	2.7

Because participants took more time answering the first few items of the survey, the amount of time taken to answer specific scales is confounded with the extra time taken to answer the first items. Therefore, the time spent answering each item was standardized to have a mean of 0 and standard deviation of 1 prior to analyses.

Across all three experimental conditions, time spent completing each scale was highly correlated, ranging from $r = 0.67$ to 0.90 for the full game condition, $r = 0.63$ to 0.89 for the environmental effects condition, and $r = 0.40$ to 0.87 for the control condition. In other words, some participants played quickly, some played more slowly, but individual participants seem to have played the game at a fairly consistent rate throughout the game.

Was time taken to answer items related to scale scores?

In general, the time spent answering items within each scale was unrelated to the scores obtained on that scale. However, several significant correlations emerged in the control condition. Table 6 contains the correlations between participants' HEXACO scores and the time taken to answer questions within scales where the p-values were less than 0.10, and correlations significant at $p < 0.05$ are bolded. All of the correlations were positive, meaning that participants with higher scores on a given scale tended to take more time to answer items on another scale. It is possible that a greater number of significant results

were seen in the Control condition because there were no distractions or obstacles that may have impacted the time taken to answer the items in the other two conditions.

It is interesting to note that most of the significant correlations found were related to the Emotionality, Openness, and Honesty-Humility scales, meaning that participants higher in these attributes tended to take longer to answer some of the items.

Table 6: Correlations between HEXACO Scores and Time to Answer Items Within Scales

Condition	HEXACO Score	Time to Complete Scale	Correlation	P-Value
Full Game	Extraversion	Emotionality	0.22	0.05
Full Game	Openness	Emotionality	0.20	0.09
Full Game	Extraversion	Openness	0.20	0.08
Full Game	Openness	Openness	0.21	0.07
Environmental	Emotionality	Extraversion	0.18	0.09
Environmental	**Emotionality**	**Openness**	**0.25**	**0.02**
Environmental	Emotionality	Total	0.18	0.09
Control	Openness	Honesty	0.21	0.07
Control	**Openness**	**Emotionality**	**0.25**	**0.03**
Control	**Honesty**	**Extraversion**	**0.24**	**0.04**
Control	**Openness**	**Extraversion**	**0.31**	**0.01**
Control	**Honesty**	**Conscientiousness**	**0.34**	**0.00**
Control	**Honesty**	**Openness**	**0.26**	**0.03**
Control	Honesty	Total	0.20	0.09
Control	**Openness**	**Total**	**0.26**	0.02

Consistency Between Survey Scores and Game Scores

Across all three conditions, participants were highly consistent in their responses to the online survey version of the assessment and the game version of the assessment. Intra-scale correlations ranged from 0.87 to 0.94 in the full game condition, 0.91 to 0.96 in the environmental condition, and 0.88 to 0.93 in the control condition. This is additional evidence that the scores obtained using the game version of the assessment are comparable to scores obtained using a traditional survey assessment.

Were HEXACO Scales Related to Game Play?

Numbers of Bubbles Collected

Participants collected significantly more bubbles, both overall and within each of the scales in the full game condition than in the environmental effects only condition (see Table 7).

Table 7: Average Numbers of Bubbles Collected

	Full Game		Environmental			
Scale	Mean	SD	Mean	SD	t-value	p-value
Honesty-Humility	10.2	5.2	6.5	4.4	4.93	< 0.0001
Emotionality	10.0	4.5	7.0	5.1	3.88	0.0002
Extraversion	10.0	4.9	6.9	4.7	4.10	< 0.0001
Agreeableness	10.9	5.7	6.4	4.6	5.39	< 0.0001
Conscientiousness	12.6	12.9	7.0	4.6	3.61	0.0005
Openness	10.0	5.3	7.5	5.6	2.97	0.004
Total	63.7	28.6	41.3	23.4	5.47	< 0.0001

In the full game condition, there were no significant relationships between numbers of bubbles collected and scores on the six HEXACO scales.

In the environmental effects condition, participants with higher Conscientiousness scores tended to collect fewer bubbles than students with lower Conscientiousness scores ($r = -0.24$, $p = 0.02$).

Jellyfish Collisions

Jellyfish were only present in the full game condition. Table 8 contains the average numbers of jellyfish collisions, by scale and across the entire game. Participants collided with significantly more jellyfish while answering questions related to Conscientiousness than when answering questions related to Emotionality ($t = 1.98$, $p < 0.05$) and Agreeableness ($t = 2.04$, $p < 0.05$).

Table 8: Average Numbers of Jellyfish Collisions

	Full Game	
Scale	Mean	SD
Honesty-Humility	1.4	1.6
Emotionality	1.2	1.4
Extraversion	1.2	1.4
Agreeableness	1.3	1.8
Conscientiousness	1.9	2.8
Openness	1.4	1.5
Total	8.3	7.0

In the full game condition, several significant relationships were found between numbers of jellyfish collisions and HEXACO scores. Participants with higher Extraversion scores tended to collide with greater numbers of jellyfish than those with lower Extraversion scores ($r = 0.23$, $p = 0.049$), while participants with lower Agreeableness scores tended to collide with greater numbers of jellyfish than those with higher Agreeableness scores ($r = -0.23$, $p = 0.049$).

Relationships Among Game Mechanics

Participants in the full game condition who tended to collect more bubbles during the game also tended to collide with more jellyfish during the game ($r = 0.33$, $p = 0.004$), and in both the full game and environmental conditions, bubble collecting was highly correlated with the amount of time taken to play the game ($r = 0.55$, $p < 0.0001$ for the full game condition, and $r = 0.61$, $p < 0.0001$ for the environmental effects condition). Jellyfish collisions were also highly correlated with the amount of time taken to play the game ($r = 0.70$, $p < 0.0001$). These results suggest that some participants may have been more engaged with the game, spending time collecting bubbles and swimming around, whereas other participants may have been more focused on completing the task quickly, avoiding interacting with the game mechanics.

Time taken to play the game was generally unrelated to participants' HEXACO scores; however, in the environmental effects condition, participants with higher Emotionality scores tended to spend more time playing the game than those with lower Emotionality scores ($r = 0.21$, $p = 0.049$). There was also a potentially interesting, albeit non-significant, relationship between Extraversion and total time spent playing the game in the full game condition ($r = 0.20$, $p = 0.08$).

Item Response Changes

Unlike a traditional paper and pencil survey, the game interface allows us to collect information about whether participants change their responses to items before making their final decision.

As shown in Table 9, participants in the full game condition were significantly more likely to make one or more changes to responses during the game than participants in the environmental and control conditions ($F = 3.22$, $p < 0.05$).

Table 9: Proportion of Examinees Making One or More Item Response Changes by Scale and Condition

	Full		Environmental		Control	
Scale	Mean	SD	Mean	SD	Mean	SD
Honesty-Humility	0.87	0.34	0.56	0.50	0.69	0.46
Emotionality	0.84	0.37	0.66	0.48	0.53	0.50
Extraversion	0.83	0.38	0.63	0.49	0.60	0.49
Agreeableness	0.86	0.35	0.62	0.49	0.58	0.50
Conscientiousness	0.93	0.25	0.65	0.48	0.65	0.48
Openness	0.83	0.38	0.64	0.48	0.53	0.50
Total	1.00	0.00	0.98	0.15	0.93	0.26

On average, participants in the full game condition made significantly more changes to items than did participants in the other two conditions ($F = 36.10$, $p < 0.0001$), and this same pattern of significance held across all six HEXACO scales (see Table 10).

Table 10: Average Number of Changes in Item Responses by Scale and Condition

	Full		Environmental		Control	
Scale	Mean	SD	Mean	SD	Mean	SD
Honesty-Humility	2.3	1.9	1.1	1.3	1.2	1.2
Emotionality	2.5	2.2	1.3	1.5	1.0	1.1
Extraversion	1.9	1.6	1.3	1.7	1.2	1.4
Agreeableness	2.8	2.2	1.2	1.5	1.0	1.2
Conscientiousness	3.0	2.5	1.2	1.3	1.4	1.6
Openness	2.7	2.4	1.3	1.4	0.9	1.1
Total	15.4	9.6	7.4	5.8	6.6	4.9

Relationships between item changes and HEXACO scores

Across the three experimental conditions combined, a significant relationship was found between Emotionality scores and the number of changes made to Conscientiousness items ($r = 0.15$, $p < 0.05$). In the full game condition, participants with higher Emotionality scores tended to make more changes to Openness items ($r = 0.23$, $p < 0.05$) and participants with higher Conscientiousness scores tended to make fewer changes to Conscientiousness items ($r = -0.26$, $p < 0.05$).

In the environmental effects condition, significant relationships were found between Honesty scores and changes to Emotionality items ($r = -0.29$, $p < 0.01$), Extraversion scores and changes to Conscientiousness items ($r = 0.23$, $p < 0.05$), Agreeableness scores and changes to Honesty items ($r = -0.23$, $p < 0.05$), and Agreeableness scores and changes to Conscientiousness items ($r = -0.24$, $p < 0.05$).

In the control condition, significant relationships were found between Agreeableness scores and changes to Emotionality items ($r = 0.23$, $p < 0.05$), Openness scores and changes to Agreeableness items ($r = -0.25$, $p < 0.05$), and Openness scores and changes to Conscientiousness items ($r = -0.25$, $p < 0.05$).

Relationships between item changes and Bubble Collection

Overall, the numbers of item changes were related to the number of bubbles collected ($r = 0.35$, $p < 0.0001$), and to the number of jellyfish collisions ($r = 0.38$, $p < 0.001$).

DISCUSSION

The hypothesis of this study was that a carefully designed game-based assessment could not only engage the examinees but also provide a valid assessment result. In this study, we transferred the HEXACO personality inventory into a game that is easily accessible using a web-browser. Compared with other personality assessment studies, within a short period of time, we successfully attracted a large number of users (over 200) and collected the data about not only the assessment item responses but also the user behavior when they interact with the game. The aforementioned hypothesis was investigated by using this rich data set. As this was an initial effort at examining the administration of a traditional assessment in a gamified environment, we took an exploratory approach to the

study instead of formally specifying hypotheses about how the different administration conditions would compare. This allowed us to broadly examine multiple instances of examinee behavior, the results of which can be used to inform future studies.

As we expected, the analysis showed that the more game elements added to the game, the more time a user played the game. There is a 12 percent increase in average playtime from the control condition to the full game condition. Additionally, the only differences between the environmental effects condition and the full game was the presence of jellyfish and a score counter; however, these features resulted in significantly greater numbers of bubbles collected. Although player engagement can be measured in a number of different ways, the time spent on the game and number of bubbles collected are arguably strong indicators of engagement. In this sense, we did observe that player's engagement increased significantly. Another promising finding was that the results of several statistical and psychometric analyses showed that this gamified assessment yields comparable scores to the original HEXACO assessment, suggesting that the validity of the assessment was not compromised by the new modality.

To our knowledge, this is the first study investigating the validity of a traditional assessment transferred into a virtual environment for the purposes of examining game mechanic effects on responses and engagement. However, some work has been done examining embedded questions into educational games in order to preserve feelings of presence, immersion, and flow (Frommel et al, 2015). While designing assessment games that match strongly with their traditional assessment "ground truth" counterparts is crucial, it is also very important to consider the appropriate design of a game like this. Many assessments of cognitive and non-cognitive variables can be tedious and long, resulting in many test-takers using inappropriate test-taking strategies. A properly instrumented game could help ameliorate this problem by providing a more engaging and immersive environment that motivates the test-taker, particularly in a "low stakes" situation. One of the most promising results from this research is that the "full game" condition yielded valid personality results for players, while also being the version that the players spent the most time with. The game elements of enemies and point collection did not appear to interfere with a player's ability to choose a response or affect their decision-making in answering the Likert-based scale. The increased time spent on the full-game version also produced an extra advantage of having more opportunity for players to demonstrate different behaviors and strategies that could then be analyzed for relationships with personality.

Besides the main hypothesis, the rich data collected by this gamified assessment helped us make some interesting discoveries. For example, in the control condition, a person with a high score in Honesty category is more likely to spend less time on questions in Conscientiousness category. While only preliminary, this finding suggests there is potential in pursuing further investigation into the relationship among the different categories in HEXACO and game behaviors.

CONCLUSIONS AND FUTURE WORK

The main focus of this paper was to establish the validity of results obtained from a gamified assessment, with some exploration of the relationships between personality characteristics and game elements. Game-based assessment is a hot topic, but with a few exceptions (e.g., Shute et al., 2015), most game producers have not provided validity evidence supporting the claims that their game is indeed measuring the intended construct(s). This study is one step in that direction. The results of this study show that an

assessment can be transformed into a game, and produce results comparable to a traditional assessment. Future studies validating the use of stealth assessment could use an embedded assessment as a ground truth on which to base validity claims.

Future work is needed to further understand the relationships between player characteristics and game play, including further exploration of player position in the game, which was beyond the scope of this paper. Also, because participants appeared to spend extra time during the first items orienting themselves to the game, it would be helpful if the game included a couple of warm-up questions so that the time spent answering items could be investigated in greater depth. Additionally, while the results of this study are compelling, this line of research should be replicated across different assessments, game themes, game mechanics, and populations, to determine the extent to which these findings are generalizable.

ACKNOWLEDGEMENTS

This work is funded through grants from ACT, Inc. and NSF# 0905127.

BIBLIOGRAPHY

Ashton, M.C., and K. Lee. (2009) "The Hexaco–60: A Short Measure of the Major Dimensions of Personality." *Journal of personality assessment* 91, no. 4: 340-45.

Basak, C., W.R. Boot, M.W. Voss, and A.F. Kramer. (2008) "Can Training in a Real-Time Strategy Video Game Attenuate Cognitive Decline in Older Adults?". *Psychology and Aging* 23, no. 4: 765-77.

Bethesda (2008) Fallout 3. [PC Computer] Maryland USA.

BioWare (2002) Neverwinter Nights. [PC Computer] Alberta Canada.

Blizzard (2014) World of Warcraft. [PC Computer] Irvine USA.

Drachen, A., C. Thurau, R. Sifa, and C. Bauckhage. "A Comparison of Methods for Player Clustering Via Behavioral Telemetry," in Proceedings of the 8th international conference on the Foundations of Digital Games (Greece, 2013), Society for the Advancement of Science of Digital Games.

EA (2011) Battlefield 3. [PC Computer, PS3, Xbox 360] Stockholm Sweden.

Finn, B. (2015) "Measuring Motivation in Low-Stakes Assessments." *ETS Research Report Series* 2015, no. 2: 1-17.

Frommel, J., K. Rogers, J. Brich, D. Besserer, L. Bradatsch, I. Ortinau, R. Schabenberger, *et al*. "Integrated Questionnaires: Maintaining Presence in Game Environments for Self-Reported Data Acquisition," in Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play (2015), ACM.

Goldberg, L.R. (1999) "A Broad-Bandwidth, Public Domain, Personality Inventory Measuring the Lower-Level Facets of Several Five-Factor Models." *Personality psychology in Europe* 7: 7-28.

Hoffman, H.G., D.R. Patterson, E. Seibel, M. Soltani, L. Jewett-Leahy, and S.R. Sharar. (2008) "Virtual Reality Pain Control During Burn Wound Debridement in the Hydrotank." *The Clinical Journal of Pain* 24, no. 4: 299-304.

Holtgraves, T. (2004) "Social Desirability and Self-Reports: Testing Models of Socially Desirable Responding." *Personality and Social Psychology Bulletin* 30, no. 2: 161-72.

Kaufman, J.C., J. Lee, J. Baer, and S. Lee. (2007) "Captions, Consistency, Creativity, and the Consensual Assessment Technique: New Evidence of Reliability." *Thinking Skills and Creativity* 2, no. 2: 96-106.

Lee, K., and M.C. Ashton. "Descriptive Statistics and Internal Consistency Reliabilities of the Hexaco-60 Scales in a College Student Sample." (accessed 1-26-2016). http://hexaco.org/downloads/descriptives_60.pdf.

Levy, L., R. Solomon, M. Gandy, J. Moore, J. Way, and R. Liu. "Actions Speak Louder Than Words: An Exploration of Game Play Behavior and Results from Traditional Assessments of Individual Differences," in Foundations of Digital Games (Pacific Grove, CA, 2015), ACM.

Matthews, G., I.J. Deary, and M.C. Whiteman.(2003) *Personality Traits*. Cambridge University Press, 2003.

Nouchi, R., Y. Taki, H. Takeuchi, H. Hashizume, T. Nozawa, T. Kambara, A. Sekiguchi, *et al*. (2013) "Brain Training Game Boosts Executive Functions, Working Memory and Processing Speed in the Young Adults: A Randomized Controlled Trial." *PloS One* 8, no. 2: e55518.

Paulhus, D.L. (1984) "Two-Component Models of Socially Desirable Responding." *Journal of Personality and Social Psychology* 46, no. 3: 598.

Pintrich, P.R., and E.V. De Groot. (1990) "Motivational and Self-Regulated Learning Components of Classroom Academic Performance." *Journal of educational psychology* 82, no. 1: 33.

Sarason, I.G. (1961) "Test Anxiety and the Intellectual Performance of College Students." *Journal of Educational Psychology* 52, no. 4: 201.

Shute, V.J. (2011) "Stealth Assessment in Computer-Based Games to Support Learning." *Computer Games and Instruction* 55, no. 2: 503-24.

Shute, V.J., M. Ventura, M. Bauer, and D. Zapata-Rivera. "Melding the Power of Serious Games and Embedded Assessment to Monitor and Foster Learning." In *Serious Games: Mechanisms and Effects*, 295-321: Routledge/LEA Philadelphia, PA, 2009.

Shute, V.J., M. Ventura, and F. Ke. (2015) "The Power of Play: The Effects of Portal 2 and Lumosity on Cognitive and Noncognitive Skills." *Computers & Education* 80: 58-67.

Spronck, P., I. Balemans, and G. van Lankveld. "Player Profiling with Fallout 3," in Eighth Annual Intelligence and Interactive Digital Entertainment Conference (2012),

Tekofsky, S., Spronck, P., Plaat, A., van den Herik, J., Broersen, J. "Personality Assessment through Gaming Behavior," in Foundations of Digital Games (Greece, 2013),

van Lankveld, G., P. Spronck, J. Van den Herik, and A. Arntz. "Games as Personality Profiling Tools," in Computational Intelligence and Games (CIG 2011), IEEE.

Webb, N.M., R.J. Shavelson, and E.H. Haertel. (2006) "Reliability Coefficients and Generalizability Theory." *Handbook of statistics* 26, no. 4: 81-124.

Wiggins, J.S.(1996) *The Five-Factor Model of Personality: Theoretical Perspectives*. Guilford Press, 1996.

Wise, S.L., and C.E. DeMars. (2005) "Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions." *Educational Assessment* 10, no. 1: 1-17.

Wise, S.L., and X. Kong. (2005) "Response Time Effort: A New Measure of Examinee Motivation in Computer-Based Tests." *Applied Measurement in Education* 18, no. 2: 163-83.

Zatz, S., and L. Chassin. (1985) "Cognitions of Test-Anxious Children under Naturalistic Test-Taking Conditions." *Journal of Consulting and Clinical Psychology* 53, no. 3: 393.