# COGNITIVE ABILITIY IN VIRTUAL REALITY: VALIDITY EVIDENCE FOR VR GAME-BASED ASSESSENT

A Thesis submitted to the faculty of
San Francisco State University
In partial fulfillment of
the requirements for
the Degree



Master of Science

In

Psychology: Industrial/Organizational Psychology

by

Erik James Weiner

San Francisco, California

May 2019

CERTIFICATION OF APPROVAL

I certify that I have read Cognitive Ability in Virtual Reality: Validity Evidence for VR Game-based Assessment by Erik James Weiner, and that in my opinion this work meets the criteria for approving a thesis submitted in partial fulfillment of the requirement for the degree Master of Science in Psychology: Industrial/Organizational Psychology at San Francisco State University.

Diana R. Sanchez, Ph.D.
Assistant Professor of Psychology

Christian Wright, Ph.D.
Professor of Psychology

# COGNITIVE ABILITIY IN VIRTUAL REALITY: VALIDITY EVIDENCE FOR VR GAME-BASED ASSESSENT

Erik James Weiner
San Francisco, California
2019

The purpose of this study is to evaluate validity evidence for assessing cognitive ability using VR game-based assessment scores. Participants completed a series of VR game-based assessments, self-report cognitive ability assessments, and additional scales. Convergent validity was evaluated through comparisons between VR scores with self-report assessment scores. Divergent validity was evaluated through comparisons between VR scores with five factors of personality (i.e., openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism) based on the Five Factor Model of personality traits. Criterion-related validity was evaluated using the associations between VR scores and academic performance (i.e., GPA). Exploratory analyses examined incremental validity and adverse impact. Results demonstrated promising relationships between the compared formats of VR and self-report. These results indicate the need for further research to examine the qualities of VR games that contribute to validity in assessing cognitive ability and other individual characteristics, seeing as the VR assessment format may provide unique benefits in assessing certain specific abilities for selection compared with traditional measures.

I certify that the Abstract is a correct representation of the content of this thesis.

Chair, Thesis Committee

5/15/19
Date

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figures Page

LIST OF APPENDICES

Appendix Page

## Introduction

There is an increasing volume of research literature exploring video games for a wide range of applications in the workplace, such as employee selection and performance management (Chamorro-Premuzic, Winsborough, Sherman, & Hogan, 2016; Lowman, 2016). Given the scope of this research, it should be anticipated that findings have supported certain applied uses of video games. Accordingly, the research examining video game-based assessments has demonstrated validity evidence comparable to that of traditional self-report assessments (Hummel, Brinke, Nadolski, & Baartman, 2017; Kiili, Devlin, Perttula, Tuomi, & Lindstedt, 2015). However, a drawback to the current evidence is the limitation to two-dimensional (2D) game formats, rather than other game formats. That is, research exploring the application of games in a virtual reality (VR) format is sparse, despite the VR format prominently featuring qualities that are broadly cited as benefits of video games, such as realism and immersion (Hvass et al., 2017; Shin, 2017). Considering the rapid rate at which VR technology is improving, VR games will likely find a place in similar settings as 2D games (Sykes, 2018). It is crucial that VR technology is evaluated for evidence to support the validity of its anticipated future use.

There exists some peer-reviewed research investigating the practical applications of current-generation, high-fidelity (i.e., hyper-realistic) VR technology (Aïm, Lonjon, Hannouche, & Nizard, 2016; Freeman et al., 2017; Gavish et al., 2015), despite its relative recency compared with older generations of VR technology that have been more widely studied in research. A majority of the research on current-generation VR

technology has focused on limited topics, such as skills training and physical rehabilitation (Aminov, Rogers, Middleton, Caeyenberghs, & Wilson, 2018; Jensen & Konradsen, 2018). As a result, there is presently a dearth of research investigating the use of VR games for assessment purposes. Given the infancy of this research, it is important to explore which specific applications warrant further investigation, such as certain assessment types that may exhibit evidence for validity and/or utility in a VR format.

The potential utility of VR games in assessment is largely generated by the unique behaviors that can be prompted and observed in VR environments, which can provide likewise unique measurements of skills and abilities that are associated with these behaviors. For example, cognitive ability encompasses multiple specific abilities, such as spatial reasoning, which may be most accurately assessed in a realistic environment. Because VR technology is capable of simulating such environments while collecting objective behavioral data points, we chose to explore the measurement of cognitive ability in VR. In addition, cognitive ability is one of the most commonly-used occupational assessments, supported by consistent evidence that cognitive ability predicts job performance (Kuncel & Hezlett, 2010). Validity coefficients, i.e., correlations between cognitive ability assessment scores and job performance ratings, have been commonly estimated at approximately $r = .50$, indicating strong predictive validity in comparison with other common selection assessments (e.g., interviews, personality measures, assessment centers; Outtz, 2002; Schmidt & Hunter, 1998).

Despite the strong evidence for cognitive ability as a predictor of job performance, a primary concern for the use of cognitive ability assessments in selection is that they can result in adverse impact against ethnic minority subgroups, often favoring White applicants over subgroups such as Black and Latino applicants (Ployhart & Holtz, 2008). Still, VR technology might help mitigate adverse impact resulting from cognitive ability scores, further justifying the exploration of VR games in this context. Research has shown that certain technological advancements have helped mitigate adverse impact in other assessment types, such as the use of a video-based assessment format in situational judgement tests as a strategy to reduce construct-irrelevant variance, i.e., variance in assessment scores that is associated with differential test performance among demographic subgroups but irrelevant to the targeted measure (Outtz, 2002). Specifically, in game-based assessments, preliminary research supports the use of video game assessments to reduce adverse impact in the assessment of cognitive ability (Montefiori, 2016). However, this research predominantly features 2D games. Thus, this study is one of the first to extend this literature into a 3D VR environment in order to explore the potential benefits of a more immersive game-based environment. We were unable to find any prior studies using VR games to assess cognitive ability.

The aim of the present study is to evaluate validity evidence for the assessment of cognitive ability through a VR game-based assessment format. In addition, measures of adverse impact in VR game-based assessment scores will be evaluated as further criteria

of interest pertinent to researchers and practitioners who wish to explore the effects of VR game-based assessments.

## VR Technological Advancements and Applications

Existing research on the utility of VR technology has focused on older generations of VR with veritable differences from the current generation of VR technology (Freina & Ott, 2015; Gigante, 1993). These differences indicate that the research findings relative to older forms of VR technology may not be extended to the potential applications of the current generation of VR technology without further research-based justification. The lack of applicable research on the utility of the current generation of VR technology, as well as the potential benefits of this technology, justify the importance of the current study. In addition, the lack of current VR assessment research warrants the consideration of other (i.e., 2D) video game formats, rather than strictly VR formats, in the discussion of video game literature throughout this introduction.

**VR definitions and history.** VR technology refers to electronic console devices that are used to simulate realistic and immersive environments through head-mounted visual displays, often involving narratives and objectives that may require input via controllers from the individual or individuals who are using the device and are immersed in the simulated environment. VR is most commonly defined by the following characteristics, as delineated by Gigante (1993): "The illusion of participation *in* a synthetic environment rather than external observation of such an environment. VR relies

on three-dimensional (3D), stereoscopic, head-tracked displays, hand/body tracking and binaural sound. VR is an immersive, multi-sensory experience" (p. 3). VR technology is differentiated from traditional 2D gaming technology by its incorporation of features such as head-mounted displays and physical tracking in order to place the player directly into a first-person perspective, rather than a third-person perspective.

VR technology has been historically less prevalent than its 2D counterparts, but VR has increased in popularity over the years along with other changes in factors relevant to its appeal to the typical consumer. For instance, VR technology has become more accessible by overcoming constraints such as cost. Older VR technology was highly inaccessible to many consumers in its earlier years of existence due to high cost (Engler, 1992). However, newer VR technology has decreased in cost and increased in unit sales, with over one million units sold across VR platforms in the third fiscal quarter of 2017 alone, constituting the first recorded instance of quarterly VR console unit sales exceeding the threshold of seven figures (Stanton et al., 2017). Similarly, generations of VR technology have varied in fidelity, i.e., the degree of realism replicated in the VR environment compared with reality. Among other limitations, older VR technology featured visual displays with low refresh rates, i.e., the rates at which individual static images shown on a visual display change in order to simulate motion or other environmental changes (Vince, 1993). Conversely, newer VR consoles feature visual displays with relatively high refresh rates, such as the Oculus Rift, which refreshes images at a rate of 90 Hz (Martindale, 2018).

In addition to increased accessibility and fidelity, current VR technology continues to reflect properties of *ludus*, which generate fun and enjoyable experiences for those who use this technology. As defined by Frasca (1999), *ludus* refers to "activity organized under a system of rules that defines a victory or a defeat, a gain or a loss." In essence, ludic properties are the qualities unique to the format of a game that facilitate interactive play, as opposed to other non-interactive media such as literature or film. Ludic properties are enjoyable by definition, so these properties yield entertainment value for any activity into which they are incorporated. This entertainment value equates to potential utility for video games in other domains apart from entertainment. Because the current body of research demonstrating the utility of VR games is limited, evidence for the utility of the current generation of 2D video games will be discussed first.

**Video games in applied settings.** Research evaluating the utility of 2D video games in various applied settings has yielded promising results. For instance, O'Connor et al. (2000) found support for the use of video games in healthcare, observing in-game skill acquisition in wheelchair users who completed a computer-based simulation game with a custom wheelchair-based controller designed to replicate real-life wheelchair operation. In the classroom, video game-based learning is associated with increases in cognitive gains and positive reactions among students, compared with traditional methods of instruction (Nte & Stephens, 2008; Vogel et al., 2006). Support has been demonstrated for video games in occupational skills training, such as the use of simulation games in combination with practical training to increase skills acquisition in

both aircraft training and surgical training, compared with practical training alone for either set of skills (Aggarwal et al., 2007; Hays, Jacobs, Prince, & Salas 1992; Lampton, Bliss, Orvis, Kring, & Martin, 2009; Orvis, Moore, Belanich, Murphy, & Horn, 2010). In general, computer-based simulation games have demonstrated utility in employee knowledge and skills training due to the intrinsically motivating ludic properties that are unique to their video game-based format (Sitzmann, 2011).

**Validity and utility of video games in assessment.** Video games have also demonstrated evidence for validity and utility in assessment across various contexts. For instance, video games developed using the EMERGO platform have been evaluated in order to establish content validity for in-game performance indicators in assessing information technology systems management skills (Hummel, Brinke, Nadolski, & Baartman, 2017). The game *Semideus* has demonstrated evidence for validity in the assessment of rational number knowledge through subject matter expert ratings of game content validity, in addition to demonstrating validity evidence through a correlational analysis comparing scores generated by the game with scores on a paper-based rational number test (Kiili, Devlin, Perttula, Tuomi, & Lindstedt, 2015). Another game called *Newton's Cradle* has similarly demonstrated validity evidence for assessing physics knowledge pertaining to Newton's three laws though a correlational analysis comparing game scores with scores on a computer-based qualitative physics test (Shute, Ventura, & Kim, 2013). Although research on the validity and utility of 2D video game-based assessments has varied in its targeted populations (e.g., occupational versus educational),

as well as in its targeted measures (e.g., skills versus knowledge), this literature has been encouraging in its early stages with respect to its support for the assessment capabilities of 2D video games.

The importance of exploring alternative assessment formats, apart from traditional paper-based and computer-based formats, has been reinforced through the literature exploring emergent properties of 2D game-based assessment formats. One considerable advantage of game-based assessments is the utility of this assessment format in mitigating attempts at faking, i.e., assessees' deliberate selection of responses perceived to yield assessment scores that indicate a socially desirable level of the construct targeted by the assessment. While traditional assessments typically feature content that is transparently linked with the targeted measures, as well as methods of measurement that are easily manipulable (e.g., through selecting a certain response on a five-point Likert scale), video game-based assessments may not be as transparent in their relationship with the targeted measures or as easily manipulable to generate scores perceived by assessees to be socially desirable (e.g., through altering in-game performance). Because faking requires some degree of knowledge on the part of the assessee regarding the construct targeted by the assessment, and video game-based assessments can obscure such knowledge by measuring multiple constructs simultaneously and by using in-game behavioral performance indicators that are less intuitively linked with these constructs than self-report responses tend to be, video game-based assessment formats can be used

to reduce faking and obtain more accurate data compared with traditional assessment formats (Bhatia & Ryan, 2018).

Another advantage of video game-based assessment formats pertains to construct-irrelevant variance in assessment scores. In paper-based and computer-based self-report assessments, construct-irrelevant variance can arise due to qualities that are innate to these formats and result in adverse impact among demographic subgroups (Outtz, 2002). To approach resolving this issue, meta-analytic research findings support the strategy of decreasing excess cognitive loading in assessments by reducing verbal and reading ability requirements, such as through the use of video-based assessments and other multimedia assessment formats (Ployhart & Holtz, 2008). Because video game-based assessments frequently rely upon visually-displayed graphic content rather than written content, these formats can similarly be used to reduce verbal and reading ability requirements, in turn reducing the construct-irrelevant variance that is often associated with traditional assessment formats (Zapata-Rivera & Bauer, 2012). Additionally, other recent research has yielded promising findings pertaining to the reduction of construct-irrelevant variance unique to video game-based assessment formats, such as the development of a video game experience (VGE) scale to statistically account for differential levels of VGE in relation to performance on a video game-based assessment (Sanchez & Langer, 2018).

Overall, this research composes a small sample of the total literature that has emerged in support of the use of video games in assessment. In the following section, applications of video games will be explored specific to VR technology.

**Current state of applied VR research.** Through the available research on the applications of VR technology, there is a limited scope of existing support for its utility in domains apart from entertainment. In healthcare, VR has demonstrated some utility as an evaluative tool, such as in the evaluation of differential locomotive patterns across age groups (Janeh, Bruder, Steinicke, Gulberti, & Poetter-Nerger, 2018). However, much of the research examining the applications of VR technology in healthcare focuses on physical rehabilitation and treatment, rather than assessment or evaluation; furthermore, this literature is imprecise in its definition of VR, often including technology that lacks head-mounted displays and other immersive characteristics that are innate to the common definition of VR technology (Aminov, Rogers, Middleton, Caeyenberghs, & Wilson, 2018; Brunner et al., 2014; Cameirão, Bermúdez i Badia, Oller, & Verschure, 2010; Gigante, 1993). Other research outside of healthcare has demonstrated the utility of VR technology in skills training across occupational and educational settings, predominantly including physical (i.e., psychomotor) and cognitive (i.e., spatial) skills (Jensen & Konradsen, 2018). Regardless, there is minimal research to support the use of VR technology in assessment.

Given the practical utility of video game-based assessment formats, additional research is warranted in order to inform the continued and future applied use of these assessment formats. VR technology is especially viable for such research, given the recency of its current high-fidelity and accessible generation of consoles such as the Oculus Rift as well as the increasing number of studies investigating the applications of

VR technology. This trend is comparable to that of the technological advancements in telecommunication and subsequent changes to the job interview process at large. While job interviews were originally conducted in person and without any technological facilitation, the interview process was changed as telephones were introduced into the workplace, later followed by video-based conferencing technology; both of these technological applications gained popularity due to practical advantages such as time and cost, ultimately changing the way that job interviews could be conducted (Blackman, 2002; Toldi, 2011). Similarly, as VR technology continues to gain popularity in overall use, along with 2D video games gaining popularity specifically in the domain of assessment, it is likely that emergent practical advantages across both formats will draw increased attention from employers. This likelihood may increase for VR technology as ongoing developments continue to address practical concerns and increase the appeal of this technology in applied settings. For instance, the upcoming Oculus Quest console will constitute the first major step toward the facilitation of meetings via telecommunication in a virtual environment (Zuckerberg et al., 2018).

Despite this trend, the literature on applications of VR technology remains minimal in scope, and there is an even more limited scope of research exploring VR games as assessment tools. If this format is to be used for similar purposes as the 2D game format, then there is a need for new research to validate the use of VR technology for these purposes, in turn informing best practices. Considering the lack of research

investigating assessment through VR technology, the immediacy of the need for research exploring this application is affirmed.

In the following section, the specific type of assessment that is examined in this study, i.e., cognitive ability assessment, will be described. This description will include the definition, validity, and utility of this assessment type, in addition to its disadvantages and the proposed mitigation of these disadvantages through a VR game-based assessment format.

## Cognitive Ability Assessment and Adverse Impact

Cognitive ability is one of the most commonly assessed constructs in occupational settings due to the strong evidence for its criterion-related validity in predicting job performance (Kuncel & Hezlett, 2010). Thus, cognitive ability assessments will serve as the assessment-type of focus for this study in order to maximize the utility of findings.

**Cognitive ability definitions and taxonomy.** Cognitive ability is a broad construct that typically functions as a measure of human intelligence. Gottfredson (1997) collaborated with 51 other intelligence experts to define intelligence, as measured by cognitive ability: "Intelligence is a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience" (p. 13). Throughout much of the intelligence literature, cognitive ability is structured hierarchically, with a general cognitive ability (or *g*, the general intelligence factor) underlying more specific abilities (Carroll, 1993). Individuals who possess the same level of *g* may differ in their specific

abilities due to differences in the extent to which *g* is uniquely applied across these abilities (Ones, Dilchert, Viswesvaran, & Salgado, 2017). Because of this, such individuals may also receive differential scores on the same cognitive ability assessment despite possessing the same level of *g*, depending on the specific abilities that are targeted by the assessment (Carroll, 1993).

**Adverse impact of cognitive ability assessments.** The literature on cognitive ability generally demonstrates strong evidence for the validity and utility of cognitive ability assessments in employee selection (Kuncel & Hezlett, 2010). Validity coefficients typically range across studies from $r = .35$ to .55, establishing relatively strong criterion-related validity evidence for cognitive ability assessments in predicting assessees' job performance (Ones, Dilchert, Viswesvaran, & Salgado, 2017). However, cognitive ability assessments are also associated with a potentially critical disadvantage. Namely, the assessment of cognitive ability often results in adverse impact between demographic subgroups.

In employment law, adverse impact refers to the negative and illegal effect resulting from a given workplace practice upon employment opportunities for any federally protected class (Equal Employment Opportunity Commission [EEOC], 1978). Adverse impact is clear in the comparison of cognitive ability assessment scores between ethnic subgroups. This comparison yields very large differences that demonstrate statistical favorability of White assessees over Black assessees ($d = .99$), medium to large differences favoring White assessees over Latino assessees ($d = .58$ to .83), and small

differences favoring Asian assessees over White assessees ($d$ = -.20; Ployhart & Holtz, 2008). Apart from these subgroups, meta-analytic research indicates that, while male and female assessees differ in some specific abilities, there is little or no adverse impact between cognitive ability assessment scores between genders (Ones, Dilchert, Viswesvaran, & Salgado, 2017). For other demographics such as age, research exploring the adverse impact of cognitive ability assessment scores between different subgroups is limited.

Regardless, adverse impact in cognitive ability assessment scores between ethnic subgroups remains a prominent concern (Outtz, 2002). Some have identified this to be an up-the-river problem, indicating that differences in cognitive ability across demographic subgroups may be real but not innate; that is, cognitive ability assessments are accurate in producing differential scores between subgroups, but these differences arise from lifelong systemic inequities which are unrelated to the fairness of an assessment (Sackett, Borneman, & Connelly, 2008). Others have focused on reducing adverse impact by proposing alternative strategies for formatting assessments of cognitive ability, such that validity is preserved while adverse impact is reduced (Ployhart & Holtz, 2008). Proponents of these strategies fundamentally seek to reduce construct-irrelevant variance (Helms, 2006).

Through the research evaluating strategies of preserving validity evidence for assessments of cognitive ability while reducing adverse impact among demographic subgroups, there is support for the use of alternative assessment formats, such as video-

based assessments (Ployhart & Holtz, 2008). As discussed previously, video game-based assessment formats have also received support for their minimization of adverse impact in cognitive ability assessment scores (Montefiori, 2016). Although research supports the use of video game-based assessments to reduce adverse impact in assessment scores between demographic subgroups, this literature remains in its early stages and pertains mostly to 2D video games. With this in mind, the present study is intended to further examine adverse impact across various demographic subgroups in the assessment of cognitive ability through a VR game-based format.

**Exploring Validity Evidence to Assess Specific Cognitive Abilities in VR**

While adverse impact will be evaluated in this study, it is critical that validity evidence for the VR game-based assessment of cognitive ability is demonstrated. In the context of assessment, validity refers to the accuracy of the conclusion ascertained by an assessment score (American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME], 2014). The traditional tripartite model of validity includes three primary sources of validity evidence: content validity, construct validity, and criterion-related validity (EEOC, 1978). Respectively, content validity evidence indicates that assessment scores are based on content that is representative of the characteristics targeted by the assessment; construct validity evidence indicates that assessment scores are related with other measures of attributes composing the construct that is targeted by the assessment; and criterion-related validity evidence indicates that assessment scores can be used to

predict a given outcome associated with the construct that is targeted by the assessment (AERA, APA, & NCME, 2014).

Recent literature and best practice standards in assessment validation have largely departed from the tripartite model of validity. Instead, validity is emphasized to be a unitary concept indicating whether an assessment measures what it purports to measure, and validity evidence should be obtained from a variety of sources that need not conform to the tripartite model (AERA, APA, & NCME, 2014). Therefore, this study includes multiple sources of validity evidence that are evaluated irrespective of the tripartite model of validity.

**Primary measures of validity.** Because the current generation of VR technology is relatively new, the validity of the content featured in VR game-based assessments has not been sufficiently explored. It is important that subject matter expert input be obtained so that the content of new potential VR game-based assessments can be evaluated by experts in fields relevant to the measures targeted by these assessments. This input can be used to establish a preliminary level of confidence that the behaviors exhibited by assessees in the VR environment will target the anticipated specific abilities. For this reason, subject matter expert input will be obtained in order to verify that the content and procedures of the VR game-based assessments used in this study will target their respective specific abilities.

Beyond this preliminary evaluation, evidence will be evaluated for convergent validity and divergent validity. Convergent validity indicates that assessment scores

demonstrate relationships with other measures that are associated with the targeted construct, while divergent validity indicates that assessment scores do not demonstrate relationships with measures that are orthogonal to the targeted construct (AERA, APA, & NCME, 2014). In order to evaluate convergent validity of the VR game-based assessments, scores for each VR assessment (i.e., for each targeted specific ability) will be compared with corresponding scores for each self-report (i.e., traditional computer-based) assessment that is measured in this study. Previous research has supported the matching of a targeted construct with a video game for assessment purposes, based on the identification of specific observable behaviors in the video game that correspond with the construct (Shute & Emihovich, 2018). This strategy was used to select the VR games intended to measure specific abilities in the present study, based on the matches between these specific abilities and the gameplay behaviors that are observable and scorable through the VR games. It is anticipated that significant relationships will be demonstrated between assessments across both formats.

> ***Hypothesis 1.*** For each specific ability, VR game-based cognitive ability assessment scores will demonstrate a significant relationship with self-report cognitive ability assessment scores.

Next, personality will be measured in order to evaluate divergence. The literature on cognitive ability supports the divergence between measures of cognitive ability and measures of the Big Five OCEAN personality traits, i.e., openness, conscientiousness, extraversion, agreeableness, and neuroticism (Barrick & Mount, 1991). Although

cognitive ability and conscientiousness are both linked with job performance, such that both measures have demonstrated criterion-related validity evidence in the prediction of performance, these two measures are minimally related to one another (Cortina, Goldstein, Payne, Davison, & Gilliland, 2000). Based on this evidence, no significant relationships are anticipated between VR game-based assessment scores and self-report levels of openness, conscientiousness, extraversion, agreeableness, or neuroticism.

> ***Hypothesis 2.*** For each specific ability, VR game-based cognitive ability assessment scores will demonstrate orthogonal relationships with self-report levels of openness, conscientiousness, extraversion, agreeableness, and neuroticism.

Additionally, criterion-related validity of the VR game-based assessments will be evaluated with participant grade point average (GPA) serving as the performance measure of interest. Although job performance is the criterion by which assessment validity is traditionally evaluated in occupational settings, GPA is often used as a proxy measure for job performance in research featuring university student samples, such as the sample featured in the current study (Cotton, Dollard, & de Jonge, 2002; Hysenbegasi, Hass, & Rowland, 2005; Schmit, Ryan, Stierwalt, & Powell, 1995). Furthermore, the established relationship between GPA and cognitive ability indicates that a valid measure of cognitive ability should be expected to demonstrate a relationship with GPA (Imose & Barber, 2015). Thus, it is anticipated that VR game-based assessment scores will be significantly related to participant GPA.

***Hypothesis 3.*** For each specific ability, VR game-based cognitive ability assessment scores will demonstrate a significant relationship with GPA.

**Additional measures of validity and adverse impact.** Incremental validity evidence indicates that assessment scores can be used to increase the predictive ability of another assessment with an established level of predictive ability, with respect to a given outcome (Hough & Dilchert, 2017). After testing the primary study hypotheses, an exploratory analysis will be conducted in order to determine the incremental validity of the VR game-based assessment scores over the predictive ability of the self-report assessment scores, with respect to the prediction of participant GPA.

***Research Question 1.*** For each specific ability, will VR game-based cognitive ability assessment scores demonstrate significant incremental validity in the prediction of participant GPA over self-report cognitive ability assessment scores?

Finally, as previously discussed, adverse impact among demographic subgroup scores constitutes one of the most prominent drawbacks of cognitive ability assessments in occupational settings. Because adverse impact yields major implications for the utility of cognitive ability assessments, adverse impact will be evaluated as an emergent effect of the VR game-based assessment format. If this format is identified as potential method of demonstrating validity in the measurement of cognitive ability while minimizing adverse impact in cognitive ability assessment scores, then the findings of this study will

provide further support for the utility of VR technology in assessment (i.e., cognitive ability assessment).

***Research Question 2.*** For each specific ability, will VR game-based cognitive ability assessment scores and self-report cognitive ability assessment scores demonstrate significant differences in adverse impact between demographic subgroups?

## Method

### Participants

Participants were 124 students from a large Western university. This sample was primarily college-aged ($M$ = 24 years, $SD$ = 7 years) and female (71%), with approximately 28% identifying as White (i.e., "White, Caucasian, or other European"; or "Middle Eastern or North African"), 28% as Latino (i.e., "Latina/o/x or Hispanic"), 25% as Asian (i.e., "Asian American or other East Asian"; "Indian, Pakistani, or other South Asian"; or "Native Hawaiian or Pacific Islander"), and 19% as Other (i.e., "Black or African American"; "Native American or Alaska Native"; or any combination of multiple ethnic subgroups). The distribution of participant demographics is provided in Table 1.

### Procedures

Prior to arrival, each participant was assigned to one of two conditions, which differed in counterbalancing the order of the two assessment formats (i.e., the VR game-based assessments and the computer-based self-report assessments). Upon arrival, the

participant signed in and completed the Consent Form and the Video Release Form. The participant then completed the two assessment formats in their assigned order.

Researchers assisted participants with the VR equipment during the VR game-based assessments. This included general instructions on using the VR controllers. For the VR assessments, participants completed three VR games, receiving verbal instructions specific to the objectives and controls of the VR games prior to beginning each respective game.

For the computer-based self-report assessments, the researcher provided the participant with verbal instructions on the general procedures for the assessments prior to beginning the first assessment. Instructions and sample items specific to each assessment were provided on-screen prior to the start of the respective assessment. There were three self-report assessments in total, each with a time limit of five minutes (not including the time spent reading instructions and completing sample items). The participant could self-pace their completion of each individual assessment within its five-minute time limit.

Lastly, participants completed a survey including demographic information (i.e., age, gender, ethnicity), as well as two video game-relevant scales (i.e., video game experience, VGE, and intimidation with games, IWG), and a personality scale (i.e., IPIP-120). Student identification numbers were temporarily recorded to retrieve and match each participant's grade point average (GPA) to their survey and assessment results. GPA served as the criterion measure for performance. At the end of the study, the researcher debriefed and excused participants. The study took approximately 75 minutes to

complete, and each participant received course credit to compensate for the time and effort required for their participation in the study.

**Materials**

**Self-Report Assessments.** The computer-based self-report assessment measures for specific cognitive abilities were taken from a cognitive ability assessment battery called the Employee Aptitude Survey (EAS). The EAS assessment battery includes ten assessments in total, and three of these assessments were selected for use in the study because they included content relevant to behaviors that were feasible to perform in a VR environment. These assessments are described below, including Space Visualization, Visual Speed & Accuracy, and Visual Pursuit.

***Space Visualization (Self-Report Assessment).*** Space Visualization (SV) refers to "the ability to imagine objects in three-dimensional space and to manipulate objects mentally" (Ruch, Stang, McKillip, & Dye, 1994, p. 20). The SV assessment included 50 items ($\alpha$ = .89). Previous validation research has shown the SV assessment to be correlated with scores on the several other cognitive abilities assessments, such as the Primary Mental Abilities (PMA) Tests for space ability ($r$ = .58) and reasoning ability ($r$ = .46; Ruch, Stang, McKillip, & Dye, 1994). For the SV assessment, participants were provided with five minutes to complete the assessment. Participants were instructed to look at a given block within an arrangement of several surrounding blocks and indicate how many other blocks it was touching. Scores were calculated as one-fifth of the total number of incorrect responses, subtracted from the total number of correct responses,

rounded to the nearest whole number, and transformed into a percentile against normative data included with the assessment. Scores ranged from 1.00 to 97.00 ($M$ = 47.44, $SD$ = 31.22). Sample items are provided in Figure 1.

***Visual Speed & Accuracy (Self-Report Assessment).*** Visual Speed & Accuracy (VSA) refers to "the ability to compare numbers or patterns quickly and accurately" (Ruch, Stang, McKillip, & Dye, 1994, p. 20). The VSA assessment included 150 items ($\alpha$ = .91). Previous validation research has shown the VSA assessment to be correlated with scores on the several other cognitive abilities assessments, such as the PMA Tests for space ability ($r$ = .30) and reasoning ability ($r$ = .50; Ruch, Stang, McKillip, & Dye, 1994). For the VSA assessment, participants were given five minutes to complete the assessment. Participants were instructed to rapidly scan through pairs of numbers and indicate whether they were the same or different. Scores were calculated as the total number of incorrect responses subtracted from the total number of correct responses, and transformed into a percentile against normative data included with the assessment. Scores ranged from 1.00 to 98.00 ($M$ = 33.94, $SD$ = 32.08). Sample items are provided in Figure 2.

***Visual Pursuit (Self-Report Assessment).*** Visual Pursuit (VP) refers to "the ability to make rapid, accurate scanning movements with the eyes" (Ruch, Stang, McKillip, & Dye, 1994, p. 20). The VP assessment included 30 items ($\alpha$ = .86). Previous validation research has shown the VP assessment to be correlated with scores on the several other cognitive abilities assessments, such as the PMA Tests for space ability ($r$ =

.53) and reasoning ability ($r$ = .44; Ruch, Stang, McKillip, & Dye, 1994). For the VR assessment, participants were given five minutes to complete the assessment. For each item, the participant was instructed to use visual scanning to follow a given line from its starting point, through a tangle of different lines, and indicate which among several endpoints was connected with the targeted starting point. Scores were calculated as one-fourth of the total number of incorrect responses, subtracted from the total number of correct responses, rounded to the nearest whole number, and transformed into a percentile against normative data included with the assessment. Scores ranged from 1.00 to 97.00 ($M$ = 37.45, $SD$ = 26.10). Sample items are provided in Figure 3.

**VR Assessments.** A series of VR games was identified by a group of researchers who explored current VR games and other programs that had possible relevance for the current research question. Several games were purchased and evaluated for their demonstration of cognitive ability through automatically generated scores and other in-game behaviors. All games considered for inclusion were played on the Oculus Rift console, requiring the use of a head-mounted display and two handheld controllers. The final battery of VR games was selected due to their demonstration of specific abilities analogous to those measured through cognitive ability assessments like those described above (i.e., VP, VSA, and SV). Each of the three VR games included in the final study is described below.

***Space Visualization (VR Assessment).*** The game *ThrounnelVR* was used to evaluate SV. In this game, participants were shown a shape made of several connected or

free-floating cubes, which they were instructed to rotate and position so that the cubes could pass through the approaching wall. The wall would contain one or more cube-shaped holes, allowing the cubes to pass through if rotated and positioned correctly. When the cubes were rotated and aligned to the correct position, this would allow them to pass through the wall so that the participant could continue playing. As the cubes passed through the wall, the game would increase in difficulty. The shape of the holes in the wall would change, which forced participants to select the correct rotation and position of the cubes as quickly as possible. Failure to select the correct position, or collision with the wall, would result in a game over, restarting all progress (i.e., resetting the number of walls that had been passed, which influenced the resultant score displayed at each game over screen). The participant was also able to instantly send the cubes to the wall once they found the position of the cubes that they believed would fit through the hole in the wall. Selecting this instant option would result in more points if correct. However, being at a greater distance from the wall could result in misjudgment and misalignment, which would result in collision with the wall, yielding a game over. Scores were calculated as an average of the in-game scores that the participant received at each game over screen, and transformed into a percentage. Scores ranged from .01 to .46 ($M = .14$, $SD = .10$). A screenshot of this game is provided in Figure 4.

***Visual Speed & Accuracy (VR Assessment).*** The game *CubeWorks* was used to evaluate VSA. In this game, participants saw a conveyor belt with cubes passing in front of them. They were instructed to retrieve cubes from the conveyor belt as they passed by

and to match cubes with corresponding patterns on their sides. Cubes were matched by pressing the matching patterns against one another, face-to-face. Cubes had multiple sides with different patterns, and the patterns could be matched in different ways (e.g., matching or opposing shapes, akin to a plug-and-socket design). The rounds became increasingly difficult. In the first round, participants were given five minutes to match 20 cubes. In the second round, participants were given five minutes to match 15 cubes, matching only a particular pattern, which would change periodically. The current pattern to match would be indicated on a virtual screen shown within the participant's in-game point of view. Scores were calculated using the game-generated score plus the time remaining from each round, then averaged across both rounds, and transformed into a percentage. Scores ranged from .08 to .91 ($M$ = .71, $SD$ = .13). A screenshot of this game is provided in Figure 5.

***Visual Pursuit (VR Assessment).*** The game *Super Amazeballs* was used to evaluate VP. In this game, participants were shown a 3D track fixed within a translucent orb that they could rotate, spin, and manipulate. The objective of the game was to rotate and balance the orb in order to direct and navigate the ball along the track, so that the ball would eventually reach the end of the course without falling over the ledges on either side of the track. Participants would progress through a series of tracks that became increasingly difficult at each level. If the ball fell off the track, participants would have to start over from the beginning or most recent checkpoint of that level. Each track required the participant to visually pursue the progression of the track. The participant was limited

to five minutes to play this game and instructed to progress through as many levels (i.e., tracks) as possible within the allotted time while preventing the ball from falling off the track. Scores were calculated as the number of balls fallen added to one, then divided by thirteen (to prevent negative scores), subtracted from the number of levels completed, and transformed into a percentage. Scores ranged from .00 to .96 ($M$ = .53, $SD$ = .22). A screenshot of this game is provided in Figure 6.

**Measures.** The following measures were obtained through a self-report survey administered at the end of the study. Excluding demographic and academic achievement measures, all of the following measures were evaluated on a 5-point Likert-type scale.

***Video game experience.*** Video game experience (VGE) refers to video game players' competence with video games based on their prior experience interacting with video games (Sanchez & Langer, 2018). The VGE scale was used in this study, including 19 items ($\alpha$ = .94). Participants were instructed to rate their agreement with each item. Subscales included *game enjoyment* ($\alpha$ = .89), *intentional game play* ($\alpha$ = .90), *game self-efficacy* ($\alpha$ = .91), and *game flow* ($\alpha$ = .67). For the *game enjoyment* subscale, a sample item includes, "I enjoy playing video games." For the *intentional game play* subscale, a sample item includes, "I spend many hours each week playing video games." For the *game self-efficacy* subscale, a sample item includes, "I am good at video games, compared to others." For the *game flow* subscale, a sample item includes, "I lose track of time when I play video games." See Appendix 1 for the full VGE scale.

***Intimidation with games.*** Intimidation with games (IWG) refers to the extent to which an individual finds it difficult to play and achieve competence in video games (Sanchez & Langer, 2018). The IWG scale was used in this study, including 6 items ($\alpha$ = .88). Participants were instructed to rate their agreement with each item. A sample item includes, "Video games are intimidating to me." See Appendix 1 for the full IWG scale.

***Personality.*** Personality was measured using the International Personality Item Pool (IPIP-NEO) version of the Revised NEO Personality Inventory (NEO PI-R). This scale measures the five-factor model of personality: *openness* ($\alpha$ = .84), *conscientiousness* ($\alpha$ = .83), *extraversion* ($\alpha$ = .88), *agreeableness* ($\alpha$ = .79), and *neuroticism* ($\alpha$ = .87; Maples, Guan, Carter, & Miller, 2014). The shortened form of the IPIP-NEO (IPIP-120) was used in this study, featuring 120 items ($\alpha$ = .88). Participants were instructed to rate the degree to which each item describes them. For the *openness* subscale, a sample item includes, "Have a vivid imagination." For the *conscientiousness* subscale, a sample item includes, "Complete tasks successfully." For the *extraversion* subscale, a sample item includes, "Make friends easily." For the *agreeableness* subscale, a sample item includes, "Trust others." For the *neuroticism* subscale, a sample item includes, "Worry about things." See Appendix 2 for the full IPIP-120 scale.

***Academic achievement.*** In order to measure academic achievement, a grade point average (GPA) was obtained from each participant, serving as the performance-related criterion for this study. GPAs were calculated on a scale from 0.0 to 4.0 ($M$ = 3.43, $SD$ = 0.56), averaging the results of all graded coursework completed at the university in which

participants were enrolled at the time of participation. GPAs were obtained using participants' student identification numbers, which were temporarily collected through the demographic questionnaire. See Appendix 3 for the full demographic questionnaire.

***Demographics.*** Demographic measures were gathered at the end of the study. These measures included age, gender, and ethnicity. See Table 1 for a full distribution of demographic results, and see Appendix 3 for the full demographic questionnaire.

## Results

Prior to analyzing study data, evidence for the content validity of the VR assessments was obtained through subject matter expert endorsement. To obtain this evidence, content from the VR assessments for Space Visualization (SV), Visual Speed & Accuracy (VSA) and Visual Pursuit (VP) was presented alongside the content and construct definitions from the self-report assessments to a doctoral-level expert in and professor of Psychology. This subject matter expert evaluated the VR assessment content and affirmed that each VR assessment seemed to represent the content targeted by its corresponding self-report assessment.

Next, an independent samples t-test was conducted to determine whether the order of the assessments (i.e., VR assessments first, self-report assessments second; or vice versa) affected the outcomes. Participants were divided by the order in which they took the assessments, and scores were compared for both the VR and self-report measures (i.e., SV, VSA, and VP). None of the resulting $t$ statistics were significant ($p > .05$), indicating that the order of the assessments (i.e., whether a given participant played the

VR games first or completed the self-report assessment first) did not significantly influence the outcome scores.

Prior to testing study hypotheses, we conducted a correlational analysis of study variables; see Table 2 for all correlations. To highlight results, age was found to be significantly related to conscientiousness ($r = .38, p < .001$), agreeableness ($r = .18, p = .04$), neuroticism ($r = -.42, p < .001$), and GPA ($r = .20, p = .02$). However, these correlations may be spurious due to range restriction in the primarily college-aged sample used for this study. Gender was analyzed as a dichotomous variable (0 = female, 1 = male) and found to be significantly correlated with agreeableness ($r = -.22, p = .02$), neuroticism ($r = -.20, p = .03$), video game experience (VGE; $r = .40$, $p < .001$), and intimidation with games (IWG; $r = -.39, p < .001$). The relationships between gender with VGE and IWG were expected due to previous research finding significantly higher rates of video game use by males than by females (Greenberg, Sherry, Lachlan, Lucas, & Holmstrom, 2010). Next, openness was found to be significantly correlated with both VGE ($r = .22, p = .01$) and IWG ($r = -.26, p = .004$), which were also significantly correlated with each other ($r = -.57, p < .001$). It makes sense that openness would be positively related to VGE and negatively related to IWG, seeing as the construct of openness as a personality trait includes experience-seeking behaviors that can occur through playing video games, as well as the tendency to approach rather than avoid new experiences. It was also expected that VGE and IWG would be strongly and negatively

related, seeing as both constructs approach familiarity and skill with video games, but VGE targets greater levels of familiarity and skill while IWG targets more limited levels.

Finally, relationships were demonstrated between assessments both within and across formats. Within the VR assessment format, VSA was found to be significantly correlated with both SV ($r = .34, p < .001$) and VP ($r = .45, p = <.001$), while SV and VP were not significantly correlated with one another ($r = .13, p = .17$). Across self-report and VR formats, significant correlations were demonstrated for VSA ($r = .32, p < .001$) and VP ($r = .23, p = .01$), but not for SV ($r = .16, p = .07$). These relationships provide some initial evidence that behaviors measured through the VR assessments are related to those measured through the self-report assessments. However, although the significance of these relationships varies across assessments, the behaviors across all three VR assessments seem to relate to one another despite being developed to measure three distinct abilities.

**Hypothesis Testing**

In order to test our hypotheses for this study, we used hierarchical regression analyses to evaluate the unique variance in outcomes accounted for by each unique predictor. For each hierarchical regression, we entered VGE and IWG scores in Step 1, then we entered the relevant predictor variable in Step 2. This allowed us to statistically account for the variance in outcomes explained by VGE and IWG, which were shown to relate to certain demographic variables. Previous research has also supported the link between demographics such as gender with performance in video games (Brown, Hall,

Holtzer, Brown, & Brown, 1997). Using this strategy, we were able to control for the effects of VGE and IWG in our analyses.

**Convergent validity.** Hypothesis 1 was intended to demonstrate convergent validity between the VR and self-report formats as assessments for cognitive ability. We predicted that scores from each VR assessment would account for a significant portion of the variance in scores from the respective self-report cognitive ability assessment. To test H1, we conducted a hierarchical regression using VGE and IWG as control variables in Step 1 of each model and VR assessment scores as Step 2, with self-report assessment scores as the dependent variable. As shown in Table 3, we found that VR assessment scores accounted for a significant portion of the variance in self-report assessment scores for VSA, $F(3,118) = 5.74$, $p = .001$, $\Delta R^2 = .95$, and for VP, $F(3,118) = 3.15$, $p = .02$, $\Delta R^2 = .04$. However, VR assessment scores for SV did not account for a significant portion of variance in self-report assessment scores, $F(3, 118) = 1.35$, $p = .12$, $\Delta R^2 = .02$. These results provide partial support for H1, showing that scores from the VR assessments generally accounted for a significant portion of the variance in scores from the respective self-report assessments.

**Divergent validity.** Hypothesis 2 was intended to demonstrate divergent validity of the VR assessment as unrelated to measures of personality. We predicted that scores from each VR assessment would not account for a significant portion of variance in each of the Big Five personality traits, i.e., openness, conscientiousness, extraversion, agreeableness, and neuroticism. To test H2, we conducted five hierarchical regressions

for each assessment, one for each personality trait, using VGE and IWG as control variables in Step 1 of each model and VR assessment scores as Step 2, with each respective personality trait as the dependent variable. Due to the high number of comparisons made to investigate this particular hypothesis (i.e., five personality traits across three assessments, totaling in fifteen comparisons), we used the Bonferroni correction to adjust the significance level for all comparisons, thereby correcting for alpha inflation and mitigating risk of type I error. The corrected significance level for the analyses conducted to test H2 was set at $\alpha = .003$.

***Space Visualization.*** As shown in Table 4, VR assessment scores for SV did not significantly predict openness, $F(3,118) = 3.3$, $p = .22$, $\Delta R^2 = .01$; conscientiousness, $F(3, 118) = 1.68$, $p = .43$, $\Delta R^2 = .01$; extraversion, $F(3, 118) = 0.74$, $p = .57$, $\Delta R^2 = <.001$; agreeableness, $F(3, 118) = 0.85$, $p = .72$, $\Delta R^2 = <.001$; or neuroticism, $F(3, 118) = 2.45$, $p = .90$, $\Delta R^2 = <.001$.

***Visual Speed & Accuracy.*** As shown in Table 5, VR assessment scores for VSA did not significantly predict openness, $F(3,118) = 5.69$, $p = .004$, $\Delta R^2 = .06$; conscientiousness, $F(3,118) = 1.83$, $p = .15$, $\Delta R^2 = .02$; extraversion, $F(3, 118) = 0.76$, $p = .52$, $\Delta R^2 = <.001$; agreeableness, $F(3,118) = 1.03$, $p = .46$, $\Delta R^2 = .01$; or neuroticism, $F(3,118) = 2.36$, $p = .28$, $\Delta R^2 = .01$.

***Visual Pursuit.*** As shown in Table 6, we found that VR assessment scores for VP did not significantly predict openness, $F(3, 118) = 3.16$, $p = .31$, $\Delta R^2 = .01$; conscientiousness, $F(3,118) = 1.54$, $p = .42$, $\Delta R^2 = .01$; extraversion, $F(3,118) = 0.74$, $p =$

.63, $\Delta R^2$ = <.001; agreeableness, $F(3,118) = 1.30$, $p = .25$, $\Delta R^2 = .01$; or neuroticism, $F(3,118) = 3.16$, $p = .09$, $\Delta R^2 = .02$.

These results provide full support for H2, with scores from all three VR assessments demonstrating non-significant relationships with openness, conscientiousness, extraversion, agreeableness, and neuroticism.

**Criterion-related validity.** For hypothesis 3, we intended to demonstrate that scores from the VR assessments could provide a meaningful prediction of a relevant variable (i.e., academic achievement). We predicted that scores from each VR assessment would account for a significant portion of variance in academic achievement indicated by GPA. To test H3, we conducted a hierarchical regression for each assessment, using VGE and IWG as control variables in Step 1 of each model and VR assessment scores as Step 2, with GPA as the dependent variable. As shown in Table 7, we found that VR assessment scores for VSA accounted for a significant portion of variance in GPA, $F(3,117) = 3.96$, $p = .01$, $\Delta R^2 = .06$. However, both SV, $F(3,117) = 2.30$, $p = .13$, $\Delta R^2 = .02$, and VP, $F(3,117) = 1.95$, $p = .26$, $\Delta R^2 = .01$, did not significantly predict GPA. These results provide partial support for H3, with only one VR assessment, intended to measure participants' speed and accuracy in pattern recognition, predicting academic achievement.

**Exploratory Analyses**

The general purpose of our exploratory analyses was to examine two research questions that were respectively too small or too broad to serve as study hypotheses.

These questions explore two additional emergent properties of VR game-based assessment: incremental predictive validity, and adverse impact.

**Research Question 1.** Research question 1 asked whether VR assessment scores would demonstrate incremental validity in predicting academic achievement over the self-report format scores. As shown in Table 8, we found that VR assessment scores accounted for a significant portion of the variance in academic achievement, measured using GPA, beyond the variance explained by the self-report assessment scores, but only for VSA, $F(3,117) = 3.56$, $p = .04$, $\Delta R^2 = .04$, and not for SV, $F(3,117) = 2.94$, $p = .22$, $\Delta R^2 = .01$, or VP, $F(3,117) = 1.68$, $p = .37$, $\Delta R^2 = .01$. When predicting academic achievement beyond what is explained using a traditional assessment of cognitive ability alone, we were again only able to find support for the speed and accuracy of pattern recognition as having incremental validity for assessment in a VR format.

**Research Question 2.** Research question 2 asked whether the self-report and VR assessment formats would produce adverse impact in terms of significant differences in scores using ethnic and gender-based comparisons. Calculating adverse impact requires designating a cut score to discriminate between *passing* and *non-passing* participants. For objectivity purposes, prior to data analyses, we predetermined to evaluate adverse impact at the 50th, 75th, and 90th percentiles for each assessment score based on the distribution of each score in the current sample. This analysis considered adverse impact between the available gender subgroups (i.e., minority female, $N = 88$, and majority male, $N = 34$) and ethnic subgroups (i.e., minority Asian, $N = 31$, or minority Latino, $N = 35$, and majority

White, $N = 35$), which were dummy coded as *majority* or *minority* for each comparison. Adverse impact analyses excluded participants who identified within the "Other" ethnic subgroup category, including "Black or African American", "Native American or Alaska Native", or any combination of multiple ethnic subgroups, due to insufficient sample size ($N = 23$) for yielding the statistical power necessary to make any meaningful comparison. Participants who identified as "Other" (i.e., "Transgender Female"; "Transgender Male"; Genderqueer, Genderfluid, or Non-Binary"; "Intersex"; or "Other"; $N = 2$) for their gender subgroup category were excluded for the same reason. Based on recommended methods for testing adverse impact, we used the following calculations: four-fifths ratio test, $\chi^2$ test, $Z$ test, and Fisher's exact probability test (see Table 9; Tippins, 2010). General findings across all calculations show a pattern of adverse impact in both cognitive ability assessment formats, self-report and VR.

***Four-Fifths Ratio Test.*** Indications of adverse impact were consistent for the four-fifths ratio testing, where patterns typically favored male participants over female participants and White participants over Latino participants. However, there was a consistent pattern favoring Asian participants over White participants.

***Chi-Squared Test.*** Based on $\chi^2$ statistics, patterns of adverse impact favored male participants over female participants for nearly all assessments, but not across all percentile cutoffs. Patterns also favored Asian participants over White participants for fewer assessments, while no adverse impact was demonstrated between White participants and Latino participants.

***Z-test.*** Because Z-test statistics were computed as transformations of $\chi^2$ statistics, identical patterns of adverse impact were demonstrated through both tests. That is, based on *Z* statistics, patterns of adverse impact favorited male participants over female participants for most assessments, Asian participants over White participants for some assessments, and White participants over Latino participants for no assessments.

***Fischer's Exact Probability.*** Based on *p* values from Fisher's exact probability test, identical patterns of adverse impact were yet again observed as seen through both the chi-squared test and Z-test results. That is, based on p values from Fisher's exact probability test, patterns of adverse impact favorited male participants over female participants for most assessments, Asian participants over White participants for some assessments, and White participants over Latino participants for no assessments.

Based on results from the four-fifths ratio tests, chi-square tests, *Z*-tests, and Fisher's exact probability tests, adverse impact was evidenced across all assessments, using all tested percentile cutoff values, and across all demographic subgroup comparisons. However, no differences were seen across formats, indicating that adverse impact was equally prevalent in self-report assessment scores as in VR assessment scores.

## Discussion

With VR technology increasing in popularity outside of purely entertainment-oriented functions, further evidence is needed to support and guide the application of VR games in these contexts. Therefore, the purpose of this study is to explore evidence for

the validity of three VR games as assessments of three respective cognitive abilities (i.e., Space Visualization, Visual Speed & Accuracy, and Visual Pursuit). We evaluated different types of validity to gather evidence regarding the application of these VR games as cognitive ability assessments. This evidence included convergent validity (i.e., significant relationships between the VR format and the self-report format of the same cognitive ability assessments), divergent validity (i.e., non-significant relationships between the VR assessment scores and OCEAN personality traits), and criterion-related validity (i.e., significant relationships between the VR assessment scores and academic achievement indicated by participants' GPAs). We also conducted exploratory analyses to address questions on the incremental validity and potential for adverse impact through the VR assessment format.

In hypothesis 1, we predicted that the similarities between the game content and traditional test content would be similar enough to show a significant relationship between the VR assessment format and the self-report assessment format when measuring the three assessments of specific cognitive abilities. We found evidence to support this assumption for two of the three assessments. It is important to consider which aspects of these games allowed them to perform better as assessments. First, consider the two assessments that demonstrated significant convergence between the VR and self-report formats. The first was Visual Speed & Accuracy, in which participants needed to distinguish between and match patterns. Similarities between the VR and traditional self-report assessments included pattern recognition and distinction

components. However, the VR format required additional skills, such as balance for standing and shifting weight, and manual dexterity for using controls to manipulate the orientation of the patterned cubes in the game.

The second of the two VR assessments that demonstrated convergence with its respective self-report assessment was Visual Pursuit. Similarities across formats for this specific ability assessment emphasized visual tracking and concentration components. However, in addition to mixing these components with the requirements for navigating VR environments such as balance and manual dexterity, the VR assessment for Visual Pursuit also involved sustained concentration, where any break could result in an error and subsequent checkpoint reset. It is important to consider that these behavioral differences may have contributed to differences in results between assessment formats, with gameplay in VR environments potentially placing additional confounding demands on participants beyond the self-report assessments (Plass, Homer, Kinzer, Frye, & Perlin, 2011). Most importantly, it should be underscored that two of the VR measures still converged significantly with their self-report counterparts.

On the other hand, one of the three VR measures, i.e., Space Visualization, did not converge significantly with its corresponding self-report measure. Beyond the potential confounding effects of video gameplay in assessment contexts as described above, the research design of this study may have contributed to the limited findings for this assessment's convergence across formats. Specifically, this was the first VR assessment that each participant completed. Considering the likely novelty of the VR

environment to many participants, this ordering may have yielded practice effects that significantly affected performance on the Space Visualization assessment, such that results may reflect participants' real-time VR gaming skill acquisition in addition to measuring their Space Visualization behaviors. These practice effects are consistently supported by prior research, particularly with spatial tests such as this one (Benedict & Zgaljardic, 1998). For this reason, limited results for Space Visualization should be anticipated across additional hypotheses, as seen below in the discussion of the remaining hypotheses and exploratory analyses.

Even so, the partial support found for hypothesis 1 is encouraging because it demonstrates that there are meaningful similarities between the two assessment formats that warrant further exploration. These findings build upon previous research-based support for the convergent validity of assessments in 2D game-based formats by extending this support to a VR game-based format (McPherson & Burns, 2008; Thompson, Barrett, Patterson, & Craig, 2012; Ventura & Shute, 2013).

In hypothesis 2, we predicted that none of the three VR assessments would demonstrate significant relationships with any of the five OCEAN personality traits, i.e., openness, conscientiousness, extraversion, agreeableness, and neuroticism. Because previous research has demonstrated orthogonal relationships between cognitive ability and personality (Barrick & Mount, 1991), we hoped to demonstrate similarly orthogonal relationships using the VR assessments as measures of cognitive ability. Accordingly, we found that all three VR assessments were non-significantly related to each of the five

OCEAN personality traits. While no significant relationships were found relative to the significance level established using the Bonferroni correction ($\alpha = .003$), it should be noted that this level was closely approximated by the significance of the relationship between Visual Speed & Accuracy and openness ($p = .004$). Although non-significant, this relationship might be explained by certain qualities of the game that was used to assess Visual Speed & Accuracy. For example, the bright colors, vibrant shapes, and other highly experiential imagery and actions featured in this game might relate to the similarly experiential elements of openness as a construct (McCrae & Costa, 1987). Additionally, while cognitive ability has typically demonstrated orthogonality with personality traits, recent research has provided some support for a potential link between cognitive ability and openness (Carretta & Ree, 2018).

Nevertheless, hypothesis 2 was supported by the non-significant relationships between all VR assessments and OCEAN personality traits. In addition to the convergence for Visual Speed & Accuracy and Space Visualization demonstrated through testing hypothesis 1, this divergence contributes to broader evidence for the construct validity of the VR assessments, showing that they conform to the constructs of their respective specific abilities (i.e., Space Visualization, Visual Speed & Accuracy, and Visual Pursuit), not by only measuring what they are intended to measure, but also by not measuring what they are not intended to measure.

In hypothesis 3, we predicted that the three VR assessments would each demonstrate predictive relationships with academic performance, such that the VR

assessment scores as distinct measures of specific cognitive abilities would account for significant portions of variance in GPA as a measure of academic performance. Evidence supported the criterion-related validity of the Visual Speed & Accuracy assessment, indicating that this assessment was shown to function as a significant predictor of academic performance. However, neither of the other two VR assessments, i.e., Space Visualization and Visual Pursuit, were significantly related with academic performance in the same way. The findings for this hypothesis were further supported by the results of the first exploratory analysis, which explored whether VR assessment scores could be used to explain a significantly increased portion of variance in GPA above the variance accounted for by self-report scores. Again, findings from the first exploratory analysis provided validity evidence for only the Visual Speed & Accuracy VR assessment.

As mentioned above, there may be certain properties unique to the game selected for the VR assessment of Visual Speed & Accuracy that yielded greater evidence for its validity, in this case, in the predication (or incremental prediction) of academic performance. Examples of such properties could pertain to features like the specific images encountered, objectives sought, actions performed, or feedback received by participants in this game. Regardless, these findings for hypothesis 3 and for the first exploratory analysis provide evidence for the criterion-related validity of the VR assessment format in predicting academic performance. This evidence is consistent with previous literature supporting the link between measures of cognitive ability with performance, indicating that VR assessments could be used to predict performance in a

similar manner as 2D game-based assessments and even self-report assessments (Bottino, Ott, Tavella, & Benigno, 2010; DeRosier & Thomas, 2018; Kiili & Ketamo, 2018).

In order to expand upon this evidence for the construct (i.e., convergent and divergent) and criterion-related (i.e., predictive and incremental) validity of the VR assessments, we focused our second and final exploratory analysis toward examining the utility of these assessments in applied settings. In particular, this analysis explored the emergence of adverse impact in assessment scores, evaluated through comparisons between male versus female, White versus Asian, and White versus Latino participants' scores. We found mixed results across the various practical and statistical tests that were run to evaluate adverse impact, but evidence for adverse impact was generally widespread for all VR and self-report assessments, typically favoring male over female participants, Asian over White participants, and White over Latino participants.

The findings for this exploratory analysis align with some previous research-based findings, such as the tendency of cognitive ability tests to favor Asian assesses over White assesses and White assesses over Latino assesses (Ployhart & Holtz, 2008). However, this analysis also found extensive adverse impact across gender subgroups, whereas the literature on adverse impact through cognitive ability testing indicates that gender differences are small (Hyde, 1981). Because adverse impact was found across both the VR and self-report assessment formats, the mixed support that these findings provide for prior research may be due to the highly visual and spatial nature of both of

these assessments compared with other tests of cognitive ability, rather than due to their formats alone.

For comparison, the use of non-verbal visual components in cognitive ability testing has been shown to lead to reductions in ethnic adverse impact by way of reducing additional cognitive load from reading and other confounds (Outtz, 2002). Also, while cognitive ability as a general mental ability (i.e., *g*) does not differ greatly between genders, there are differences for specific abilities, with female assesses typically outperforming males in verbal reasoning assessments but underperforming in spatial reasoning assessments (Ones, Dilchert, Viswesvaran, & Salgado, 2017). It makes sense that the focus upon graphics, rather than text, in both the VR and self-report formats used in this study would yield similarly lower than expected rates of adverse impact comparing White and Latino participants' scores and higher than expected rates comparing male and female participants' scores. Regardless, these findings do not support the use of a VR game-based assessment format to reduce adverse impact compared with a traditional self-report assessment.

**Future Directions**

Overall, the results from this study provide mixed but promising support for the validity of the VR assessments in assessing their respective specific cognitive abilities. This support extends to the convergent validity of the Visual Speed & Accuracy and Visual Pursuit assessments; the divergent validity of all VR assessments, compared for orthogonality with OCEAN personality traits; the criterion-related validity of the Visual

Speed & Accuracy assessment in predicting GPA; and the incremental validity of the Visual Speed & Accuracy VR assessment in predicting GPA over the corresponding self-report assessment. We also found evidence for adverse impact through all VR and self-report assessment scores, with greater evidence seen when comparing genders than when comparing ethnicities. Taken together, these results build upon past research demonstrating the validity and utility of 2D game-based assessments in measuring constructs associated with cognitive ability by providing evidence for the validity of a similar cognitive ability measurement in a VR game-based format (Hummel, Brinke, Nadolski, & Baartman, 2017; Kiili, Devlin, Perttula, Tuomi, & Lindstedt, 2015; Shute, Ventura, & Kim, 2013). This might be used to support future research exploring the validity and utility of VR technology to assess cognitive ability.

For instance, future research should examine other specific abilities in the VR assessment format, i.e., beyond Space Visualization, Visual Speed & Accuracy, and Visual Pursuit. This study primarily found support for the validity of the VR assessment format for measuring Visual Speed & Accuracy, but not Space Visualization or Visual Pursuit. Therefore, other studies could explore validity evidence for VR assessments in measuring constructs more similar to Visual Speed & Accuracy than to Space Visualization or Visual Pursuit, or otherwise measuring constructs associated with cognitive ability but entirely distinct from Space Visualization and Visual Pursuit. Such research should be conducted in order to further establish links between VR game-based assessments and specific cognitive abilities.

Future research should also explore other forms of validity evidence demonstrated by the VR assessment format, i.e., beyond convergent, divergent, criterion-related, and incremental validity. In line with the prevailing view of validity as a unitary concept rather than a shortlist of common sources for evidence (such as the outdated tripartite model of validity), forms of validity beyond content, construct, and criterion-related validity should be taken into consideration. For instance, test validity has been explored in recent research through methods such as nomological networks and multitrait-multimethod matrices (Engellant, Holland, & Piper, 2016; Miller et al., 2018). Depending on the specific constructs being assessed, these methods might be appropriate for evaluating validity through similar research on VR assessments.

Additionally, future research should emphasize participants' experiences with and reactions to VR assessments, as well as relevant organizational outcomes. The current literature supports the effects of certain 2D game-based assessments in increasing positive reactions among job candidates (Armstrong, Ferrell, Collmus, & Landers, 2016). It should be determined whether these effects will translate to VR assessment formats. These outcomes are particularly important for organizations and practitioners relying upon the use of assessments for employee selection, considering the potential impact of candidate reactions upon factors like organizational attractiveness as well as selected employees' intent to accept employment offers based on experiences throughout the selection process (Highhouse, Lievens, & Sinar, 2003).

**Limitations**

Furthermore, future research should address the limitations of this study. A major limitation is the use of commercial off-the-shelf games rather than games that were custom developed or otherwise specialized to assess the targeted constructs. This may have affected assessment scores, seeing as the scoring for some of the VR assessments utilized in this study relied upon manual behavioral coding rather than cleaner, more objective scores that could have been produced automatically within the VR games. For instance, scores for the Visual Pursuit assessment were calculated based on a criticality index combining different counts of in-game behaviors, rather than scores generated by the game, and this scoring system may be related to the lack of positive findings to support the criterion-related or incremental validity of this assessment. Also, the games chosen for this study were originally developed and published exclusively for entertainment purposes, so it is likely that the match between the constructs targeted by this study could be better assessed through gameplay in VR games that were specifically designed to invoke behaviors that are representative of these constructs. Thus, the lack of intentionality in game design, with respect to the measurement of specific cognitive abilities, likely entailed the inclusion of unrelated behaviors that may have confounded the measurement of these abilities.

Another limitation of this study is the use of a sample composed of participants who are disproportionately White, educated, industrialized, rich, and Democratic, with respect to the global population, otherwise known as a WEIRD sample. This limitation is common to studies such as the current one, which relied upon a convenience sample of

undergraduate college students, and this most likely impacted results (Henrich, Heine, & Norenzayan, 2010). The limitations involved with this sample also included the need to use GPA as a proxy measure for job performance, seeing as many participants may not have possessed significant job performance history to allow for this measure to serve as a targeted criterion of interest. The use of GPA in this context is supported by use in previous research, but as a proxy, this measure is still an imperfect representation of the targeted measure of job performance (Imose & Barber, 2015). As a result, the findings of this study may be limited in generalizability from the sampled undergraduate student population to a wider workforce population, in turn limiting the ecological validity of these findings.

Finally, certain aspects of the research design may have affected study outcomes, potentially limiting the extent to which findings supported hypotheses. For instance, while participants completed self-report assessments and VR assessments as two distinct assessment batteries that were delivered in counterbalanced orders, the order of assessments within these batteries was not counterbalanced. Participants completed the VR assessments in a fixed order (i.e., Space Visualization, Visual Speed & Accuracy, Visual Pursuit). This ordering may have affected how participants performed on the individual assessments. Taking into consideration the novelty of the VR technology that was featured in the study and the prior unfamiliarity that many participants may have possessed with this technology, it is possible that results may be confounded by practice effects, with participants increasing in successful in-game behaviors as they progressed,

and/or fatigue effects, with participants decreasing in successful in-game behaviors as they tired from the unfamiliar and relatively lengthy experience.

**Conclusion**

This study is intended to provide evidence for the validity of a set of three VR game-based assessments in measuring three respective constructs associated with cognitive ability, i.e., Space Visualization, Visual Speed & Accuracy, and Visual Pursuit. Findings provide some validity evidence for these VR assessments, especially Visual Speed & Accuracy, as evaluated through relationships demonstrated with self-report assessment scores, orthogonality with OCEAN personality traits, and prediction of GPA. In addition, findings provide evidence for the incremental validity of VR assessment scores for Visual Speed & Accuracy over self-report assessment scores to predict GPA, as well as demonstrating moderate adverse impact in gender and ethnic subgroup comparisons across both assessment formats, i.e., VR and self-report. Future research should continue to explore the VR game-based assessment format in measuring additional specific cognitive abilities, supported by additional forms of validity evidence, and using specialized or custom-made VR games, so that more robust support for validity might be anticipated. Also, future studies should explore VR assessment among applied samples, rather than student samples, in order to increase the ecological validity of findings supporting the use of VR game-based assessments in workplace contexts, and to explore additional outcomes such as reactions. Regardless, this study constitutes a

valuable early step forward to help drive research and practice aligned with the proactive rather than reactive incorporation of novel and useful technology into personnel practices.

# References

Aggarwal, R., Ward, J., Balasundaram, I., Sains, P., Athanasiou, T., Darzai, A. (2007). Proving the effectiveness of virtual reality simulation for training in laparoscopic surgery. *Annals of Surgery, 246*(5), 771-779. doi:10.1097/SLA.0b013e3180f61b09

Aïm, F., Lonjon, G., Hannouche, D., & Nizard, R. (2016). Effectiveness of virtual reality training in orthopaedic surgery. *Arthroscopy: The Journal of Arthroscopic & Related Surgery, 31*(1), 224-232. https://doi.org/10.1016/j.arthro.2015.07.023

American Educational Research Association (AERA), American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association,

Aminov, A., Rogers, J. M., Middleton, S., Caeyenberghs, K., & Wilson, P. H. (2018). What do randomized controlled trials say about virtual rehabilitation in stroke? A systematic literature review and meta-analysis of upper-limb and cognitive outcomes. *Journal of NeuroEngineering and Rehabilitation, 15*(29), 2-24. https://doi.org/10.1186/s12984-018-0370-2

Armstrong, M. B., Ferrell, J. Z., Collmus, A. B., & Landers, R. N. (2016). Correcting misconceptions about gamification of assessment: More than SJTs and badges. *Industrial and Organizational Psychology, 9*(3), 671-677. https://doi.org/10.1017/iop.2016.69

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*(1), 1-26. https://doi.org/10.1111/j.1744-6570.1991.tb00688.x

Benedict, R. H. B., & Zgaljardic, D. J. (1998). Practice effects during repeated administrations of memory tests with and without alternate forms. *Journal of Clinical and Experimental Neuropsychology, 20*(3), 339-352. https://doi.org/10.1076/jcen.20.3.339.822

Bhatia, S., & Ryan, A. M. (2018). Hiring for the win: Game-based assessment in employee selection. In J. H. Dulebohn & D. L. Stone (Eds.), *The brave new world of eHRM 2.0* (pp. 81-110). Charlotte, NC: Information Age Publishing, Inc.

Blackman, M. C. (2002). The employment interview via the telephone: Are we sacrificing accurate personality judgments for cost efficiency? *Journal of Research in Personality, 36*(3), 208-223. doi:10.1006/jrpe.2001.2347

Bottino, R. M., Ott, M., Tavella, M., & Benigno, V. (2010, October). *Can digital mind games be used to investigate children's reasoning abilities?* Paper presented at the 4th annual meeting of the European Conference on Games Based Learning, Copenhagen, Denmark.

Brown, R. M., Hall, L. R., Holtzer, R. Brown, S. L., & Brown, N. L. (1997). Gender and video game performance. *Sex Roles, 36*(11-12), 793-812.

Brunner, I., Skouen, J. S., Hofstad, H., Strand, L. I., Becker, F., Sanders, A.-M., Pallesen, H., Kristensen, T., Michielsen, M., & Verheyden, G. (2014). Virtual reality

training for upper extremity in subacute stroke (VIRTUES): Study protocol for a randomized controlled multicenter trial. *BMC Neurology, 14*(186), 1-5. doi:10.1186/s12883-014-0186-z

Cameirão, M. S., Bermúdez i Badia, S., Duarte Oller, E., & Verschure, P. F. M. J. (2010). Neurorehabilitation using the virtual reality based Rehabilitation Gaming System: Methodology, design, psychometrics, usability and validation. *Journal of NeuroEngineering and Rehabilitation, 7*(48), 1-14. doi:10.1186/1743-0003-7-48

Carretta, T. R., & Ree, M. J. (2018). The relations between cognitive ability and personality: Convergent results across measures. *International Journal of Selection and Assessment, 26*(2-4), 133-144. doi:10.1111/ijsa.12224

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York, NY: Cambridge University Press.

Chamorro-Premuzic, T., Winsborough, D., Sherman, R. A., & Hogan, R. (2016) New talent signals: Shiny new objects or a brave new world? *Industrial and Organizational Psychology, 9*(3), 621-640. doi:10.1017/iop.2016.6

Cortina, J. M., Goldstein, N. B., Payne, S. C., Davison, H. K., & Gilliland, S. W. (2000). The incremental validity of interview scores over and above cognitive ability and Conscientiousness scores. *Personnel Psychology, 53*(2),325-351. https://doi.org/10.1111/j.1744-6570.2000.tb00204.x

Cotton, S. J., Dollard, M. F., & de Jonge, J. (2002). Stress and student job design: Satisfaction, well-being, and performance in university students. *International*

*Journal of Stress Management, 9*(3), 147-162. https://doi.org/10.1023/A:1015515714410

DeRosier, M. E., & Thomas, J. M. (2018). Establishing the criterion validity of Zoo U's game-based social emotional skills assessment for school-based outcomes. *Journal of Applied Development Psychology, 55*, 52-61. https://doi.org/10.1016/j.appdev.2017.03.001

Engellant, K. A., Holland, D. D., & Piper, R. T. (2016). Assessing convergent and discriminant validity of the motivation construct for the technology integration education (TIE) model. *Journal of Higher Education Theory and Practice, 16*(1), 37-50. https://doi.org/10.1080/15391523.2016.1172448

Engler, C. E. (1992). Affordable VR by 1994. *Computer Gaming World, 100*, 80-81. Retrieved from http://www.cgwmuseum.org/galleries/index.php?year=1992&pub=2&id=100

Equal Employment Opportunity Commission (EEOC). (1978). *Uniform guidelines on employee selection procedures*. Washington, DC: Equal Employment Opportunity Commission.

Frasca, G. (1999). Ludology meets narratology: Similitude and differences between (video)games and narrative. *Parnasso, 3*, 365-371. Retrieved from http://www.ludology.org/articles/ludology.htm

Freeman, D., Reeve, S., Robinson, A., Ehlers, A., Clark, D., Spanlang, B., & Slater, M. (2017). Virtual reality in the assessment, understanding, and treatment of mental

health disorders. *Psychological Medicine, 47*, 2393-2400. doi:10.1017/S003329171700040X

Freina, L., & Ott, M. (2015). *A literature review on immersive virtual reality in education: State of the art and perspectives.* Paper presented at the 7th annual meeting of the Annual International Scientific Conference on eLearning and Software for Education (eLSE), Bucharest, Romania.

Gavish, N., Gutiérrez, T., Webel, S., Rodríguez, J., Peveri, M., Bockholt, U., & Tecchia, F. (2015). Evaluating virtual reality and augmented reality training for industrial maintenance and assembly tasks. *Interactive Learning Environments, 23*(6), 778-798. https://doi.org/10.1080/10494820.2013.815221

Gigante, M. A. (1993). Virtual Reality: Definitions, history and applications. In R. A. Earnshaw, M. A. Gigante, & H. Jones (Eds.), *Virtual reality systems* (pp. 3-14). London, UK: Academic Press Limited.

Greenberg, B. S., Sherry, J., Lachlan, K., Lucas, K., & Holmstrom, A. (2010). Orientations to video games among gender and age groups. *Simulation & Gaming, 41*(2), 238-259. doi:10.1177/1046878108319930

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history and bibliography. *Intelligence, 24*, 13-23. https://doi.org/10.1016/S0160-2896(97)90011-8

Hays, R. T., Jacobs, J. W., Prince, C., & Salas, E. (1992). Flight simulator training effectiveness: A meta-analysis. *Military Psychology, 4*(2), 63-74. https://doi.org/10.1207/s15327876mp0402_1

Helms, J. E. (2006). Fairness is not validity or cultural bias in racial-group assessment: A quantitative perspective. *American Psychologist, 61*(8), 845-859. doi:10.1037/0003-066X.61.8.859

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Beyond WEIRD: Towards a broad-based behavioral science. *Behavioral and Brain Sciences, 33*(2-3), 111-135. https://doi.org/10.1017/S0140525X10000725

Highhouse, S., Lievens, F., & Sinar, E. F. (2003). Measuring attraction to organizations. *Education and Psychological Measurement, 63*(6), 986-1001. doi:10.1177/0013164403258403

Hough, L. & Dilchert, S. (2017). Personality: Its measurement and validity for employee selection. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (2nd ed., pp. 251-276). New York, NY: Rutledge.

Hummel, H. G. K., Brinke, D. J., Nadolski, R. J., & Baartman, L. K. J. (2017). Content validity of game-based assessment: case study of a serious game for ICT managers in training. *Technology, Pedagogy and Education, 26*(2), 225-240. https://doi.org/10.1080/1475939X.2016.1192060

Hvass, J. S., Larsen, O. S., Vendelbo, K. B., Nilsson, N. C., Nordahl, R., & Serafin, S. (2017, October). *Virtual realism and presence in a virtual reality game*. Paper

presented at the 11th annual meeting of the 3DTV-Conference of the Institute of Electrical and Electronics Engineers (IEEE), Copenhagen, Denmark. doi:10.1109/3DTV.2017.8280421

Hyde, J. S. (1981). How large are cognitive gender differences? A meta-analysis using $\omega^2$ and *d*. *American Psychologist, 36*(8), 892-901. doi:10.1037/0003-066X.36.8.892

Hysenbegasi, A., Hass, S. L., & Rowland, C. R. (2005). The impact of depression on the academic productivity of university students. *The Journal of Mental Health Policy & Economics, 8*(3), 145-151.

Imose, R. A., & Barber, L. K. (2015). Using undergraduate grade point average as a selection tool: A synthesis of the literature. *The Psychologist-Manager Journal, 18*(1), 1-11. http://dx.doi.org/10.1037/mgr0000025

Janeh, O., Bruder, G., Steinicke, F., Gulberti, A., & Poetter-Nerger, M. (2018). Analyses of gait parameters of younger and older adults during (non-)isometric virtual walking. *IEEE Transactions on Visualization and Computer Graphics, 24*(10), 2663-2674. doi:10.1109/TVCG.2017.2771520

Jensen, L., & Konradsen, F. (2018). A review of the use of virtual reality head-mounted displays in education and training. *Education and Information Technologies, 23*(4), 1515-1529. https://doi.org/10.1007/s10639-017-9676-0

Kiili, K., Devlin, K., Perttula, A., Tuomi, P., & Lindstedt, A. (2015). Using video games to combine learning and assessment in mathematics education. *International Journal of Serious Games, 2*(4), 37-55. http://dx.doi.org/10.17083/ijsg.v%vi%i.98

Kiili, K., & Ketamo, H. (2018). Evaluating cognitive and affective outcomes of a digital game-based math test. *IEEE Transactions on Learning Technologies, 11*(2), 255-263. doi:10.1109/TLT.2017.2687458

Kuncel, N. R., & Hezlett, S. A. (2010). Fact and fiction in cognitive ability testing for admissions and hiring decisions. *Current Directions in Psychological Science, 19*(6), 339-345. doi:10.1177/0963721410389459

Lampton, D., Bliss, J., Orvis, K., Kring, J., & Martin, G. A. (2009, October). *A distributed game-based simulation training research testbed.* Paper presented at the 35th annual meeting of the Human Factors and Ergonomics Society, San Antonio, TX. https://doi.org/10.1177/154193120905302703

Lowman, G. H. (2016). Moving beyond identification: Using gamification to attract and retain talent. *Industrial and Organizational Psychology, 9*(3), 677-682. doi:10.1017/iop.2016.70

Maples, J. L., Guan, L., Carter, N. T., & Miller, J. D. (2014). A test of the International Personality Item Pool representation of the Revised NEO Personality Inventory and development of a 120-item IPIP-based measure of the five-factor model. *Psychological Assessment, 26*(4), 1070-1084. http://dx.doi.org/10.1037/pas0000004

Martindale, J. (2018, April 4). *Oculus Rift vs. HTC Vive*. Retrieved from https://www.digitaltrends.com/virtual-reality/oculus-rift-vs-htc-vive/

McCrae, R. R., & Costa, P. T., Jr. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52*(1), 81-90. http://dx.doi.org/10.1037/0022-3514.52.1.81

McPherson, J., & Burns, N. R. (2008). Assessing the validity of computer-game-like tests of processing speed and working memory. *Behavior Research Methods, 40*(4), 969-981. doi:10.3758/BRM.40.4.969

Miller, J. D., Gentile, B., Carter, N. T., Crowe, M., Hoffman, B. J., & Campbell, W. K. (2018). A comparison of the nomological networks associated with forced-choice and Likert formats of the Narcissistic Personality Inventory. *Journal of Personality Assessment, 100*(3), 259-267. https://doi.org/10.1080/00223891.2017.1310731

Montefiori, L. (2016, September). *Psychometrically valid game-based assessment: Fact or Fiction?* Presented at the 8th annual meeting of the Europe Association of Test Publishers (E-ATP) Conference, Lisbon, Portugal.

Nte, S., & Stephens, R. (2008). Videogame aesthetics and e-learning: A retro-looking computer game to explain the normal distribution in statistics teaching. In T. Conolly & M. Stansfield (Eds.), *2nd European Conference on Games Based Learning* (pp. 341-348). Reading, UK: Academic Publishing Limited.

O'Connor, T. J., Cooper, R. A., Fitzgerald, S. G., Dvorznak, M. J., Boninger, M. L., VanSickle, D. P., & Glass, L. (2000). Evaluation of a manual wheelchair interface

to computer games. *Neurorehabilitation and Neural Repair, 14*(1), 12-31. doi:10.1177/154596830001400103

Ones, D. S., Dilchert, S., Viswesvaran, C., & Salgado, J. F. (2017). Cognitive ability: Measurement and validity for employee selection. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (2nd ed., pp. 251-276). New York, NY: Rutledge.

Orvis, K. A., Moore, J. C., Belanich, J., Murphy, J. S., & Horn, D. B. (2010). Are soldiers gamers? Videogame usage among soldiers and implications for the effective use of serious videogames for military training. *Psychology, 22*(2), 143-157. doi:10.1080/08995600903417225

Outtz, J. L. (2002). The role of cognitive ability tests in employment selection. *Human Performance, 15*(1-2), 161-171. http://dx.doi.org/10.1207/S15327043HUP1501&02_10

Plass, J. L, Homer, B. D., Kinzer, C., Frye, J. M., & Perlin, K. (2011). *Learning mechanisms and assessment mechanics for games for learning* [White paper]. Teachers College Columbia University, New York, NY: Institute for Games for Learning. doi:10.13140/2.1.3127.1201

Ployhart, R. E., & Holtz, B. C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61*(1), 153-172. https://doi.org/10.1111/j.1744-6570.2008.00109.x

Ruch, W. W., Stang, S. W., McKillip, R. H., & Dye, D. A. (1994). *Employee Aptitude Survey: Technical manual* (2nd ed.). Burbank, CA: PSI Services LLC.

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High-stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist, 63*(4), 21-227. doi:10.1037/0003-066X.63.4.215

Sanchez, D. R. & Langer, M. (2018). *Video game experience (VGE): Designing and validating a scale for game-based application.* Paper presented at the 11th annual meeting of the International Testing Commission (ITC) Conference, Montreal, Canada.

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:0033-2909/98/S3.00

Schmit, M. J., Ryan, A. M., Stierwalt, S. L., & Powell, A. B. (1995). Frame-of-reference effects on personality scale scores and criterion-related validity. *Journal of Applied Psychology, 80*(5), 607-620. http://dx.doi.org/10.1037/0021-9010.80.5.607

Shin, D. (2017). How does immersion work in augmented reality games? A user-centric view of immersion and engagement. *Information, Communication & Society, 20*, 1-18. https://doi.org/10.1080/1369118X.2017.1411519

Shute, V. J., & Emihovich, B. (2018). Assessing problem-solving skills in game-based immersive environments. In J. Voogt, G. Knezek, R. Christensen, & K.-W. Lai (Eds.), *Second handbook of information technology in primary and secondary education* (pp. 635-648). Berlin, Germany: Springer.

Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and learning of qualitative physics in Newton's Playground. *The Journal of Educational Research, 106*, 423-430. doi:10.1080/00220671.2013.832970

Sitzmann, T. (2011). A meta-analytic examination of the instructional effectiveness of computer-based simulation games. *Personnel Psychology, 64*(2), 489-528. https://doi.org/10.1111/j.1744-6570.2011.01190.x

Stanton, B., Jia, M. Low, J., Nguyen, T., Chen, L., & Thielke, V. (2017, November 27). *Media alert: Virtual reality headset shipments top 1 million for the first time.* Retrieved from https://www.canalys.com/newsroom/media-alert-virtual-reality-headset-shipments-top-1-million-first-time

Sykes, J. M. (2018). Digital games and language teaching and learning. *Foreign Language Annals, 51*(1), 219-224. https://doi.org/10.1111/flan.12325

Thompson, O., Barrett, S., Patterson, C., & Craig, D. (2012). Examining the neurocognitive validity of commercially available, smartphone-based puzzle games. *Psychology, 3*(7) 525-526. http://dx.doi.org/10.4236/psych.2012.37076

Tippins, N. (2010). Adverse impact in employee selection procedures from the perspective of an organizational consultant. In J. L. Outtz (Ed.), *Adverse impact:*

*Implications for organizational staffing and high stakes selection* (pp. 201-225). New York, NY: Rutledge.

Toldi, N. L. (2011). Job applicants favor video interviewing in the candidate-selection process. *Employment Relations, 38*(3), 19-27. https://doi.org/10.1002/ert.20351

Ventura, M., & Shute, V. (2013). The validity of a game-based assessment of persistence. *Computers in Human Behavior, 29*(6), 2568-2572. http://dx.doi.org/10.1016/j.chb.2013.06.033

Vince, J. (1993). Virtual reality techniques in flight simulation. In R. A. Earnshaw, M. A. Gigante, & H. Jones (Eds.), *Virtual reality systems* (pp. 135-142). London: Academic Press Limited.

Vogel, J. J., Vogel, D. S., Cannon-Bowers, J., Bowers, C. A., Muse, K., & Wright, M. (2006). Computer gaming and interactive simulations for learning: A meta-analysis. *Journal of Educational Computing Research, 34*, 229-243. https://doi.org/10.2190/FLHV-K4WA-WPVQ-H0YM

Zapata-Rivera, D. & Bauer, M. (2012). Exploring the role of games in educational assessment. In M. C. Mayrath, J. Clarke-Midura, D. H. Robinson, & G. Schraw (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research.* (pp. 147–169). Charlotte, NC: Information Age Publishing.

Zuckerberg, M., Bosworth, A., Bram, R., Chen, L., Liu, S., Beck, V. D., ... & Abrash, M. (2018, September). *Keynote day 01*. Presented at the 5th annual Oculus Connect (Oculus Connect 5) Conference, San Jose, CA.

**Table 1. Participant demographics**

| | *N* | % |
|---|---|---|
| Gender | | |
| Female | 88 | 70.97 |
| Male | 34 | 27.42 |
| Other | 2 | 1.61 |
| Ethnicity | | |
| Asian | 31 | 25.00 |
| Latino | 35 | 28.23 |
| White | 35 | 28.23 |
| Other | 23 | 18.55 |
| Study Order | | |
| VR assessments first | 62 | 50.00 |
| Self-report assessments first | 62 | 50.00 |
| Total | 124 | |

**Table 2. Descriptive statistics and study correlations**

| | *M* | *SD* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. Age | 24.15 | 6.88 | — | | | | | | | | | | | | | | |
| 2. Gender | .28 | 0.45 | .07 | — | | | | | | | | | | | | | |
| 3. GPA | 3.43 | 0.56 | .20* | -.05 | — | | | | | | | | | | | | |
| 4. Video game experience | 3.00 | 0.77 | -.12 | .40** | -.16 | (.94) | | | | | | | | | | | |
| 5. Intimidation with games | 2.50 | 0.36 | .02 | -.39** | .14 | -.57** | (.88) | | | | | | | | | | |
| 6. Openness | 3.64 | 0.50 | -.05 | -.00 | .06 | .22* | -.26** | (.84) | | | | | | | | | |
| 7. Conscientiousness | 3.66 | 0.45 | .38** | .06 | .15 | -.16 | -.00 | -.01 | (.83) | | | | | | | | |
| 8. Extraversion | 3.38 | 0.50 | .12 | .07 | -.14 | .10 | -.11 | .29** | .31** | (.88) | | | | | | | |
| 9. Agreeableness | 3.72 | 0.40 | .18* | -.22* | .05 | -.14 | .09 | .09 | .43** | .08 | (.79) | | | | | | |
| 10. Neuroticism | 2.78 | 0.56 | -.42** | -.20* | .06 | .08 | .14 | -.12 | -.51** | -.44** | -.10 | (.87) | | | | | |
| 11. SV (Self-Report) | 47.44 | 31.22 | -.01 | .34** | .20* | .06 | -.13 | .29** | -.07 | .10 | .03 | -.03 | — | | | | |
| 12. VSA (Self-Report) | 33.94 | 32.08 | -.24** | .09 | .16 | .20* | -.08 | .10 | -.24** | -.09 | -.13 | .18* | .47** | — | | | |
| 13. VP (Self-Report) | 37.45 | 26.10 | .05 | .23* | .07 | .20* | -.10 | .21* | -.02 | .19* | -.01 | -.11 | .54** | .31** | — | | |
| 16. SV (VR) | .14 | .10 | -.10 | .26** | .08 | .21* | -.29** | .18* | -.07 | -.01 | .00 | -.04 | .14 | .25** | .16 | — | |
| 15. VSA (VR) | .71 | .13 | -.26** | .25** | .19* | .18 | -.22* | .30** | -.13 | -.03 | .04 | .08 | .41** | .32** | .45** | .45** | — |
| 14. VP (VR) | .53 | .22 | .01 | .12 | .05 | .19* | -.26** | .16 | -.08 | .08 | .08 | .13 | .23* | .04 | .39** | .13 | .34** |

*$p < .05$, **$p < .01$.

*Note.* Gender: 0 = Female; 1 = Male. GPA evaluated on a 0.0. to 4.0 scale. Cognitive Ability Assessments: SV = Space Visualization, VSA = Visual Speed & Accuracy, and VP = Visual Pursuit. Reliability values are provided in parentheses along the diagonal.

**Table 3. Hypothesis 1: Convergent validity evidence**

| | | *B* | *SE* | *β* | 95% CI | $\Delta R^2$ | *p* | *F(df)* |
|---|---|---|---|---|---|---|---|---|
| Step 1 | Video game experience | -2.36 | 4.59 | -.06 | -11.45, 6.73 | .01 | .48 | 0.79(2, 118) |
| | + Intimidation with games | -4.83 | 3.89 | -.14 | -12.54, 2.88 | | | |
| Step 2 | SV (VR) | 47.14 | 30.11 | .15 | -12.50, 106.78 | .02 | .12 | 1.35(3, 118) |
| Step 1 | Video game experience | 9.11 | 4.65 | .21 | -0.10, 18.32 | .04 | .12 | 2.12(2, 118) |
| | + Intimidation with games | 2.12 | 3.94 | .06 | -5.68, 9.91 | | | |
| Step 2 | VSA (VR) | 76.51 | 21.60 | .32 | 33.73, 119.29 | .10 | .001 | 5.74(3, 118) |
| Step 1 | Video game experience | 7.03 | 3.81 | .20 | -0.51, 14.57 | .03 | .14 | 1.97(2, 118) |
| | + Intimidation with games | 1.30 | 3.24 | .04 | -5.11, 7.71 | | | |
| Step 2 | VP (VR) | 26.02 | 11.25 | .22 | 3.73, 48.31 | .04 | .02 | 3.15(3, 118) |

*Note.* Hierarchical regression using VR assessment scores to predict self-report assessment scores for Space Visualization (SV), Visual Speed & Accuracy (VSA), and Visual Pursuit (VP), controlling for video game experience and intimidation with games.

**Table 4. Hypothesis 2a: Divergent validity evidence (Space Visualization)**

| | | *B* | *SE* | *β* | 95% CI | $\Delta R^2$ | *p* | *F(df)* |
|---|---|---|---|---|---|---|---|---|
| *Openness* | | | | | | | | |
| Step 1 | Video game experience | 0.06 | 0.07 | .09 | -0.08, 0.20 | .07 | .02 | 4.21(2, 118) |
| | + Intimidation with games | -0.11 | 0.06 | -.20 | -0.23, 0.01 | | | |
| Step 2 | SV (VR) | 0.58 | 0.47 | .17 | -0.35, 1.51 | .01 | .22 | 3.33(3, 118) |
| *Conscientiousness* | | | | | | | | |
| Step 1 | Video game experience | -0.14 | 0.07 | -.23 | -0.27, -0.01 | .04 | .11 | 2.21(2, 118) |
| | + Intimidation with games | -0.08 | 0.06 | -.16 | -0.19, 0.03 | | | |
| Step 2 | SV (VR) | -0.35 | 0.44 | -.08 | -1.21, 0.52 | .01 | .43 | 1.68(3, 118) |
| *Extraversion* | | | | | | | | |
| Step 1 | Video game experience | 0.04 | 0.07 | .06 | -0.11, 0.18 | .02 | .39 | 0.95(2, 118) |
| | + Intimidation with games | -0.05 | 0.06 | -.09 | -0.17, 0.07 | | | |
| Step 2 | SV (VR) | -0.27 | 0.49 | -.06 | -1.24, 0.69 | <.001 | .57 | 0.74(3, 118) |
| *Agreeableness* | | | | | | | | |
| Step 1 | Video game experience | -0.08 | 0.06 | -.14 | -0.19, 0.04 | .02 | .30 | 1.22(2, 118) |
| | + Intimidation with games | <0.01 | 0.05 | <.01 | -0.10, 0.10 | | | |
| Step 2 | SV (VR) | 0.14 | 0.40 | .03 | -0.64, 0.92 | <.001 | .72 | 0.85(3, 118) |
| *Neuroticism* | | | | | | | | |
| Step 1 | Video game experience | 0.18 | 0.08 | .25 | 0.02, .034 | .06 | .03 | 3.54(2, 118) |
| | + Intimidation with games | 0.16 | 0.07 | .26 | 0.03, 0.30 | | | |
| Step 2 | SV (VR) | -0.07 | 0.53 | -.01 | -1.12, 0.99 | <.001 | .90 | 2.45(3, 118) |

*Note.* Hierarchical regression using VR assessment scores for Space Visualization (SV) to predict OCEAN personality trait scores, accounting for video game experience and intimidation with games.

**Table 5. Hypothesis 2b: Divergent validity evidence (Visual Speed & Accuracy)**

| | | *B* | *SE* | *β* | 95% CI | $\Delta R^2$ | *p* | *F(df)* |
|---|---|---|---|---|---|---|---|---|
| *Openness* | | | | | | | | |
| Step 1 | Video game experience | 0.06 | 0.07 | .09 | -0.08, 0.20 | .07 | .02 | 4.08(2, 118) |
| | + Intimidation with games | -0.11 | 0.06 | -.19 | -0.23, 0.01 | | | |
| Step 2 | VSA (VR) | 0.98 | 0.34 | .26 | 0.31, 1.65 | .06 | .004 | 5.69(3, 118) |
| *Conscientiousness* | | | | | | | | |
| Step 1 | Video game experience | -0.12 | 0.07 | -.20 | -0.25, 0.01 | .03 | .20 | 1.65(2, 118) |
| | + Intimidation with games | -0.07 | 0.06 | -.14 | -0.18, 0.04 | | | |
| Step 2 | VSA (VR) | -0.47 | 0.32 | -.14 | -1.10, 0.17 | .02 | .15 | 1.83(3, 118) |
| *Extraversion* | | | | | | | | |
| Step 1 | Video game experience | 0.05 | 0.07 | .07 | -0.10, 0.20 | .02 | .39 | 0.95(2, 118) |
| | + Intimidation with games | -0.04 | 0.06 | -.07 | -0.17, 0.08 | | | |
| Step 2 | VSA (VR) | -0.23 | 0.36 | -.06 | -0.95, 0.49 | <.001 | .52 | 0.76(3, 118) |
| *Agreeableness* | | | | | | | | |
| Step 1 | Video game experience | -0.08 | 0.06 | -.15 | -0.20, 0.04 | .02 | .28 | 1.28(2, 118) |
| | + Intimidation with games | 0.00 | 0.05 | <.01 | -0.10, 0.10 | | | |
| Step 2 | VSA (VR) | 0.22 | 0.29 | .07 | -0.36, 0.79 | .01 | .46 | 1.03(3, 118) |
| *Neuroticism* | | | | | | | | |
| Step 1 | Video game experience | 0.16 | 0.08 | .22 | 0.00, 0.32 | .05 | .06 | 2.93(2, 118) |
| | + Intimidation with games | 0.15 | 0.07 | .24 | 0.02, 0.29 | | | |
| Step 2 | VSA (VR) | 0.43 | 0.39 | .10 | -0.35, 1.20 | .01 | .28 | 2.36(3, 118) |

*Note.* Hierarchical regression using VR assessment scores for Visual Speed & Accuracy (VSA) to predict OCEAN personality trait scores, accounting for video game experience and intimidation with games.

**Table 6. Hypothesis 2c: Divergent validity evidence (Visual Pursuit)**

| | | *B* | *SE* | *β* | 95% CI | $\Delta R^2$ | *p* | *F(df)* |
|---|---|---|---|---|---|---|---|---|
| *Openness* | | | | | | | | |
| Step 1 | Video game experience | 0.07 | 0.07 | .10 | -0.07, .209 | .07 | .02 | 4.21(2, 118) |
| | + Intimidation with games | -0.11 | 0.06 | -.19 | -0.23, 0.01 | | | |
| Step 2 | VP (VR) | 0.22 | 0.22 | .10 | -0.21, 0.65 | .01 | .31 | 3.16(3, 118) |
| *Conscientiousness* | | | | | | | | |
| Step 1 | Video game experience | -0.13 | 0.07 | -.21 | -0.26, 0.00 | .03 | .14 | 1.99(2, 118) |
| | + Intimidation with games | -0.08 | 0.06 | -.15 | -0.19, 0.04 | | | |
| Step 2 | VP (VR) | -0.16 | 0.20 | -.08 | -0.56, 0.23 | .01 | .42 | 1.54(3, 118) |
| *Extraversion* | | | | | | | | |
| Step 1 | Video game experience | 0.05 | 0.07 | .08 | -0.10, 0.20 | .02 | .37 | 0.99(2, 118) |
| | + Intimidation with games | -0.04 | 0.06 | -.07 | -0.17, 0.08 | | | |
| Step 2 | VP (VR) | 0.11 | 0.22 | .05 | -0.33, 0.55 | <.001 | .63 | 0.74(3, 118) |
| *Agreeableness* | | | | | | | | |
| Step 1 | Video game experience | -0.08 | 0.06 | -.15 | -0.20, 0.04 | .02 | .28 | 1.27(2, 118) |
| | + Intimidation with games | 0.00 | 0.05 | <.01 | -0.10, 0.10 | | | |
| Step 2 | VP (VR) | 0.21 | 0.18 | .11 | -0.15, 0.56 | .01 | .25 | 1.30(3, 118) |
| *Neuroticism* | | | | | | | | |
| Step 1 | Video game experience | 0.17 | 0.08 | .23 | 0.01, 0.33 | .05 | .04 | 3.23(2, 118) |
| | + Intimidation with games | 0.16 | 0.07 | .25 | 0.02, 0.29 | | | |
| Step 2 | VP (VR) | 0.41 | 0.24 | .16 | -0.07, 0.88 | .02 | .09 | 3.16(3, 118) |

*Note.* Hierarchical regression using VR assessment scores for Visual Pursuit (VP) to predict OCEAN personality trait scores, accounting for video game experience and intimidation with games.

**Table 7. Hypothesis 3: Criterion-related validity evidence**

| | | *B* | *SE* | *β* | 95% CI | $\Delta R^2$ | *p* | *F(df)* |
|---|---|---|---|---|---|---|---|---|
| Step 1 | Video game experience | -0.10 | 0.08 | -.13 | -0.26, 0.06 | .04 | .11 | 2.22(2, 117) |
| | + Intimidation with games | 0.05 | 0.07 | .08 | -0.09, 0.19 | | | |
| Step 2 | SV (VR) | 0.83 | 0.54 | .15 | -0.23, 1.90 | .02 | .13 | 2.30(3, 117) |
| Step 1 | Video game experience | -0.10 | 0.08 | -.13 | -0.26, 0.07 | .04 | .12 | 2.13(2, 117) |
| | + Intimidation with games | 0.06 | 0.07 | .09 | -0.09, 0.19 | | | |
| Step 2 | VSA (VR) | 1.05 | 0.39 | .25 | 0.28, 1.82 | .06 | .01 | 3.96(3, 117) |
| Step 1 | Video game experience | -0.11 | 0.08 | -.14 | -0.27, 0.06 | .04 | .11 | 2.27(2, 117) |
| | + Intimidation with games | 0.05 | 0.07 | .08 | -0.09, 0.19 | | | |
| Step 2 | VP (VR) | 0.28 | 0.25 | .11 | -0.21, 0.76 | .01 | .26 | 1.95(3, 117) |

*Note.* Hierarchical regression using VR assessment scores for Space Visualization (SV), Visual Speed & Accuracy (VSA), and Visual Pursuit (VP) to predict GPA, accounting for video game experience and intimidation with games.

**Table 8. Incremental predictive validity evidence**

| | | *B* | *SE* | *β* | 95% CI | $\Delta R^2$ | *p* | *F(df)* |
|---|---|---|---|---|---|---|---|---|
| Step 1 | Video game experience | -0.10 | 0.08 | -0.13 | -0.26, 0.08 | .04 | .11 | 2.22(2, 117) |
| | + Intimidation with games | 0.05 | 0.07 | 0.08 | -0.09, 0.19 | | | |
| Step 2 | SV (Self-Report) | 0.01 | 0.01 | 0.21 | <0.01, 0.01 | .05 | .02 | 3.41(3, 117) |
| Step 3 | SV (VR) | 0.65 | 0.54 | 0.12 | -0.41, 1.71 | .01 | .22 | 2.94(4, 117) |
| Step 1 | Video game experience | -0.10 | 0.08 | -0.13 | -0.26, 0.07 | .04 | .12 | 2.13(2, 117) |
| | + Intimidation with games | 0.06 | 0.07 | 0.07 | -0.09, 0.19 | | | |
| Step 2 | VSA (Self-Report) | 0.01 | 0.01 | 0.21 | <0.01, 0.01 | .04 | .03 | 3.14(3, 117) |
| Step 3 | VSA (VR) | 0.86 | 0.41 | 0.20 | 0.06, 1.66 | .04 | .04 | 3.56(4, 117) |
| Step 1 | Video game experience | -0.11 | 0.08 | -0.14 | -0.27, 0.06 | .04 | .11 | 2.27(2, 117) |
| | + Intimidation with games | 0.05 | 0.07 | 0.08 | -0.09, 0.19 | | | |
| Step 2 | VP (Self-Report) | 0.01 | 0.01 | 0.11 | -0.01, 0.01 | .01 | .24 | 1.97(3, 117) |
| Step 3 | VP (VR) | 0.23 | 0.25 | 0.09 | -0.27, 0.72 | .01 | .37 | 1.68(4, 117) |

*Note.* Hierarchical regression using VR assessment scores for Space Visualization (SV), Visual Speed & Accuracy (VSA), and Visual Pursuit (VP) to predict GPA, accounting for video game experience and intimidation with games in Step 1 and self-report assessment scores in Step 2 of each model.

**Table 9. Adverse impact calculation across demographic comparisons**

| | Percentile | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 50% | | | | 75% | | | | 90% | | | |
| | *4/5* | $\chi^2$ | *Z* | *Fischer* | *4/5* | $\chi^2$ | *Z* | *Fischer* | *4/5* | $\chi^2$ | *Z* | *Fischer* |
| **Space Visualization (SV)** | | | | | | | | | | | | |
| *Self-Report Assessment* | | | | | | | | | | | | |
| Female/Male | 0.56† | .001** | 3.23† | <.01† | 0.51† | .03* | 2.18† | .03† | 0.33† | .03* | 2.21† | .05† |
| Asian/White | 0.95 | .81 | 0.25 | .99 | 1.01 | .99 | <0.01 | .99 | 1.41 | .58 | 0.56 | .72 |
| Latino/White | 0.72† | .23 | 1.20 | .34 | 0.78† | .57 | 0.57 | .78 | 0.50† | .39 | 0.85 | .67 |
| *VR Assessment* | | | | | | | | | | | | |
| Female/Male | 0.57† | .002** | 3.08† | <.01† | 0.44† | .008** | 2.64† | .02† | 0.54† | .27 | 1.12 | .31 |
| Asian/White | 1.06 | .79 | 0.27 | .81 | 1.22 | .67 | 0.43 | .77 | 6.26 | .04* | 2.09† | .05† |
| Latino/White | 1.00 | .99 | <0.01 | .99 | 1.14 | .77 | 0.29 | .99 | 2.00 | .56 | 0.59 | .99 |
| **Visual Speed & Accuracy (VSA)** | | | | | | | | | | | | |
| *Self-Report Assessment* | | | | | | | | | | | | |
| Female/Male | 0.76† | .13 | 1.50 | .16 | 0.94 | .87 | 0.18 | .99 | 1.16 | .82 | 0.23 | .99 |
| Asian/White | 1.92 | .002** | 3.07† | <.01† | 2.44 | .03* | 2.22† | .03† | 2.86 | .17 | 1.37 | .24 |
| Latino/White | 0.93 | .81 | 0.24 | .99 | 0.67† | .50 | 0.68 | .73 | 1.00 | .99 | <0.01 | .99 |
| *VR Assessment* | | | | | | | | | | | | |
| Female/Male | 0.69† | .04* | 2.10† | .04† | 0.48† | .02* | 2.33† | .03† | 0.54† | .27 | 1.12 | .31 |
| Asian/White | 0.94 | .81 | 0.23 | .99 | 0.91 | .81 | 0.24 | .99 | 0.85 | .82 | 0.23 | .99 |
| Latino/White | 0.97 | .91 | 0.12 | .99 | 0.52† | .16 | 1.40 | .24 | 0.77† | .72 | 0.36 | .99 |
| **Visual Pursuit (VP)** | | | | | | | | | | | | |
| *Self-Report Assessment* | | | | | | | | | | | | |
| Female/Male | 0.59† | .03* | 3.01† | <.01† | 0.51† | .03* | 2.18† | .04† | 0.54† | .26 | 1.12 | .31 |
| Asian/White | 1.06 | .81 | 0.25 | .99 | 1.25 | .56 | 0.59 | .60 | 0.56† | .48 | 0.70 | .68 |
| Latino/White | 1.00 | .99 | <0.01 | .99 | 0.67† | .38 | 0.87 | .56 | 0.25† | .16 | 1.39 | .36 |
| *VR Assessment* | | | | | | | | | | | | |
| Female/Male | 0.74† | .14 | 1.49 | .15 | 0.82 | .57 | 0.56 | .63 | 0.28† | .01* | 2.47† | .04† |
| Asian/White | 1.06 | .80 | 0.25 | .99 | 0.80 | .64 | 0.47 | .77 | 0.53† | .44 | 0.78 | .67 |
| Latino/White | 0.88 | .62 | 0.49 | .81 | 1.00 | .99 | <0.01 | .99 | 0.50† | .39 | 0.09 | .67 |

*$p > .05$, **$p > .01$.

*Note.* † = calculations that indicate adverse impact.

*4/5* = Four-fifths ratio test, adverse impact for values below .80.

$\chi^2$ = p-values for Chi-squared test, adverse impact for values below .05.

*Z* = Z test, adverse impact for values above 1.96.

*Fischer* = p-values for Fisher's exact probability test, adverse impact for values below .05.

**Figure 1. Space Visualization (Self-Report Assessment) — Sample Content**

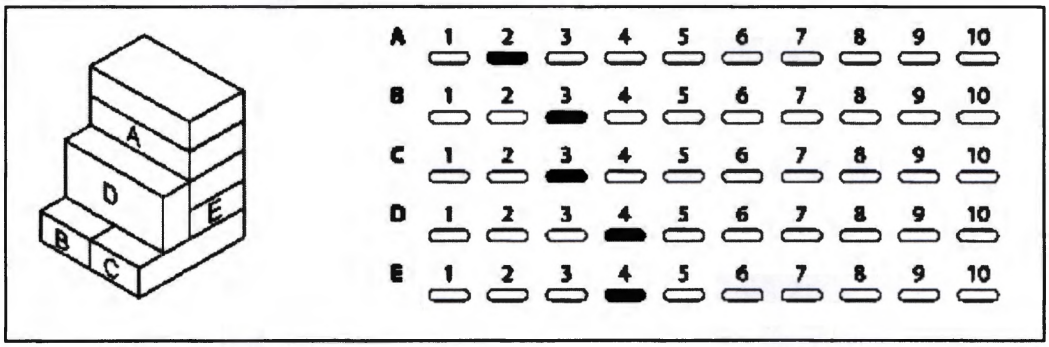


**Figure 2. Visual Speed & Accuracy (Self-Report Assessment) — Sample Content**

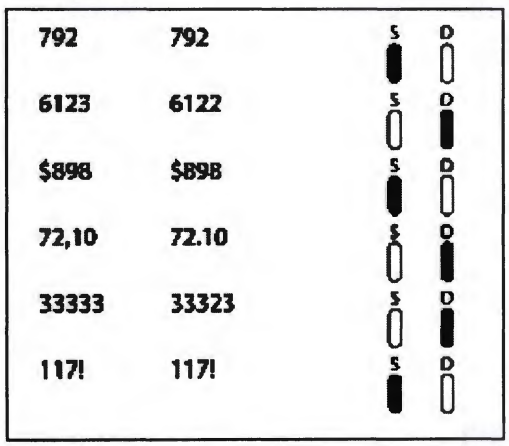


**Figure 3. Visual Pursuit (Self-Report Assessment) — Sample Content**

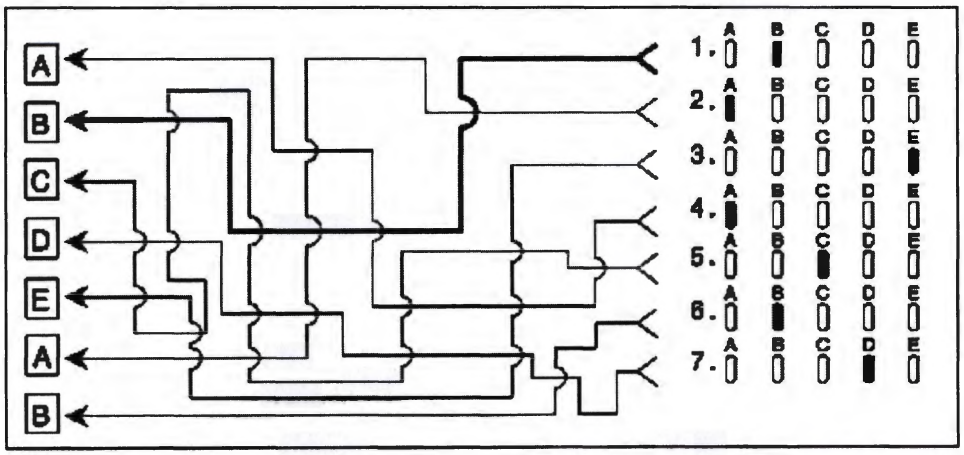


**Figure 4. Space Visualization (VR Assessment) — Sample Content**

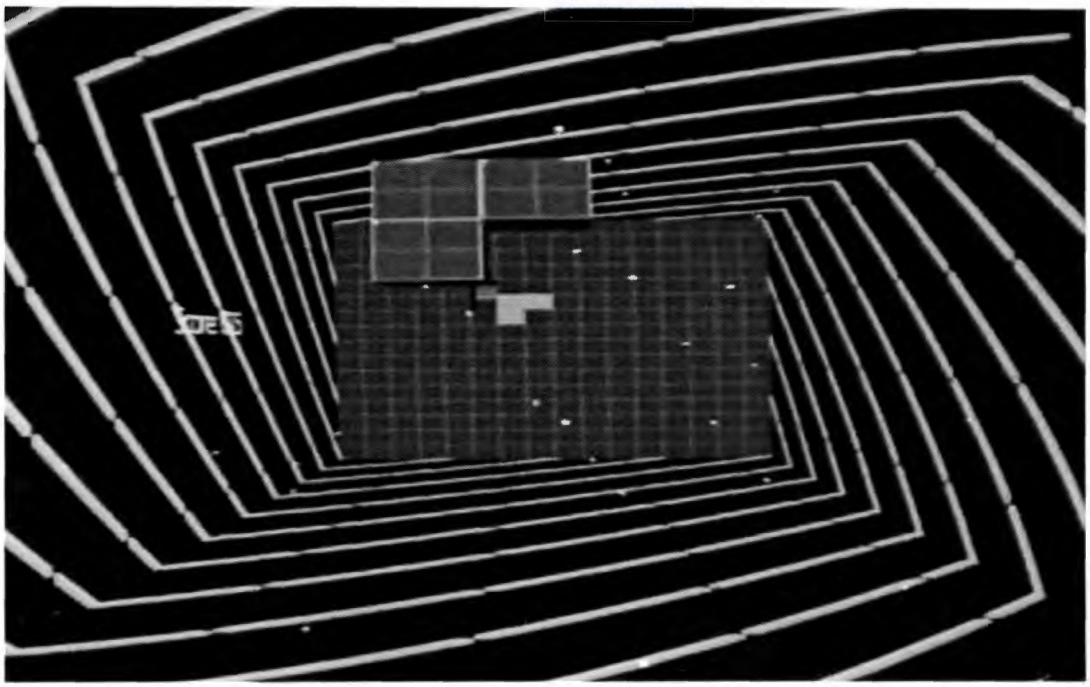
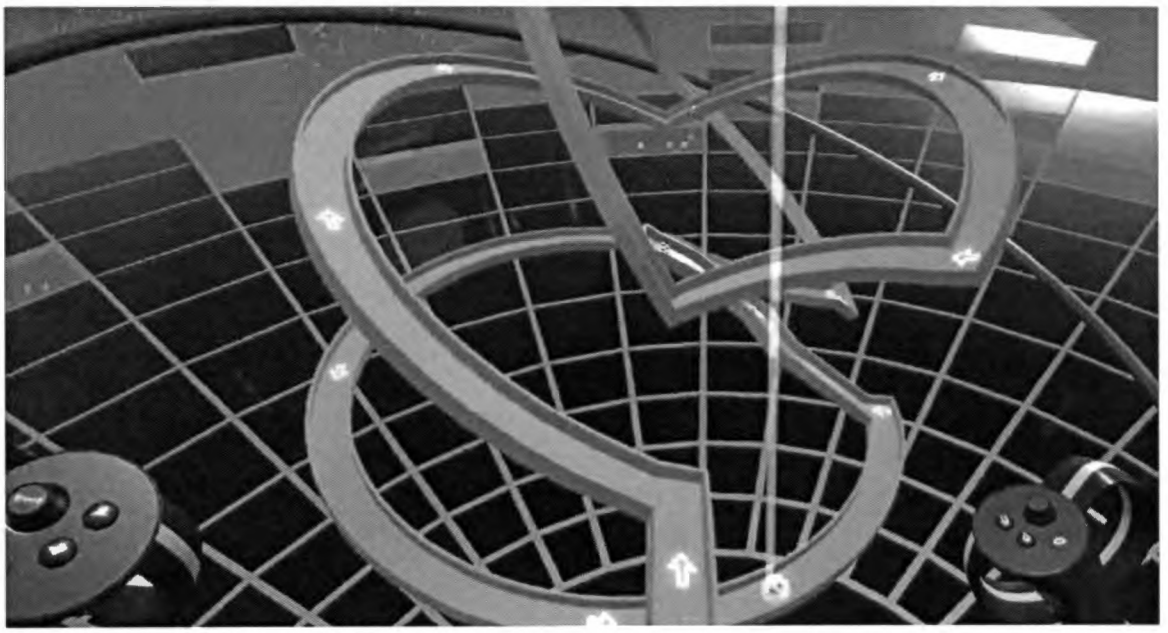


**Figure 5. Visual Speed & Accuracy (VR Assessment) — Sample Content**

**Figure 6. Visual Pursuit (VR Assessment) — Sample Content**



# Appendix 1. Video game experience scale and intimidation with games scale

This survey is not specific to any of your experiences in the study today. Instead, focus on what you know about yourself, and try to answer as honestly as possible. Please remember to notify the researcher when you have completed the survey.

Please read each statement carefully, in reference to your general experience playing video games. Select the option that best corresponds to your agreement or disagreement on the scale provided, ranging from "Strongly disagree" on the left to "Strongly agree" on the right.

| | Strongly disagree | Disagree | Neither agree nor disagree | Agree | Strongly agree |
|---|---|---|---|---|---|
| I plan to continue improving my video game skills. | ○ | ○ | ○ | ○ | ○ |
| I lose track of time when I play video games. | ○ | ○ | ○ | ○ | ○ |
| I am proactive in seeking ways to improve my video game skills. | ○ | ○ | ○ | ○ | ○ |
| I find video game rules confusing. | ○ | ○ | ○ | ○ | ○ |
| Based on my knowledge of previous video games, I can easily see through the rules of a game. | ○ | ○ | ○ | ○ | ○ |
| I am good at video games, compared to others. | ○ | ○ | ○ | ○ | ○ |
| I spend many hours each week playing video games. | ○ | ○ | ○ | ○ | ○ |
| Learning how to play a video game is confusing to me. | ○ | ○ | ○ | ○ | ○ |
| I think video games are entertaining. | ○ | ○ | ○ | ○ | ○ |
| I deliberately seek out video games to play. | ○ | ○ | ○ | ○ | ○ |
| Video games are fun. | ○ | ○ | ○ | ○ | ○ |
| When I play video games I lose track of my senses (e.g., can't tell if I am getting hungry or tired). | ○ | ○ | ○ | ○ | ○ |
| I would call myself a "serious gamer." | ○ | ○ | ○ | ○ | ○ |
| I have searched for information (e.g., magazine or websites) to improve my gaming skills. | ○ | ○ | ○ | ○ | ○ |
| It takes me a long time to understand the controls of a video game. | ○ | ○ | ○ | ○ | ○ |
| I find it difficult to understand video games. | ○ | ○ | ○ | ○ | ○ |
| I have good video game skills. | ○ | ○ | ○ | ○ | ○ |
| I would need help to figure out the controls of a video game. | ○ | ○ | ○ | ○ | ○ |
| Video games are intimidating to me. | ○ | ○ | ○ | ○ | ○ |
| I am confident playing video games. | ○ | ○ | ○ | ○ | ○ |
| I like playing video games. | ○ | ○ | ○ | ○ | ○ |
| I enjoy playing video games. | ○ | ○ | ○ | ○ | ○ |
| I can keep up with a video game that moves quickly. | ○ | ○ | ○ | ○ | ○ |
| I have a lot of experience with playing video games. | ○ | ○ | ○ | ○ | ○ |
| I am fully immersed when I play video games. | ○ | ○ | ○ | ○ | ○ |

## Appendix 2. International Personality Item Pool 120-item scale

The following statements describe people's behaviors. Please use the rating scale next to each phrase to describe how accurately each statement describes you. Describe yourself honestly and as you generally are now, not as you wish to be in the future.

Please read each statement carefully and select the option that best corresponds to your agreement or disagreement on the scale provided, ranging from "Strongly disagree" on the left to "Strongly agree" on the right.

| | Strongly disagree | Disagree | Neutral | Agree | Strongly agree |
|---|---|---|---|---|---|
| Go on binges. | ○ | ○ | ○ | ○ | ○ |
| Have a vivid imagination. | ○ | ○ | ○ | ○ | ○ |
| Love to help others. | ○ | ○ | ○ | ○ | ○ |
| Know how to get things done. | ○ | ○ | ○ | ○ | ○ |
| Find it difficult to approach others. | ○ | ○ | ○ | ○ | ○ |
| Try to lead others. | ○ | ○ | ○ | ○ | ○ |
| Do more than what's expected of me. | ○ | ○ | ○ | ○ | ○ |
| Believe in one true religion. | ○ | ○ | ○ | ○ | ○ |
| Am not interested in abstract ideas. | ○ | ○ | ○ | ○ | ○ |
| Cheat to get ahead. | ○ | ○ | ○ | ○ | ○ |
| Feel sympathy for those who are worse off than myself. | ○ | ○ | ○ | ○ | ○ |
| Do a lot in my spare time. | ○ | ○ | ○ | ○ | ○ |
| Need a push to get started. | ○ | ○ | ○ | ○ | ○ |
| Excel in what I do. | ○ | ○ | ○ | ○ | ○ |
| Warm up quickly to others. | ○ | ○ | ○ | ○ | ○ |
| Am able to stand up for myself. | ○ | ○ | ○ | ○ | ○ |
| Distrust people. | ○ | ○ | ○ | ○ | ○ |
| Laugh aloud. | ○ | ○ | ○ | ○ | ○ |
| Enjoy being reckless. | ○ | ○ | ○ | ○ | ○ |
| Love action. | ○ | ○ | ○ | ○ | ○ |
| Set high standards for myself and others. | ○ | ○ | ○ | ○ | ○ |
| Love life. | ○ | ○ | ○ | ○ | ○ |
| Rarely get irritated. | ○ | ○ | ○ | ○ | ○ |
| Complete tasks successfully. | ○ | ○ | ○ | ○ | ○ |

| | Strongly disagree | Disagree | Neutral | Agree | Strongly agree |
|---|---|---|---|---|---|
| Am attached to conventional ways. | | | | | |
| Worry about things. | | | | | |
| Rush into things. | | | | | |
| Seldom get emotional. | | | | | |
| Jump into things without thinking. | | | | | |
| Tend to vote for conservative political candidates. | | | | | |
| Trust what people say. | | | | | |
| Have difficulty understanding abstract ideas. | | | | | |
| Am calm even in tense situations. | | | | | |
| Use flattery to get ahead. | | | | | |
| Don't like crowded events. | | | | | |
| Like to get lost in thought. | | | | | |
| Seek adventure. | | | | | |
| Feel comfortable around people. | | | | | |
| Fear for the worst. | | | | | |
| Often feel blue. | | | | | |
| Get back at others. | | | | | |
| Rarely overindulge. | | | | | |
| Tend to vote for liberal political candidates. | | | | | |
| Love a good fight. | | | | | |
| Act without thinking. | | | | | |
| Dislike changes. | | | | | |
| Am concerned about others. | | | | | |
| Often eat too much. | | | | | |
| Am not easily affected by my emotions. | | | | | |
| Trust others. | | | | | |
| Know how to get around the rules. | | | | | |
| Start tasks right away. | | | | | |
| Find it difficult to get down to work. | | | | | |
| Tell the truth. | | | | | |

| | Strongly disagree | Disagree | Neutral | Agree | Strongly agree |
|---|---|---|---|---|---|
| Dislike myself. | | | | | |
| Avoid philosophical discussions. | | | | | |
| Am not interested in theoretical discussions. | | | | | |
| Love to daydream. | | | | | |
| Get others to do my duties. | | | | | |
| Have a high opinion of myself. | | | | | |
| Leave a mess in my room. | | | | | |
| Am not interested in other people's problems. | | | | | |
| Make people feel welcome. | | | | | |
| Talk to a lot of different people at parties. | | | | | |
| Like to tidy up. | | | | | |
| Break my promises. | | | | | |
| Do not enjoy going to art museums. | | | | | |
| Remain calm under pressure. | | | | | |
| Feel that I'm unable to deal with things. | | | | | |
| Get angry easily. | | | | | |
| Do not like poetry. | | | | | |
| Am always on the go. | | | | | |
| Am not highly motivated to succeed. | | | | | |
| Make myself the center of attention. | | | | | |
| Insult people. | | | | | |
| Get irritated easily. | | | | | |
| Make rash decisions. | | | | | |
| Can manage many things at the same time. | | | | | |
| Like order. | | | | | |
| Work hard. | | | | | |
| Keep my promises. | | | | | |
| See beauty in things that others might not notice. | | | | | |
| Believe that others have good intentions. | | | | | |
| Enjoy wild flights of fantasy. | | | | | |

| | Strongly disagree | Disagree | Neutral | Agree | Strongly agree |
|---|---|---|---|---|---|
| Act comfortably with others. | | | | | |
| Take advantage of others. | | | | | |
| Handle tasks smoothly. | | | | | |
| Am often down in the dumps. | | | | | |
| Take charge. | | | | | |
| Get stressed out easily. | | | | | |
| Sympathize with the homeless. | | | | | |
| Have a lot of fun. | | | | | |
| Don't like the idea of change. | | | | | |
| Turn my back on others. | | | | | |
| Take control of things. | | | | | |
| Experience very few emotional highs and lows. | | | | | |
| Am afraid of many things. | | | | | |
| Am not embarrassed easily. | | | | | |
| Experience my emotions intensely. | | | | | |
| Know how to cope. | | | | | |
| Make friends easily. | | | | | |
| Radiate joy. | | | | | |
| Love excitement. | | | | | |
| Like to stand during the national anthem. | | | | | |
| Lose my temper. | | | | | |
| Am able to control my cravings. | | | | | |
| Have a low opinion of myself. | | | | | |
| Am always busy. | | | | | |
| Love large parties. | | | | | |
| Have difficulty starting tasks. | | | | | |
| Think highly of myself. | | | | | |
| Avoid crowds. | | | | | |
| Am easily intimidated. | | | | | |
| Leave my belongings around. | | | | | |
| Prefer to stick with things that I know. | | | | | |
| Wait for others to lead the way. | | | | | |
| Yell at people. | | | | | |
| Do not like art. | | | | | |
| Suffer from others' sorrows. | | | | | |
| Believe that I am better than others. | | | | | |

## Appendix 3. Demographic questionnaire

Please enter your SFSU Student ID Number.

Please enter your current age in years (e.g., "28").

Please indicate the gender with which you identify.

- Female
- Male
- Transgender Female
- Transgender Male
- Genderqueer, Genderfluid, or Non-Binary
- Intersex
- Other (please specify)

Please indicate your ethnicity (select all that apply).

- Asian American or other East Asian
- Black or African American
- Indian, Pakistani, or other South Asian
- Latina/o/x or Hispanic
- Native American or Alaska Native
- Native Hawaiian or Pacific Islander
- Middle Eastern or North African
- White, Caucasian, or other European