9-2016

Gamification of Individual Differences

Carolyn A. Fiore
Industrial Organizational Psychology, cakardong@gmail.com

Gamification of Individual Differences Inventories

by

Carolyn Fiore

A Thesis

Submitted to the Graduate Faculty of

St. Cloud State University

in Partial Fulfillment of the Requirements

for the Degree of

Master of Science

in Industrial Organizational Psychology

September, 2016

Thesis Committee:
Dr. Jody Illies, Chairperson
Dr. Daren Protolipac
Dr. Edward Ward

Abstract

This study examines the new method of gamification for measuring individual differences for personal decision making. The first purpose was to examine if gamification, compared to generally accepted self-report inventories, has convergent and differential evidence that will support gamification's assessment ability. Results did not find support for gamification's ability to measure individual differences. Participants' ability to fake was also measured to see if gamification is less susceptible to this problem that plagues self-report methods. Support was found for this hypothesis, which potentially provides a bright outlook for gamification. Casting a shadow on this hope though, participants' perceptions of face validity was also measured and found that gamification was rated as appearing less valid. Last but not least, how enjoyable participants found gamification and self-report measures was compared and it was found that gamification was more enjoyable.

Table of Contents

Page

List of Tables 4

Gamification of Individual Differences Inventories 5

Assessment of Individual Differences 5

Limitations of Self-Report Inventories 7

Alternatives to Self-Report Inventories 9

Gamifications 13

Method 17

Participants 17

Gamification Measures 18

Self-report Measures 19

Procedure 23

Data Analyses 25

Results 26

Discussion 32

Limitations 35

Conclusion 37

References 38

Appendix A 44

Appendix B 47

Appendix C 49

List of Tables

Table Page

1. Means, Standard Deviations, and Correlations for Gamification and Self-report Constructs 28

Gamification of Individual Differences Inventories

In the early 2000s, employers began looking at using social media profiles, such as Facebook, when considering applicants for hiring. A survey in 2009 asked over 2,600 employers if they explore social media sites to obtain a better understanding of their applicants; 45% of them replied that they do (Brown & Vaughn, 2011). In 2013, the odds of a prospective employer looking at your social media profiles went up according to a survey by Jobvite, which reported that "93% of recruiters said they were likely to look at the social media profiles of applicants, and 43% have reconsidered a candidate (both in the negative and positive direction) based on the candidates' social media profile" (Drouin, O'Connor, Schmidt, & Miller, 2015, p. 1-2). Some studies also suggest that using social media in employee selection might even be useful at predicting performance (e.g., Kluemper & Rosen, 2009) and assessing personality (e.g., Drouin et al., 2015).

Assessment of Individual Differences

Although Brown and Vaughn (2011) and Drouin et al. (2015) suggested that social media is becoming a popular tool for selection, there are many more options for assessing individual differences that have been around longer. Some individual assessments that most job applicants have experienced would be recommendations, references, and interviews (Cascio & Aguinis, 2011). Wilk and Cappelli's (2003) findings support this assertion, noting that employers report that they regularly ask for references and almost always interview their applicants. The average validities for several of these methods, though, are mediocre at best, ranging from .14 and .26 (Cascio & Aguinis, 2011). Interviews, however, have been found to have considerably better validity with corrected validities of .38 for unstructured interviews and .51 for structured interviews (Cascio & Aguinis, 2011; Schmidt & Hunter, 1998). McDaniel, Whetzel, Schmidt,

and Mauer (1994) also found that when it comes to predicting job performance criteria, the corrected validity scores for structured interviews (.44) are better than unstructured (.33).

In addition to the previously mentioned methods, tests and inventories can also be used to assess individual differences. Cognitive ability tests, for example, have been found to be powerful predictors of performance in applicants with a corrected validity of .51 (Cascio & Aguinis, 2011; Schmidt & Hunter, 1998). The problem with cognitive ability tests, though, is that they are also linked with adverse impact (Cascio & Aguinis, 2011). Validated tests with adverse impact have utility, however, and are legal to use because they are accurately predicting performance (Cascio & Aguinis, 2011), but it would still stand to reason that it might be a good idea to take additional measures to address this adverse impact problem. There are a few options for mitigating this problem, one of which is to supplement cognitive ability tests with other measures (Cascio & Aguinis, 2011). One common supplement is using personality or other individual difference inventories as predictors in addition to cognitive ability tests (Cascio & Aguinis, 2011). Many of these inventories do not have the same adverse impact problems as cognitive ability (e.g., Mount & Barrick, 1995). Schmidt and Hunter (1998) found that conscientiousness tests alone were found to add .09 validity above and beyond general mental ability tests let alone what an entire personality test might be able to add in the form of incremental validity.

Individual difference inventories, particularly personality inventories, have also been argued to be able to predict criteria such as attitudes, behaviors, performance, and leadership (Ones, Dilchert, Viswesvaran, & Judge, 2007). Among the Big Five dimensions, research tends to support the trait of conscientiousness extensively. Mount and Barrick (1995) found that conscientiousness was a valid predictor of job and training proficiency in all occupation groups.

Hurtz and Donovan (2000) found support for conscientiousness as well, finding that it had the highest validity when predicting job performance among the Big Five dimensions as well as the highest estimated true validity across all four of the occupation categories they researched. Extraversion and openness to experience have also been found to be valid predictors of training proficiency across occupations (Mount & Barrick, 1995). Schmidt and Hunter (1998) even found that personality can predict above and beyond cognitive ability. It should be noted that measures of cognitive ability are referred to as tests, which indicates that there are right and wrong answers, while personality measures are called inventories because they measure preferences where there are technically no right or wrong answers (Cascio & Aguinis, 2011).

Limitations of Self-Report Inventories

When responding to a self-report inventory in an evaluation situation, such as selection, individuals know that evaluators are looking for certain results. This may lead responders to ask themselves if they should be honest in their answers or fake their answers to match the characteristics they think the assessor is evaluating. There is a great deal of research on whether or not job applicants are able to fake inventory responses as well as whether or not they actually make the attempt to do so. The literature seems to have come to the general conclusion that applicants can fake on personality inventories (Cascio & Aguinis, 2011; Ones et al., 2007; Rees & Metcalfe, 2003). McFarland and Ryan (2000) mentioned that researchers have found that participants can increase their scores on non-cognitive measures "by as much as one standard deviation through faking" (p. 813). Whether or not applicants actually fake their answers, though, is still unclear along with if this distortion decreases the validity of the inventory. Ones et al. (2007) found that even though participants demonstrated that they could fake on a personality inventory, they showed a general tendency not to do so in real employment

situations. However, in the real employment situation, some participants still did attempt to fake their answers, so although Ones et al. (2007) suggested faking is not as big of a problem as others make it out to be, it does exist and needs to be considered.

Faking is potentially problematic for many reasons. Mueller-Hanson, Heggestad, and Thornton (2003) along with McFarland and Ryan (2000) discussed how faking can decrease criterion-related validity. The criterion-related validity is decreased in these cases because the results from the personality inventory would not be accurately depicting the applicant's personality. Mueller-Hanson et al. (2003) also suggested that criterion-related validity might be misleading when socially desirable measures are used in an attempt to correct faking. It was suggested that this decreases criterion-related validity because social desirability scales do not tend to correlate well with each other; therefore, they lack convergent validity. In other words, when using these measures, it is unclear exactly what is being corrected and if that correction is appropriate.

Many studies also have found that faking can have vast effects on selection decisions, and that these effects are made even worse by the inconsistency of faking across people (Mueller-Hanson et al., 2003). If everyone were to fake equally, it would not necessarily matter because everyone would still maintain the same rank order. However, if people are inconsistent in their faking, it will cause the order of preferred candidates to be shuffled (McFarland & Ryan, 2000; Muller-Hanson et al., 2003). This can also negatively impact criterion-related validity because the personality scores will not be able to predict performance as accurately. The best performer might no longer have the best personality score due to a lower performer faking on his/her inventory, resulting in that person obtaining a seemingly better score (McFarland & Ryan, 2000).

McFarland and Ryan (2000; 2006) have looked at what predicts faking in applicants and found that people's beliefs in regard to faking have an influence on their intentions to fake. This, in turn, predicts if they will fake or not. However, the ability to fake moderates the relationship between faking intention and faking success. Therefore, regardless of how much a person wants to fake their results, if they do not know how to fake or do not have the ability to fake, they will not be successful in their faking attempt (McFarland & Ryan, 2006). Attitudes toward faking (such as if they feel faking is acceptable or not), subjective norms towards faking (such as if they feel it is socially acceptable to fake or not), and perceived behavioral control toward faking (whether or not they think they are able to fake or not) were also significantly related to the intention to fake (McFarland & Ryan, 2006). In fact, those three variables were found to account for 45% to 57% of the variance in intention to fake (McFarland & Ryan, 2006). Telling participants that a measure included a social desirability scale intended to measure faking behaviors, though, was found to lower both the intention to fake and faking behaviors; however, beliefs about faking were not altered (McFarland & Ryan, 2006). Although the warning method did mitigate the faking problem to an extent, there are other, potentially more effective ways of dealing with this problem.

Alternatives to Self-Report Inventories

The faking problems related to personality inventories stem from the self-report nature of these assessments (Cascio & Aguinis, 2011). As a result of this, one solution to the faking problem is to consider alternatives to the self-report method. In 1963, forced-choice personality scales were developed, but it was not until the early 2000s that they were considered for use during selection procedures (Goffin, Jang, & Skinner, 2011). Forced-choice response scales are often considered a good way to measure personality while avoiding the faking problem that

plagues validity (Cascio & Aguinis, 2011; Goffin et al., 2011). Forced-choice personality scales give the applicant two statements that are considered to be equally desirable, and the applicant is forced to choose the one statement that is most fitting for themselves (Goffin et al., 2011). Assuming both statements are equally desirable, the applicant should be honest in their selection, therefore reducing or eliminating faking, as has been found by various studies (Goffin et al., 2011). On the negative side, this response option has been known to lead to negative reactions on the part of the applicants as they dislike being forced to choose among only two options (Cascio & Aguinis, 2011). This dislike is associated with a variety of other problems that will be discussed later.

A more recently developed method for measuring personality while avoiding faking is called conditional-reasoning (Berry, Sackett, & Wiemann, 2007; Cascio & Aguinis, 2011; James et al., 2005). In this method, the focus is on how people with different personality traits will use different justification systems to explain behaviors instead of focusing on the behaviors themselves (Berry et al., 2007 ; Cascio & Aguinis, 2011; James et al., 2005). To make faking even more difficult, the measurement of justification systems can be hidden behind what seems to be a question measuring inductive reasoning (Berry et al., 2007; DeSimone & James, 2015; James et al., 2005). Therefore, conditional-reasoning lacks obvious face-validity to those taking the inventory and, as a result, this circumvents the faking problem (DeSimone & James, 2015). An example conditional reasoning question that James et al. (2005) provided is:

> American cars have gotten better in the past 15 years. American carmakers started to build better cars when they began to lose business to the Japanese. Many American buyers thought that foreign cars were better made. Which of the following is the most logical conclusion based on the above?
> a. America was the world's largest producer of airplanes 15 years ago.
> b. Swedish carmakers lost business to America 15 years ago.

c. The Japanese knew more than Americans about building good cars 15 years ago.
d. American carmakers built cars to wear out 15 years ago so they could make a lot of money selling parts. (p.76)

The first two answers are essentially filler responses not meant to be picked, but the second two answers are actually measuring personality, not logical reasoning as the instructions would suggest (James et al., 2005). Answer d would suggest a hostile personality because it gets at the underlying belief that powerful people will victimize others when it benefits them while answer c would not (James et al., 2005).

DeSimone and James (2015) tested and found support that conditional reasoning does indeed work as it was intended, so people with latent personality traits will select the logical response option that addresses that trait. Studies have found that conditional reasoning inventories are acceptable by psychometric standards, and based on 11 studies, has an average uncorrected validity of .44 when it comes to measuring aggression and as high as .52 when measuring academic achievement (Berry et al., 2007; James et al., 2005). On the other hand, Berry, Sackett, and Tobares (2010) ran a meta-analysis on conditional reasoning tests of aggression and only found criterion-related validity of .16. There appears to be no conclusive information yet on the predictive validity of conditional reasoning but, Berry et al. did report mean correlations of .14 and .21 for conditional reasoning test of aggression predicting job performance.

Studies have also supported the idea that conditional reasoning is not easy to fake (Bowler & Bowler, 2014; Bowler, Bowler, & Cope, 2013). Although it was noted that this is only true when indirect measurement is sustained given that "when the construct of interest was made explicit, participants could identify the keyed response options when instructed to do so"

(Bowler & Bowler, 2014, p. 415). If conditional reasoning instruments become more commonly used, this could become an issue as people might become informed on the method and then possibly be able to fake these tests as well. There is also the problem that developing conditional reasoning measure can be quite difficult and time consuming (Cascio & Aguinis, 2011).

A variation of conditional reasoning that is also being considered is the Differential Framing Test (DFT) (Berry et al., 2007). This test appears to be a synonyms test on the surface in which the applicant is given a word and two options to pick from as synonyms for the word. The trick is that both words are synonyms but the word that is selected will reflect a person's personality as a result of the connotations attached to the word (Berry et al., 2007). For example, the original word might be "critique" and people need to pick if the synonym is "criticize" or "evaluate" (Berry et al., 2007). Technically, both words are correct but, similar to conditional reasoning, the expectation is that a person's underlying personality will influence the word they will select. In this example, someone that selects "criticize" would be expected to be more aggressive than someone who opts for "evaluate" (Berry et al. 2007). Initial validity tests for the DFT show promise. For example, Berry et al. (2007) reported that when predicting conduct violations in an academic setting, DFT resulted in cross-validities in the .30-.50 range across two samples (Berry et al., 2007, p.286). Other studies have shown DFT has acceptable internal consistency and test-retest reliability but has low correlation with tests measuring similar constructs using different methods, such as conditional reasoning (Berry et al., 2007). In the end, the jury is still out on the DFT.

The development of conditional reasoning inventories and DFTs has provided evidence that there are potentially equally or even more effective methods of assessing individual differences than the traditional, direct self-report inventory. In fact, there are many more options

than just self-report measures of personality that currently exist. In the end though, the search for the "golden measure" continues as problems of applicant reaction, development time, and lack of validity evidence plague these options. Therefore, psychologists still need to look to the future for other possible solutions.

Gamification

As it turns out, the future might be closer than originally thought. Gamification is an up and coming method for measuring individual differences that appears to offer the potential for accurate assessment while decreasing or eliminating the disadvantages associated with self-report assessment. Gamification is when the mechanics of playing games are used in business applications (Herzig, Strahringer, & Ameling, 2012). Game mechanics are defined as methods used by people for interacting with games (Sicart, 2008). There are even mobile application software (app) games where personality and other individual difference constructs are assessed (Computer Basics, n.d., Noguchi, 2015). In these cases, personality and individual differences are measured based on the performance on specific tasks completed on a mobile device (Noguchi, 2015).

Gamification methods have also been found to improve variables such as enjoyment, flow (a mental state in which a person is fully focused and energized by their task), and perceived ease of use (Herzig et al., 2012). Herzig et al. (2012) looked at how gamification could influence behavioral intentions to use a particular software system. The results showed that enjoyment was a strong antecedent of perceived ease of use while flow was a weak antecedent (Herzig et al., 2012).

First things first though, for starters it is important to look at if gamification is even capable of sufficiently measuring individual differences such as personality. Considering how

new gamification is, it is not surprising that research on its measurement abilities is lacking. As a result, this study will compare gamification to standard self-report assessments to see if it is a comparable option for measuring individual differences. It is expected that gamification measures and self-report measures of similar constructs will show convergent evidence (Cascio & Aguinis, 2011). On the other hand, gamification measures and self-report measures of theoretically unrelated constructs should show discriminate evidence (Cascio & Aguinis, 2011). By demonstrating that gamification measures share appropriate convergent and discriminate evidence to already accepted measures, it is reasonable to consider that gamification holds potential as an individual difference measurement method. The threshold for being considered as appropriate levels of convergent and discriminate evidence are correlations of .70 or higher and .50 or lower, respectively, as is recommended by Carlson and Herdman (2012).

H1: Results will show appropriate convergent and discriminant validity evidence reflecting that gamification has the ability to assess individual differences to the same extent as self-report personality measures.

It has also been suggested that gamification can reduce applicants' ability to fake on the measurement (Armstrong, Landers, & Collmus, 2016). Noguchi (2015) argued that gamification will reduce the ability to fake as a result of an applicant's inability to tell what the games are measuring. Therefore, if test takers do not know what is being measured, they will not be able to fake their responses to try and to make their scores better. One company that uses gamification reported that it is much harder to fake a game because "playing a game involves thousands of decisions and actions;" in addition, they noted that "sophisticated data-mining algorithms look at many different characteristics of a person's game play, some of which people are not even aware of" (" New to Knack?" 2015, What if someone "games" the game?).

It has also been argued that games might be able to evoke behaviors reflective of various constructs better than normal questionnaires, therefore making these constructs easier to measure (Armstrong et al., 2016). Without obvious links between game behaviors and personality results, it is impossible for players to fake; they just have to try their best to "win" the game to the best of their abilities. Even if they do try to fake, they probably have no idea if their efforts are actually helping or hurting the achievement of a desired score. Unfortunately, because gamification is such a new field, there is a lack of scientific research on the ability of people to fake their scores. To this end, the present study will explore faking in a gamification assessment as related to a self-report assessment. For the purpose of this study, faking will be defined as inflating responses on constructs deemed desirable by subject matter experts while decreasing responses on constructs deemed undesirable.

H2: Gamification will have less faking than self-report personality measures.

Although not understanding how the games are measuring personality might help reduce faking, it also might reduce face validity. This could be problematic for companies because participants' ratings of face validity are linked with a variety of good and bad outcomes. Some of the outcomes of face validity found by Shotland, Alliger, and Sales (1998) include level of comfort administering a test, motivation to perform on the test, attractiveness of an organization, favorability of selection procedures, chance of the selection procedures being challenged in court, and perceptions of fairness, ethicalness, and morals. Hausknecht, Day, and Thomas (2004) also reported that face validity could be related to legal challenges, test performance, and company perceptions. They also discussed that face validity may be related to an applicant's likelihood of accepting a job offer, willingness to recommend an employer to others, and perception of procedural and distributive justice (Hauckneche et al., 2004).

When considering the relationship between face validity and motivation, it has been suggested that when an applicant can understand that an assessment is related to the job, they have no reason not to perform their best, but if they cannot see this relationship, they may feel that they are being put through unnecessary stress, which might lead to feeling demoralized and demotivated (Shotland et al., 1998). Using face valid tests might also facilitate feelings of fairness and trust in an organization, which should increase the likelihood of accepting job offers, whereas it would seem odd for a person to want to work for a company they do not feel is fair and that they cannot trust (Shotland et al., 1998).

If low face validity is linked to an applicant's likelihood of not accepting a job, it could cause a company to lose top applicants to competitors (Hausknecht et al., 2004). Armstrong et al. (2016) added that even if applicants accept a job offer, there could eventually be decreased job satisfaction and performance as well as increased turnover due to attitudes resulting from a lack of face validity of selection predictors. These are supposedly a result of decreased self-efficacy, self-esteem, and organizational attractiveness (Armstrong et al., 2016). Applicants that were hired may not believe they got the job because they were the most qualified because they do not understand the purpose of the selection assessment, which could lower their self-efficacy and self-esteem.

Many individual difference assessments, particularly in the area of personality, are not rated well on face validity (Steiner & Gilliland, 1996). Hoang, Truxillo, Erdogan, and Bauer (2012) found that among participants from the United States and Vietnam, personality tests and honesty tests were perceived to be in the bottom half of a list of ten selection methods for process favorability. This has been found despite the fact that personality inventories are believed to be made up of questions that are susceptible to faking, indicating that one can decipher the intent of

the measurement. Considering that gamification is thought to result in less response distortion due to an inability of an individual to understand what is being measured, it stands to reason that these assessments would result in less face validity than standard individual different inventories.

H3: Participants will rate gamification measures as less face valid than self-report measures.

As mentioned previously, it has also been suggested that gamification should increase peoples' enjoyment over mundane self-report inventories (Attali & Arieli-Attali, 2015; Herzig et al., 2012). Herzig et al. (2012) found that enjoyment through gamification increased behavior intentions through perceived ease of use. Attali and Arieli-Attali (2015) found greater likeability ratings under some testing conditions with middle school participants where points were awarded for accuracy and speed during testing compared to simply saying if the responses were correct or incorrect. Hamari, Koivisot, and Sarsa (2014) completed a literature review of gamification articles and claimed that "gamification does work, but some caveats exist" (p. 5) in relation to producing positive psychological and behavioral outcomes such as motivation for example. In general, however, there is a lack of published research on the enjoyment of gamification methods as compared to traditional assessment methods. Therefore, this study will also explore this aspect.

H4: Enjoyment ratings will be higher for the gamification measures than the self-report measures.

Method

Participants

A convenience sample was used which included 35 undergraduate students at a Midwestern university who received course extra-credit for completing the study. Of the 35

participants, 17 were in the faking condition and 18 were in the non-faking condition. The majority of participants selected the age bracket of 18-25, while four participants were in the 26-35 bracket, and one participant selected the 36-45 age bracket. Over three quarters of the participants were females with a total of 27 females and 8 males. As for average hours playing app games, 14 participants reported that they do not play app games, 12 reported 1 to 3 hours per week, 8 reported 4-6 hours per week, and 1 reported 7 to 10 hours per week. None of the participants claimed to have played the app games using in this study, so all participants were new to the game and came in with no prior experience.

Gamification Measures

Individual differences measured by gamification were obtained using a set of app games that measure personality based off of the way people play app games (the identity of the company who developed this gamification assessment has been intentionally ommitted). The gamification assessment measured 37 individual difference traits, including personality, interests, motivation, reasoning, and problem-solving preferences. Considering that assessment is marketed by a private, for profit business, extensive information about how these constructs are assessed is not provided. It is stated that algorithms are used to measure these constructs based on how people play the app games. It is noted that data-mining methodology is also utilized to to aid in the assessment results.

Originally, players were provided with all of the scores on the constructs measured while playing the games. Results were provided on a one to five star system. Shortly before this study began, this changed and players were only providing with information on the constructs on which they scored five stars. A request for full result information was denied, leading to a decision to score results on a dichotomous scale That is, if the participant got a five-star score on

the construct, they were rated as high on that construct, but if they did not receive a score on the construct, they were rated as being low on the construct because they were below the five-star rating.

Definitions for the constructs were developed based off of high and low end scale descriptions obtained after playing the games to help players interpret their results. These definitions were then matched with similar construct definitions from self-report measures. The comparable construct measures were combined to form a self-report personality test that measured several of the same personality characteristics as provided by the gamification measure. Appendix A provides a table of the gamification constructs with their definitions and the comparable self-report construct measures that were used in this study

Three games were included in the gamification assessment. One was a puzzle style game where the goal is to make a path so that two robots can connect. In a second game, the player controls a waiter who is tasked with reading customers' emotions, placing orders for food based off the customers' emotions, serving the food, clearing dirty dishes, and washing those dishes. Finally, a third game tasks players with throwing water balloons at fire demons who are trying to destroy the water balloon machine, all while also trying to save the flowers between the fire demons and the machine.

Self-report Measures

General Self-Efficacy Scale (GSE). The GSE is an eight-item scale that can be used to measure people's level of general self-efficacy (Chen, Gully, & Doy, 2001). The scale uses a five-point Likert scale system that goes from "disagree strongly" to "agree strongly" (Chen, Gully, & Doy, 2001). A sample question from this scale is "I will be able to achieve most of the goals that I have set for myself" (Chen, Gully, & Doy, 2001). Chen et al. (2001) defined self-

efficacy as "beliefs in ones capabilities to mobilize the motivation, cognitive resources, and courses of action needed to meet given situational demands" (p. 62). A test of content validity by Chen et al. (2001) provided support for both "discriminant and content validity of the GSE and self-esteem measures" (p. 69), and internal consistency was found to be.86 for the GSE and test-retest reliability was adequate (r = .67). In this study, the internal consistency of the GSE was .79.

Big Five Inventory (BFI)-short version. The short version of the BFI was used excluding the neuroticism factor. With alpha reliabilities generally ranging from .75 to .90 and three month test-retest reliabilities ranging from .80 to .90, the BFI is a widely accepted personality measure (John & Srivastava, 1999). In this study, the lowest internal consistency for the BFI was the conscientiousness factor with an alpha of .78. For extraversion, agreeableness, neuroticism, and openness this study found internal consistencies of .91, .78, .84, and .79 respectively. Considering the length of this study, time was a concern so that participants do not experience boredom and fatigue. Therefore, the short version of the BFI was used which is composed of 44 items. Most of the neuroticism items were removed as they were not compared to a gamification construct, so only 39 items were administered, the few neuroticism items that were left in were used to test faking. This inventory uses a five-point response scale that ranges from "disagree strongly" to "agree strongly" (John & Srivastava, 1999). All of the items begin with "I see myself as someone who…" and one example item is "is talkative" (John & Srivastava, 1999).

International Personality Item Pool (IPIP). Three scales from the IPIP were used, achievement striving, artistic interests, and imagination. Each scale comprises ten questions (Goldberg, 1999). Goldberg (1999) found the average alpha coefficient was .80 for the IPIP

across all scales. For this study, the internal consistencies were even higher with alphas of .83 for both achievement striving and artistic interests and .88 for imagination. The IPIP was chosen for its relatively short scales as well as its general acceptance as a self-report personality measure. All three scales from the IPIP use a five-point response scale system that ranges from "very inaccurate" to "very accurate" (Goldberg, 1999). This inventory asks participants to rate how accurate each statement is to them; an example statement is "plunge into tasks with all my heart" which is used to measure achievement striving (Goldberg, 1999).

Life Orientation Test-Revised (LOT-R). The LOT-R is a measure of dispositional optimism. Questions directly ask people if they expect good or bad outcomes out of their lives (Carver, Scheir, & Segerstrom , 2010; Scheier, Carver, & Bridges, 1994). This measure is comprised of ten items, four of which are filler questions. The LOT-R uses a five-point response scale ranging from "I agree a lot" to "I disagree a lot," which is used for items such as "In uncertain times, I usually expect the best" (Scheier et al., 1994). The LOT-R has been shown to have fairly high internal consistency ($\alpha = .78$) and test-retest reliability ($r = .79$ after 28 months) (Scheier et al., 1994). In this study an alpha of .78 was found.

Sensation Seeking Scale. The Sensation Seeking Scale measures the extent to which a person actively searches for experiences and feelings of pleasure and excitement. It measures this through 14, forced-choice options (Zuckerman, 1979). The alpha for the Sensation Seeking Scale with all the items was .73, but item one showed a negative corrected item-total correlation. After reviewing the problem item it was determined that the item could be interpreted so that both response options could be selected by someone with high sensation seeking. Once this item was removed the alpha jumped to .77.

Self-Monitoring Scale. A person's ability to regulate their behaviors to be socially appropriate is measured by the Self-Monitoring Scale. The Self-Monitoring Scale uses 25 true-false questions that assess participants' personal reactions to different situations to get a measure of self-monitoring ability (Snyder, 1974). Snyder (1974) found a Kuder-Richardson 20 reliability of .70 as well as a test-retest reliability of .83 for this scale. This study was not so fortunate though, and only found an alpha of .59 when using the full scale. Three items were found to have negative corrected item-total correlations: (a) "When I am uncertain how to act in social situations, I look to the behavior of others for cues," (b) "I rarely need the advice of my friends to choose movies, books, or music," and (c) "I sometimes appear to others to be experiencing deeper emotions than I actually am." It was determined that these items could be confusing to participants and not accurately measuring their self-monitoring abilities and therefore were removed. By removing the items with negative item-total correlations in this study, the alpha improved to .66.

Social Intelligence Scale. Social intelligence is conventionally defined as the combination of both social perceptiveness and behavior flexibility (Marlowe, 1986). This biodata scale consists of 30 questions that deal with life events, assessing if these events have happened to a participant, and if so, how often. Of the 30 questions, 15 measure behavioral flexibility while the other 15 measure social perceptiveness. This measure was developed by CPS Human Resource Services. This scale uses six different five-point response scales such as "never" to "very often" and "not at all likely" to "extremely likely." The overall scale has been found to have an internal consistency of .74 while the subscales of behavioral flexibility and social perceptiveness had reliabilities of .61 and .57 respectively (Illies, Basarich, Young Illies, & Reiter-Palmon, 2007). For the present study, the overall scale was used for all analyses and

was found to have an alpha of .62. After removing five items with negative item-total correlations, the alpha became .73. The scale with the removed items was used for hypothesis testing.

Adolescent Leadership Measure. The Adolescent Leadership Measure was used to assess leader emergence in college-aged individuals (Mumford, O'Conner, Clifton, Connelly, & Zaccaro, 1993). It is a biographical data scale that measures the frequency of behaviors or activities that are consistent with leadership development using 19 items with a variety of five-point response options such as "very often" to "never", "very likely" to "not at all likely", "very effective" to "never effective", "never" to "six or more times", and "great extent" to "not at all". A sample item from this measure is "how often have you guided or directed other in group activities." Mumford et al. (1993) found alpha coefficients of .81 for men and .83 for women, showing strong internal consistency. This study found an alpha of .88 for the Adolescent Leadership Measure.

Procedure

It was estimated that the study would take approximately an hour and a half for participants to complete but the average time it took most participants was around an hour and fifteen minutes. Upon entering the study, participants were asked to read and sign an informed consent sheet that told them they would be completing self-report questionnaires and playing app games. After providing informed consent, the participants were asked to fill out a four-question demographics survey that asked for their age, their gender, the approximate hours per week they spent playing app games, and their experience playing the app game using in this study. The order in which the participants played the app games or completed the questionnaire was counterbalanced to avoid order effects. The order that the three app games were played was also

rotated. Player accounts were predeveloped before the participants' arrival to save on time and to maintain confidentiality. This also allowed the researcher to have access to the accounts to obtain the necessary data. A Samsung Galaxy 4 phone using an Android operating system was provided to students that already had all of the app games downloaded on it. This procedure was used for multiple reasons. First, this saved on time in that each participant did not need to download the games. Second, this prevented participants from having to use their own phone memory space or phone data allocation to play the games. Finally, and most importantly, this allowed the researcher to control for variability due to type of phone, which can affect the app version utilized. After finishing each assessment method, participants were asked to fill out a short, three-question survey about the method they just completed. This survey asked one question about how enjoyable the method was and two questions about the perceived face validity. The two survey items on face validity were based off questions used by Steiner and Gilliland (1996). The entire survey can be found in Appendix B.

The faking manipulation occurred before participants began each method of measuring individual differences. Half of the participants were read a script asking them to act as though they were taking the questionnaire or playing the games for their own self-learning. The other half were asked to respond or play as though the results would be used to decide whether or not they would be offered a high paying position. All of the different rotations as well as the faking condition were coded. Faking scripts are provided in Appendix C. The use of more blatant faking scripts in which participants would be directly asked to fake were considered but a more realistic idea of faking in an application setting was desired. McFarland and Ryan (2002) also found that scripts asking participants to act as though they were an applicant showed similar faking results found in a selection context.

As previously mentioned, a demographics survey was used to collect data on several possible control variables, including gender and age. Research is mixed on the influence of these factors. A report by Entertainment Software Association in 2014 found that gender and age were fairly even distributed among gamers, with males accounting for 52% of game players. The age distribution for game players under the age of 18, between 18-34, and over 36 was 29%, 32%, and 39%, respectively. The developer of the gamification assessment used in this study claims that because the games are developed around how people think and act as opposed to how well they play the game, the assessment results are not affected by demographics. Similarly they claim that previous gaming experience does not influence results because, again, how one plays the game is used to make the assessments, not how well it is played. Despite these claims, these variables were assessed as part of this study due to the lack of scientific, empirical research in this area. The demographics survey can be found in Appendix B.

Data Analyses

When this study was initially planned, all the raw score data from all the assessments was going to be used for hypothesis testing. Due to changes made by the company with proprietary rights to the gamification method, though, all of the raw data was not able to be obtained by the time participants were being recruited. That is, originally the gamification method provided results on all 34 constructs that the games measure but changed to only provide the constructs on which participants scored 5 stars, the highest score. This caused complications for assessing hypotheses one and two as there was a considerable reduction in variability.

For hypothesis one, which addressed convergent and divergent validity, data were coded such that the appearance of a gamification construct for a participant was coded as 1 and all other

constructs were coded as zero. Thus, correlations among the gamification constructs and the self-report constructs, in effect, represent point biserial correlations.

Analyzing hypothesis two presented more difficult challenges. Due to being unable to compare mean scores across the assessment methods and faking conditions, another method had to be thought of. In the end, it was decided that exploring the desirability of the gamification constructs and their frequency between faking conditions was the best method. Therefore, the frequency of appearance of the most desirable gamification constructs was compared between faking and non-faking conditions. More information about this analysis is provided in the results section.

The change in the gamification results provided luckily did not impact hypotheses three and four. These hypotheses were tested based off survey results answered by participants and therefore were able to be analyzed through t-tests as originally planned.

Results

In order to test hypothesis one, a correlation analysis was used (see Table 1). Similar and dissimilar constructs between the gamification personality measures and the self-report measures were examined for convergent and divergent evidence. Using the standards recommended by Carlson and Herman (2012) of .70 or higher for convergent validity and .50 or lower for divergent validity, hypothesis one was not supported. The IPIP Imagination scale was significantly correlated to the gamification construct of Creative Expression ($r = .39$, $p < .05$) but it did not meet the .70 convergent validity rule. This was the only predicted significant correlation between the gamification and self-report methods of measuring individual differences. Some other significant correlations were found between constructs that made sense, such as agreeableness and optimism ($r = .35$, $p < .05$), but still none of these correlations meet the

.70 rule. In addition to the lack of desired convergent validity, multiple negative correlations were found where convergent correlations were hypothesized.

Considering the hypothesis that faking would be less viable for the gamification method versus the self-report method, it was considered that the faking manipulation might be interfering with these correlations. Therefore, the correlation analysis was also conducted with only the non-faking participants; however, hypothesis one was still not supported.

To be extra thorough, a chi squared test was also completed. Considering the dichotomous nature of the gamification measure, the self-report measures were split into high and low based off being below or above the 67th percentile for that scale. As it was unknown how difficult it is to achieve a five-star rating on a construct, the researcher did not want to set the bar too high or too low, so the 67th percentile was chosen. For the majority of the tests, the chi squared minimum expected frequency assumption was violated. With the cutoff set at 20%, only 4 tests did not violate this assumption. Of these four, only the test between the gamification construct of empathy and the IPIP achievement striving scale was significant, χ^2 (1, N=35) = 5.22, $p < .05$. Unfortunately, these two constructs were not meant to have convergent validity, so this did not support hypothesis one. Even when considering the results of the tests where the

Table 1

Means, Standard Deviations, and Correlations for Gamification and Self-Report Constructs

Construct	Mean	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Self Confidence	0.03	0.17															
2. Open Mindedness	0.06	0.24	-0.04														
3. Motivation	0.03	0.17	-0.03	-0.04													
4. Optimism	0.34	0.48	0.24	0.08	0.24												
5. Risk Taking	0.14	0.36	-0.07	-0.10	-0.07	-0.12											
6. Exploring Opportunities	0.29	0.46	-0.11	0.12	-0.11	-0.06	-0.08										
7.Consensus Building	0.11	0.32	-0.06	0.30	-0.06	-0.07	-0.15	0.17									
8. Extraversion	0.23	0.43	-0.09	-0.13	-0.09	-0.11	0.17	-0.19	0.02								
9. Anticipating Emotions	0.17	0.38	-0.08	-0.11	0.38*	-0.01	-0.19	0.05	0.08	-0.07							
10. Empathy	0.37	0.49	-0.13	-0.19	-0.13	-0.06	-0.15	-0.09	-0.09	0.00	0.12						
11. Reading Emotions	0.66	0.48	0.12	-0.08	-0.24	-0.37*	-0.05	0.19	0.07	-0.18	-0.15	0.18					
12. Social Intelligence	0.23	0.43	-0.09	-0.13	-0.09	-0.25	0.17	-0.34*	-0.20	0.19	0.11	0.15	0.11				
13. Leadership Initiative	0.11	0.32	-0.06	0.30	-0.06	0.12	0.11	-0.23	-0.13	-0.20	-0.16	0.28	-0.12	-0.20			
14. Perseverance	0.23	0.43	-0.09	-0.13	-0.09	-0.25	-0.03	0.11	-0.20	0.03	-0.07	-0.14	0.11	0.03	-0.20		
15. Diligence	0.40	0.50	-0.14	0.30	-0.14	0.03	-0.33	0.13	0.26	-0.03	0.09	0.10	-0.15	-0.03	-0.11	-0.31	
16. Integrity	0.11	0.32	-0.06	-0.09	-0.06	0.12	-0.15	-0.23	-0.13	-0.20	0.08	0.28	-0.12	0.23	0.15	-0.20	0.26
17. Tenacity	0.20	0.41	-0.09	0.19	-0.09	-0.06	0.41*	0.00	0.05	0.07	-0.23	-0.24	-0.24	0.07	0.05	-0.10	-0.26
18. Creative Initiative	0.31	0.47	-0.12	0.10	-0.12	0.16	0.08	-0.16	-0.05	-0.22	-0.15	-0.27	-0.03	0.22	-0.05	-0.22	-0.18
19. Creative Expression	0.11	0.32	0.48**	-0.09	-0.06	-0.07	-0.15	-0.03	-0.13	0.23	-0.16	-0.09	0.07	0.02	-0.13	0.02	-0.11
20. General Self- Efficacy	4.61	0.36	-0.29	0.06	0.19	0.04	-0.09	-0.06	-0.14	0.15	0.08	-0.13	0.01	0.12	-0.11	0.22	0.02
21. BFI Extraversion	3.36	0.93	0.12	0.34*	-0.16	0.06	-0.04	0.06	-0.02	0.21	-0.17	-0.13	-0.29	0.08	-0.09	-0.01	0.27
22. BFI Agreeableness	4.13	0.50	0.15	0.13	0.15	0.35*	-0.16	-0.20	0.05	-0.08	-0.22	-0.03	0.01	0.12	0.07	-0.31	0.10
23. BFI Conscientiousness	4.10	0.56	-0.34*	-0.12	0.14	-0.06	-0.05	-0.30	-0.27	0.31	-0.18	0.05	-0.06	0.13	0.04	-0.16	0.08
24. BFI Openness	2.09	0.61	0.03	-0.30	0.03	-0.23	-0.10	-0.08	-0.19	0.21	-0.08	-0.03	0.08	0.19	-0.42*	0.08	0.04
25. IPIP Achievement Striving	4.39	0.48	-0.18	0.08	0.18	-0.04	0.07	-0.21	-0.13	0.31	-0.03	-0.17	0.00	-0.01	-0.11	-0.02	0.17
26. IPIP Artistic Interests	4.11	0.62	0.20	-0.22	0.14	-0.02	0.00	-0.11	0.09	-0.04	-0.14	0.08	0.00	-0.03	-0.27	-0.28	0.28
27. IPIP Imagination	3.83	0.69	-0.11	-0.19	-0.06	-0.18	0.02	-0.10	-0.19	0.09	-0.19	0.08	0.13	0.20	-0.19	0.29	-0.16
28. LOT-R Optimism	3.83	0.97	-0.04	0.22	0.13	-0.08	0.02	0.18	0.18	-0.07	-0.04	-0.05	-0.11	0.34*	-0.11	-0.26	0.21
29. Sensation Seeking	1.52	0.25	0.05	0.11	0.05	0.13	0.33	0.27	-0.09	0.19	0.04	-0.30	-0.17	-0.23	-0.24	0.03	0.00
30. Self-Monitoring	1.49	0.20	0.00	0.07	-0.19	-0.15	0.17	0.15	-0.16	0.18	-0.26	-0.13	-0.02	0.22	-0.14	0.07	0.10
31. Social Intelligence Scale	3.41	0.37	-0.02	0.32	0.17	0.04	-0.08	-0.12	0.18	0.00	0.07	-0.25	-0.17	0.13	-0.22	-0.10	0.32
32. Adolescent Leadership Measure	3.25	0.61	0.20	-0.16	-0.15	0.08	0.05	-0.23	0.05	0.18	-0.22	-0.31	-0.03	0.25	-0.31	0.21	-0.02

Note. Constructs 1 through 19 are only rated as a 0 or 1 depending on if the participant was given a 5 star rating on the construct or not. Five star ratings were coded as a one while everything else was coded as a zero.

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

Table 1

Means, Standard Deviations, and Correlations for Gamification and Self-Report Constructs

Construct	Mean	SD	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
1. Self Confidence	0.03	0.17																	
2. Open Mindedness	0.06	0.24																	
3. Motivation	0.03	0.17																	
4. Optimism	0.34	0.48																	
5. Risk Taking	0.14	0.36																	
6. Exploring Opportunities	0.29	0.46																	
7.Consensus Building	0.11	0.32																	
8. Extraversion	0.23	0.43																	
9. Anticipating Emotions	0.17	0.38																	
10. Empathy	0.37	0.49																	
11. Reading Emotions	0.66	0.48																	
12. Social Intelligence	0.23	0.43																	
13. Leadership Initiative	0.11	0.32																	
14. Perseverance	0.23	0.43																	
15. Diligence	0.40	0.50																	
16. Integrity	0.11	0.32																	
17. Tenacity	0.20	0.41	0.05																
18. Creative Initiative	0.31	0.47	-0.05	0.12															
19. Creative Expression	0.11	0.32	-0.13	-0.18	-0.24														
20. General Self- Efficacy	4.61	0.36	0.11	0.08	-0.03	-0.33													
21. BFI Extraversion	3.36	0.93	0.12	0.19	0.08	-0.07	0.18												
22. BFI Agreeableness	4.13	0.50	-0.05	0.00	0.43**	-0.24	0.19	0.05											
23. BFI Conscientiousness	4.10	0.56	0.17	0.21	0.10	-0.21	0.51**	0.10	0.32										
24. BFI Openness	2.09	0.61	0.10	0.25	-0.12	-0.07	0.31	0.31	0.02	0.47**									
25. IPIP Achievement Striving	4.39	0.48	0.14	0.19	-0.04	-0.32	0.76**	0.27	0.26	0.72**	0.43*								
26. IPIP Artistic Interests	4.11	0.62	-0.03	0.17	-0.09	0.01	-0.07	-0.02	0.19	0.25	0.57**	0.15							
27. IPIP Imagination	3.83	0.69	0.06	0.23	0.00	-0.36*	0.50**	0.39*	0.07	0.43**	0.80**	0.45**	0.30						
28. LOT-R Optimism	3.83	0.97	0.05	0.22	0.12	-0.16	0.15	0.22	0.48**	0.21	0.20	0.09	0.28	0.14					
29. Sensation Seeking	1.52	0.25	-0.27	-0.04	0.07	0.00	0.08	0.39*	-0.22	-0.10	0.15	0.13	-0.03	0.08	0.04				
30. Self-Monitoring	1.49	0.20	-0.04	0.23	0.27	-0.02	0.05	0.60**	0.13	0.10	0.39*	0.17	0.22	0.40*	0.16	0.17			
31. Social Intelligence Scale	3.41	0.37	0.11	-0.12	0.22	-0.16	0.22	0.45**	0.02	0.09	0.13	0.21	0.09	0.17	0.23	0.21	0.30		
32. Adolescent Leadership Measure	3.25	0.61	0.03	-0.11	0.22	0.05	0.26	0.42*	0.22	0.07	0.25	0.19	0.03	0.38*	0.07	0.05	0.45**	0.49**	

Note. Constructs 1 through 19 are only rated as a 0 or 1 depending on if the participant was given a 5 star rating on the construct or not. Five star ratings were coded as a one while everything else was coded as a zero.

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

frequency violation was violated, none supported hypothesis one. As a result, in the end, the chi-squared tests also did not support hypothesis one.

In testing faking between measurement methods, desirable and undesirable constructs were determined by two subject matter experts (SMEs) rating the constructs on a scale from "very undesirable" to "very desirable." One SME was a professor with a doctorate in Industrial and Organizational Psychology while the other was a second-year master's degree student in an Industrial and Organizational Psychology program. An Intraclass correlation was computed (ICC (3,2)) which found an average measure ICC was .61 with an 95% confidence interval from .22 to .80 (F (36,36) = 2.29, p <.01). Average ratings between the two SMEs ratings were used to determine the top socially desirable gamification constructs.

In the end there were seven gamification constructs that both raters determined "very desirable. By comparing which of these constructs appear in the faking and honest conditions, we can look at how easy it is to fake using this method. For these seven constructs, three (resourcefulness, perseverance, and integrity) were found more frequently in the faking condition while four (motivation, action oriented, tenacity, and self-confidence) were found to be more frequent in the honest condition. With these seven constructs there were two self-report measures that aligned, the BFI conscientiousness scale as well as the IPIP achievement striving scale. Both of these self-report measures indicated greater faking than honesty. As we will discuss next though the frequencies for self-report should be considered lightly.

When independent samples t-tests were conducted on the self-report methods of the study, no significant differences between the faking and honest conditions were found. Unfortunately, due to the nature of how the gamification constructs were scaled, t-tests could not be conducted on those constructs. The fact that none of the self-report measures in this study

were not found to have significant difference in faking between the conditions should be kept in mind when considering the frequencies data previously mentioned.

A paired-samples t-test was used to examine the mean differences between perceived face validity of the gamification and self-report methods. The t-test on data from the survey question "Do you think this could be an effective method for identifying your individual characteristics?" supported hypothesis three (t (35) = 5.90, $p < .01$). Results showed that the self-report method (M = 4.23, SD = .09) was considered to be more face valid than the gamification method (M = 3.06, SD = .19). A post hoc power analysis was also conducted considering the small sample size. The post hoc power analysis revealed that this hypothesis had a power level of .97 (Buchner, Erdelder, Faul, & Lang, 2014).

T-test results for the data on survey question "If you did not get a job based on this selection method, what would you think of the fairness of this procedure?" also supported hypothesis three (t (35) = 4.99, $p < .01$). The means of the self-report (M = 3.57, SD = .17) and gamification method (M = 2.34, SD = .12) again showed that participants found the self-report method to have more face validity than gamification. For this t-test the power analysis found a power level of .96 (Buchner et al., 2014). When averaging responses to these two questions, results confirmed that participants found the self-report method (M = 3.90, SD = .63) to be more face valid than the gamification method (M = 2.7, SD = .1.18; t (35) = 6.23, $p < .01$). Supporting hypothesis three.

To test enjoyment of measures, a paired samples t-test on the survey question addressing how enjoyable participants found both methods was used. This t-test was found to support hypothesis four (t (35) = 2.63, $p < .05$). From this, we can see that enjoyment was significantly higher for gamification (M = 4.11, SD = 0.87) than for the self-report method (M = 3.57, SD =

0.92). Although lower than the power level for hypothesis three, the power test for hypothesis four still found a good power level of .72 (Buchner et al., 2014).

Discussion

The results of this study suggest that gamification does not show much promise as a measure of individual differences. None of the gamification assessed constructs showed convergent validity with their respective self-report assessed constructs. Although this is an unfortunate result for the gamification method this was one the result of one study through the testing of one gamification method. Based off other more promising results from this study it might be worthwhile for companies to keep working on the reliability and validity of this method. There is the idea, though, that even when people are told that their results will only be used for personal growth, they still will want to see certain results for their own sanity. If this were the case, perhaps gamification is actually still better than self-report measures at measuring personality and only appears to be unable to measure individual differences because it is being compared to biased measures. This would be an idea for future research.

Another strike against the gamification method is the face validity reported by the participants. Both survey items aimed at measuring the face validity exhibited higher face validity results for the self-report method over the gamification method. Interestingly, the first question asking how well the participant thought the method could measure their individual differences had higher means than the second question related to the fairness of the measure. Therefore, even though participants thought that the measures were able to measure their individual differences, they did not think it would be very fair if they were not selected for a position based off these methods. The self-report method was considered somewhat fair by participants while the gamification method was considered slightly unfair. These results would

suggest a research study on process favorability, similar to that done by Shotland et al. (1998), that includes gamification methods could be important in the future.

With previous findings on the importance of fairness in relation to legal cases and perceptions of the organization (Hausknecht et al., 2004; Shotland et al., 1998) this should concern companies that might consider using the gamification method in hiring. With these results it would be extremely important for any company looking at using the gamification method to have proven predictive validity so that if the system was brought to court the company would be able to show the legality of it. However, this still would not help with the prospective problem of damaged perceptions of fairness towards the company that Hauskenecht et al. (2004) and Shotland et al. (1998) found in their research. Hiring top performers is highly important to most companies, and it can be very important not to damage the perceived fairness of the organization among potential applicants. If an applicant applies for a position at a time where there is a better applicant for the position, it would be undesirable to damage the image of the company in the eyes of that individual because the company might want to hire them when another opening becomes available. If the applicant has a negative perception of the company after the first time applying, the company might lose that potential employee to a competitor as a result. Considering Hauckneche et al. (2004) and Shotland et al. (1998) both discussed they thought low face validity in the hiring process might lower an applicant's likelihood of accepting a job offer, as well as their willingness to recommend the employer to others, this could cause a problem with trying to hire the best performers (Armstrong et al., 2016).

Also, if Armstrong et al. (2016) is correct and that even if applicants do accept job offers despite feeling like the selection process had low face validity their still might be problems. These might include low performance, turnover, decreased self-efficacy, and self-esteem. With

the low face validity for gamification found in this study these are things to be concerned about if used in a selection process.

In favor of gamification, hypothesis four showed that gamification was considered significantly more enjoyable than the self-report method. As discussed by Attali and Arieli-Attali (2015) as well as Herzig et al. (2012), enjoyment of gamification is thought to increase behavioral intentions. It is also possible that by making the selection process a little more enjoyable, applicants will have an increased positive perception of the application process, and therefore the company as well. This also might especially be true if they are selected for hire.

Although most results showed little to no support that gamification is something worth considering when it comes to measuring individual differences, results from hypothesis two were more positive. This study suggested there might be some truth to the theory that gamification might provide an assessment method that could avoid the faking problem. Not that the results for hypothesis two were very positive but they did suggest gamification might be less fakable based off frequency results, but t-test results on the self-report measures mitigated these results. If this is the case, it would be worth looking into fixing the convergent validation problems found in this study. By gamification being less susceptible to faking, it could also could reduce the predictive validation and reliability problems that plague self-report measures. Considering the measurement problems that occurred with the data needed to test faking these results should be considered very lightly though and much more research is necessary.

Faking has been mainly discussed as a negative aspect of self-reports and something to be avoided (references). There is another theory, though, that faking is not such a negative behavior. In impression management, the idea is that faking doesn't really exist, but that people are monitoring their behaviors to match the situation that they are in, such as at a job (Hogan et al.,

2012). To build off this the idea, if applicants can monitor their answers to demonstrate what they know an employer wants, they will also be able to monitor their behaviors on the job to match what is desired. Personally, I believe that there is a limit to how long a person can fake their behaviors from how they really act, and therefore, impression management is still faking and should be avoided. Additional research in this area is still needed.

Limitations

There is a concern with this study in regards to the number of participants included in the data analysis. With only 35 participants total, there is a concern with the power and generalizability of this study. Considering this study was conducted in a university setting using students rather than career professionals, it is possible this could be another issue related to the generalizability of the sample. Conducting this study with a larger and more ideal sample is recommended. With the small sample size there was considerably better power than expected for hypotheses three and four. It is still recommended to conduct this study with a larger sample size to ensure greater power.

Another recommendation for future research is to compare more of the gamification constructs to already accredited methods. For hypothesis one only 19 of 37 constructs were tested for divergent and convergent validity. Another study to look at other constructs would add to the research on whether or not gamification is a viable method for measuring individual differences. Additional studies could look at other small sections of the overall gamification results or perhaps a large study could be conducted were there are multiple conditions measuring the small sections of the results and participants are randomly assigned to those small sections so that fatigue does not become an issue. Multiple sections of the study would also be another viable option where participants are asked to come back multiple times to conduct each section

so that everything does not need to be tested at once but ideally each participant still has data across both gamification and already accredited methods in each area.

As with everything in life, this study is not without its flaws. Unlike the self-report measures used in this study, I was not able to obtain reliability information for the gamification measures. The reliability and consistency of these games was also beyond the scope of this study but would be an excellent separate study. One complication of this though is that the app is constantly being updated so the results of the study would only technically be applicable to the version of the app used in the study. The same can be said of this study as well. Once the games have been updated, which is only a matter of time considering the main app used in this research was updated during the duration of this study, the way the app calculates scores is different and would affect all the results of the study. It is also unclear if gamification has predictive validity for measuring job performance and would be another recommended area of research for the future.

Considering the immense literature supporting the idea that people can fake self-report measures (Cascio & Aguinis, 2011; McFarland & Ryan, 2000; Ones et al., 2007; Rees & Metcalfe, 2003), and the fact that this study did not find significant faking differences between the faking and honest condition, it would seem that there might be a problem with the methods used in this study in regards to the faking. Perhaps with this sample the scripts used to try and entice faking and honest answers was not effective, or it's even possible that participants did not really listen to the scripts. Considering the length of the self-report measure given it is also possible that after a while participants forgot how they were told to answer the questions. It is impossible to say for sure what caused the lack of difference between the faking and honest

conditions but based off previous studies the results did not turn out as expected. A similar study address the potential limitations in regards to the faking would be beneficial.

Due to the concerns about the length of the study another avenue that was not explored in this study was adding in a performance criteria. A recommended future research avenue would be to add in a performance criteria and look at predictive performance aspects of both gamification and self-report methods. This would provide potentially additional benefits for gamification.

Conclusion

Despite the findings of this study, gamification still has positive possibilities. As gamification is early in its development, especially in relation to measuring individual differences, it might still be a viable option in the selection process. Until the gamification method can be shown to be a valid system for measuring individual differences, though, it will not matter how enjoyable the method is. Only continued research in this area will show the future of gamification as a method of measuring individual differences.

References

Armstrong, M., Landers, R., & Collmus, A. (2016). Gamifying Recruitment, Selection, Training, and Performance Management. *Emerging Research and Trends in Gamification,* 140-165. doi:10.4018/978-1-4666-8651-9.ch007

Attali, Y., & Arieli-Attali, M. (2015). Gamification in assessment: Do points affect test performance?. *Computers & Education, 83,* 57-63. doi:10.1016/j.compedu.2014.12.012

Berry, C. M., Sackett, P. R., & Tobares, V. (2010). A meta-analysis of conditional reasoning tests of aggression. *Personnel Psychology*, *63*(2), 361-384. doi:10.1111/j.1744-6570.2010.01173.x

Berry, C., Sackett, P., & Wiemann, S. (2007). A review of recent developments in integrity test research. *Personnel Psychology, 60,* 271-301.

Bowler, J. L., & Bowler, M. C. (2014). Evaluating the fakability of a conditional reasoning test of addiction proneness. *International Journal of Psychology, 49*(5), 415-419. doi:10.1002/ijop.12030

Bowler, J. L., Bowler, M. C., & Cope, J. G. (2013). Measurement issues associated with conditional reasoning tests: An examination of faking. *Personality and Individual Differences*, *55*(5), 459-464. doi:10.1016/j.paid.2013.04.011

Brown, V. R., & Vaughn, E. D. (2011). The writing on the (Facebook) wall: The use of social networking sites in hiring decisions. *Journal of Business And Psychology, 26*(2), 219-225. doi:10.1007/s10869-011-9221-x

Buchner, A., Erdfelder, E., Faul, F., & Lang, A. (2014, January 31). G*Power 3.1 manual. Retrieved April 29, 2016, from

http://www.gpower.hhu.de/fileadmin/redaktion/Fakultaeten/Mathematisch-Naturwissenschaftliche_Fakultaet/Psychologie/AAP/gpower/GPowerManual.pdf

Carlson, K. D., & Herdman, A. O. (2012). Understanding the impact of convergent validity on research results. *Organizational Research Methods, 15*(1), 17-32. doi:10.1177/1094428110392383

Carver, C. S., Scheier, M. F., & Segerstrom, S. C. (2010). Optimism. Clinical Psychology Review, 30, 889.

Cascio, W.F. & Aguinis, H. (2011). Applied psychology in human resource management (7th Edition). Prentice Hall.

Chen, G., Gully, S. M., & Doy, E. (2001). Validation of a new general self-efficacy scale. *Organizational Research Methods*, 4, 62-83

Computer basics: Understanding applications. (n.d.). Retrieved November 20, 2015, from http://www.gcflearnfree.org/computerbasics/3

Desimone, J., & James, L. (2015). An item analysis of the Conditional Reasoning Test of Aggression. *Journal of Applied Psychology, 100*(6), 1872-1886. doi:10.1037/ap10000026

Drouin, M., O'Connor, K. W., Schmidt, G. B., & Miller, D. A. (2015). Facebook fired: Legal perspectives and young adults' opinions on the use of social media in hiring and firing decisions. *Computers In Human Behavior*, *46*123-128. doi:10.1016/j.chb.2015.01.011

Goffin, R. D., Jang, I., & Skinner, E. (2011). Forced-choice and conventional personality assessment: Each may have unique value in pre-employment testing. *Personality And Individual Differences*, *51*(7), 840-844. doi:10.1016/j.paid.2011.07.012

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In Mervielde, I., Deary, I., F. De Fruyt,

& Ostendorf, F. (Eds.), *Personality Psychology in Europe*, Vol. 7 (pp. 7-28). Tilburg, The Netherlands: Tilburg University Press.

Hamari, J., Koivisto, J., & Sarsa, H. (2014, January). Does gamification work?--a literature review of empirical studies on gamification. In *System Sciences (HICSS), 2014 47th Hawaii International Conference on* (pp. 3025-3034). IEEE.

Herzig, P., Strahringer, S., & Ameling, M. (2012). Gamification of ERP systems- Exploring gamification effects on user acceptance constructs. *Multikonferenz Wirtschaftsinformatik.*

Hoang, T. G., Truxillo, D. M., Erdogan, B., & Bauer, T. N. (2012). Cross-cultural examination of applicant reactions to selection methods: United States and Vietnam. *International Journal Of Selection And Assessment*, *20*(2), 209-219. doi:10.1111/j.1468-2389.2012.00593.x

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology, 85,* 869-879.

Illies, J. J., Basarich, A. D., Young Illies, M., & Reiter-Palmon, R. (2007, April). *Creativity: The influence of social intelligence, openness, and pressure.* Presentation at the annual convention of the Society of Industrial and Organizational Psychology, New York, NY.

James, L. R., McIntyre, M. D., Glisson, C. A., Green, P. D., Patton, T. W., LeBreton, J. M., & ... Williams, L. J. (2005). A Conditional Reasoning Measure for Aggression. *Organizational Research Methods*, *8*(1), 69-99. doi:10.1177/1094428104272182

John, O. P., & Srivastava, S. (1999). The Big Five Trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin, & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102-138). New York: Guilford Press.

Kluemper, D. H., & Rosen, P. A. (2009). Future employment selection methods: Evaluating social networking web sites. *Journal Of Managerial Psychology*, *24*(6), 567-580. doi:10.1108/02683940910974134

Marlowe, H. A. (1986). Social intelligence: Evidence for multidimensionality and construct independence. *Journal Of Educational Psychology*, *78*(1), 52-58. doi:10.1037/0022-0663.78.1.52

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal Of Applied Psychology*, *79*(4), 599-616. doi:10.1037/0021-9010.79.4.599

McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal Of Applied Psychology*, *85*(5), 812-821. doi:10.1037/0021-9010.85.5.812

McFarland, L. A., Ryan, A. M., & Ellis, A. (2002). Item placement on a personality measure: Effects on faking behavior and test measurement properties. *Journal Of Personality Assessment*, *78*(2), 348-369. doi:10.1207/S15327752JPA7802_09

McFarland, L. A., & Ryan, A. M. (2006). Toward an integrated model of applicant faking behavior. *Journal of Applied Social Psychology*, 36(4), 979-1016.

Mount, M. K., & Barrick, M. R. (1995). The Big Five personality dimensions: Implications for research and practice in human resource management. *Research in Personnel and Human Resources Management, 13*, 153-200.

Mueller-Hanson, R., Heggestad, E. D., & Thornton, G. I. (2003). Faking and selection: Considering the use of personality from select-in and select-out perspectives. *Journal Of Applied Psychology*, *88*(2), 348-355. doi:10.1037/0021-9010.88.2.348

Mumford, M. D., O'Connor, J., Clifton, T. C., Connelly, M. S., & Zaccaro, S. J. (1993). Background data constructs as predictors of leadership behavior. *Human Performance, 6*, 151-195.

"New to Knack?" (2015). Retrieved October 21, 2015, from https://www.knack.it/faq/index.html

Noguchi, Y. (2015, February 25). Recruiting better talent with brain games and big data. Retrieved October 6, 2015, from http://www.npr.org/sections/alltechconsidered/2015/02/25/388698620/recruiting-better-talent-with-brain-games-and-big-data

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology*, *60*(4), 995-1027. doi:10.1111/j.1744-6570.2007.00099.x

Rees, C. J., & Metcalfe, B. (2003). The faking of personality questionnaire results: Who's kidding whom?. *Journal Of Managerial Psychology*, *18*(2), 156-165. doi:10.1108/02683940310465045

Scheier, M. F., Carver, C. S., & Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A re-evaluation of the Life Orientation Test. *Journal of Personality and Social Psychology*, 67, 1063-1078.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings.*Psychological Bulletin, 124*(2), 262-274. doi:10.1037/0033-2909.124.2.262

Sicart, M. (2008, December). Defining Game Mechanics. Retrieved April 05, 2016, from http://gamestudies.org/0802/articles/sicart

Snyder, M. (1974). Self-monitoring of expressive behavior. *Journal of Personality & Social Psychology*, 30, 526-537.

Steiner, D. D., & Gilliland, S. W. (1996). Fairness reactions to personnel selection techniques in France and the United States. *Journal Of Applied Psychology*, *81*(2), 134-141. doi:10.1037/0021-9010.81.2.134

Wilk, S. L., & Cappelli, P. (2003). Understanding the determinants of employer use of selection methods. *Personnel Psychology*, *56*(1), 103-124. doi:10.1111/j.1744-6570.2003.tb00145.x

Zuckerman, M. (1979). Sensation-seeking: beyond the optimal level of arousal. Hillsdale, New Jersey: Erlbaum.

Appendix A: Gamification Constructs with Definitions and Comparison Self-Report Measure

Gamification Construct	Gamification Definition	Comparison Construct
Self-confidence	One's level of trust in their abilities, qualities, and judgements.	General Self-Efficacy Scale
Resourcefulness	Ability to see how to accomplish one's goals and how realistic those goals are.	
Open Mindedness	Receptiveness to new ideas and willingness to follow through on those ideas.	BFI Openness Scale / IPIP Imagination Scales
Resilience	Ability to bounce back from disappoints and learn from their failures.	
Motivation	Whether you feel in control of your circumstances and try to change them or accept them as they are.	IPIP Achievement Striving
Optimism	Whether you are more likely to see the good or bad in a situation.	LOTR Optimism Scale
Self-regulation	One's ability to control their emotions.	
Risk Taking	A person's willingness to take actions with uncertain outcomes.	Sensation Seeking Scale
Exploring Opportunities	Whether a person prefers to stick to the familiar or seek new experiences.	BFI Openness Scale / IPIP Imagination Scales
Self-control	If a person is spontaneous or first thinks through their actions.	
Planning	If a person starts something with or without having everything planned out.	
Action Oriented	How much a person prefers to contemplate all possible outcomes before acting.	
Self-restraint	How one handles exciting situations, by staying calm and collected or act rashly.	
Consensus Building	If a person prefers to work in a group where everyone agrees on the plan or if they are ok with disagreement.	BFI Agreeableness
Extraversion	A preference for being around lots of people compared to in small groups or alone.	BFI Extroversion
Anticipating Emotions	The natural ability to understand how things will affect others and the emotions it will cause them.	Self-Monitoring Scale
Empathy	If a person makes decisions based off of other's viewpoints more or less than their own.	Self-Monitoring Scale

Reading Emotions	How well a person can understand others emotions through subtle cues.	Social Intelligence Scale
Social Intelligence	How much a person can determine social norms and incorporates them into their behaviors.	Social Intelligence Scale
Inspirational Leadership	The extent to which a person is capable of motivating others to complete tasks through their vision and passion.	
Leadership Initiative	The likelihood of an individual taking over a leadership role rather than a submissive role.	Adolescent Leadership Measure
Perseverance	A person's level of determination and resilience to finish task even through difficult circumstances.	IPIP achievement striving
Diligence	How disciplined and organized a person is towards completing their workload.	BFI Conscientiousness
Composure	How well a person handles negative situations.	
Integrity	A preference for fairness, modesty, and sincerity over breaking rules to gain an advantage.	BFI Conscientiousness
Spatial Thinking	Ability to see how things work and how changing one thing will affect other things along with being able to visualize spatial relations.	
Problem Solving	Level of ability to learn quickly, being mentally flexible, to come up with clever solutions, and can adapt to new situations.	
Logical Reasoning	How well a person can identify patterns or rules and apply previously gained information to new problems compared to seeing problems as unique and trying to come up with new creative ways to solve the problem.	
Numbers	Level of mathematics abilities and understanding the relationships among multiple moving pieces.	
Quick Thinking	If a person performs better in fast or slow paced environments.	
Playing to Win	If a person tends to focus on ensuring good outcomes or preventing negative outcomes.	
Managing Ambiguity	How comfortable a person is at making decisions in ambiguous situations.	
Tenacity	How determined a person is to achieve their goals despite challenges.	IPIP Achievement Striving

Creative Initiative	If a person seeks out opportunities to be creative as well as trying it impact others with their creativity or not.	IPIP Artistic Interests / Imagination
Creative Expression	How creative a person is and if they like to be creative in a variety of ways or in limited ways.	IPIP Artistic Interests / Imagination
Systems Thinking	Talent for looking at a whole system and understanding how it works compared to being better at individual components.	
Growth Mindset	Extent to which a person believes they can learn the things they need to meet their goals	

Note. Big Five Inventory is represented by BFI, International Personality Item Pool is IPIP, LOTR is Life Orientation Test-Revised,

Appendix B: Study Surveys

Demographics Survey Participant Code_____

Please circle the response that answers each question.

1. What age category you fall in? 18-25 26-35 36-45 46+

2. What is your gender? Male Female

3. Approximately how many hours a week do you play app games?

0 1-3 4-6 7-10 10+

4. Have you previously played the app ----------:

Yes No

Gamification Survey Participant Code_____

1. Do you think this could be an effective method for identifying your individual characteristics?

0- Very ineffective 1 2 3 4 5- Very effective

2. If you did not get a job based on this selection method, what would you think of the fairness of this procedure?

0- Very unfair 1 2 3 4 5- Very fair

3. I found this test to be…

0- Very unenjoyable 1 2 3 4 5- Very enjoyable

Self-Report Questionnaire Survey Participant Code_____

1. Do you think this method could be an effective for identifying your individual characteristics?

0- Very ineffective 1 2 3 4 5- Very effective

2. If you did not get a job based on this selection method, what would you think of the fairness of this procedure?

0- Very unfair 1 2 3 4 5- Very fair

3. I found this test to be…

0- Very unenjoyable 1 2 3 4 5- Very enjoyable

Appendix C: Study Scripts

Induce honesty scripts:

Test: You are about to take a personality test. I would like you to respond to the questions as though the results will only be used to better understand yourself. That is, the results will only be reviewed by you, and you responses will only be used to help you gain a better understanding of yourself.

Game: You are about to play three personality games. Please play these games naturally as though the information provided by the game will only be used to better understand yourself. That is, the results will only be reviewed by you, and you responses will only be used to help you gain a better understanding of yourself.

Induce faking scripts:

Test: You are about to take a personality test. I would like you to respond to the questions as though the results will be used to decide whether or not you will be offered a high-paying position. That is, your results will only be seen by a hiring manager, and this profile will determine whether or not you will be offered the job.

Game: You are about to play three personality games. I would like you to play the games as though the results will be used to decide whether or not you will be offered a high-paying position. That is, your results will only be seen by a hiring manager, and this profile will determine whether or not you will be offered the job.