COGNITIVE ABILITY TEST MEAN SCORE DIFFERENCS ON UNPROCTORED INTERNET-BASED TESTS: SELF-SELECTION OR DEVICE-TYPE EFFECTS?

A Thesis

by

ASHLEIGH SHONTAYAH WILLIAMS

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Winfred Arthur, Jr.
Committee Members,	Stephanie C. Payne
	Hart Blanton
Head of Department,	Rebecca Brooker

May 2022

Major Subject: Industrial/Organizational Psychology

ABSTRACT

Within the assessment literature, there is an ongoing conversation pertaining to observed device-type mean score differences for unproctored internet-based tests (UIT), specifically on cognitive assessments. However, it remains unclear whether observed differences are the result of an aspect of the UIT device used to complete the assessments or some other aspect of the testing situation. Arthur, Keiser, and Doverspike' (2018) literature review concluded that mean score differences have been consistently observed in operational settings for cognitive ability UITs. However, these differences have not been obtained in laboratory, non-operational studies. The present study examines one of the explanations that have been advanced to account for this discrepancy, specifically, the self-selection hypothesis. This explanation posits that cognitive ability differences between those who choose to complete high-stakes cognitively demanding assessments on smartphones (a poor decision or choice) versus those who use devices more conducive to effective performance (e.g., laptops, desktops) may account for the observed differences in that in lab studies participants are randomly assigned to devices whereas in operational studies, participants self-select the device to use when taking the assessment. Using a sample of 488 participants, a 2-wave study was conducted in which participants completed a series of cognitive and noncognitive assessments on a device of their choosing at Time 1 and via paper-and-pencil at Time 2. Evidence of a self-selection effect would be indicated by participants who elected to complete the cognitive ability UIT via a high cognitively demanding device (e.g., smartphones or phablets) scoring lower at both Time 1 and Time 2 than participants who elected to complete the cognitive ability UIT via a low cognitively demanding device (e.g., tablets, desktops and laptops). Contrary to what was expected, the obtained pattern of results was not supportive of a device self-selection effect. Implications and limitations are discussed.

ACKNOWLEDGEMENTS

Completing this thesis was no easy task, and I would not have been able to accomplish it without the support of so many people along the way. First and foremost, all thanks be to God for continuing to sustain me throughout this entire process, even when I doubted myself.

I would like to thank my committee members Dr. Payne and Dr. Blanton for their time and insight and for going through this process with me, and I would like to thank my committee chair Dr. Arthur especially for his mentorship, for not giving up on me, and for continuing to be patient with me every step of the way.

Many thanks with love to my friends for encouraging me and reminding me that I was capable of meeting this goal at many points when I doubted myself, and many thanks to my family for believing in me and encouraging me, even when I did not believe in myself.

I would like to thank my counselors and other mental health professionals that helped me to move through difficult moments and keep going on this academic journey, and I would like to thank all the faculty in the I/O area of the Psychological and Brain Sciences Department for taking a chance and giving me this opportunity.

Last but surely not least, with a heart full of love and gratitude, I would like to thank my mom and dad who went through the fire, wind, and rain with me and never once doubted my ability.

Thank you to everyone who helped me through this stage of my academic and career journey. I am truly, forever grateful.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a thesis committee consisting of the student's committee chair Professor Winfred Arthur, Jr., and committee members Professor Stephanie C. Payne of the Department of Psychological and Brain Sciences and Professor Hart Blanton of the Department of Communication.

Funding Sources

The thesis study was supported by departmental funds from the student's faculty advisor.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

The increasing role of technology continues to shape and reshape the structure and functionality of organizations. Consequently, efforts to understand how these new innovations affect the world of work continue to be both challenging and necessary. Within the context of personnel selection, unproctored internet-based tests (UITs) have become an increasingly prevalent tool for measurement and assessment. As such UIT has sparked a wave of research interest over the last decade or more.

Questions and concerns accompanying the introduction of UIT to the assessment world were very similar to those following the introduction of computer-based assessments more broadly. Primary concerns largely pertained to the measurement equivalence of computer-based assessments to traditional paper-and-pencil methods (Parker & Meade, 2015). Over time, evidence has accumulated to support the measurement equivalence of computerized methods, and the use of computer-based assessments has become mainstream for many organizations (Ryan & Ployhart, 2014). With increased accessibility to the internet, concerns soon shifted to a focus on UIT in relation to proctored assessments.

Soon after its introduction, UIT quickly grew in popularity with organizations because it allowed many employers to access larger pools of job applicants and also reduced the amount of time it took to conduct applicant testing and assessments (Tippins, 2015). Accompanying these advantages, however, were a number of concerns regarding the validity and accuracy of UIT scores resulting from opportunities for malfeasant behaviors, specifically cheating (Arthur et al., 2010; Beaty et al., 2011; Tippins, 2009). Despite concerns, Arthur et al. (2010) did not find any evidence for score differences on cognitive and noncognitive ability tests due to malfeasance in unproctored settings which differed from that reported in proctored settings. Using a within-subjects design, participants completed a speeded cognitive ability test and personality measures at two time points. For the cognitive ability measures, the pattern of score differences resembled that of a practice effect rather than cheating. Arthur et al. (2010) suggested that the pattern of results was likely a function of the speeded time constraints impeding the ability to cheat. Steger et al. (2018) conducted a meta-analytic review of the unproctored assessment literature and obtained differences in proctored and unproctored test performance with higher scores for unproctored assessments. This effect was moderated by the feasibility of using the internet to search for test answers, a finding that is consistent with Arthur et al.'s (2010) use of speeded testing as a means to mitigate cheating on UITs.

Although researchers continue to study the use of UIT for selection assessment, rapid advancements in technology have led to a shift in focus in recent years from issues regarding UIT more generally to issues concerning the devices used to complete said assessments. Due to the unproctored nature of UIT, test takers are no longer limited to completing UITs onsite via desktop computers. What was once an electronic assessment accessible only via wired desktop connection is now accessible via a variety of internet compatible devices. With the growing number of people owning devices such as smartphones, tablets, and phablets, employers are witnessing the completion of UITs on a wide range of devices (Arthur et al., 2017; Arthur et al., 2014; Arthur, Keiser, & Doverspike, 2018; McClure & Boyce, 2015).

It is important to note, for the purpose of the discussion of past research, that internet compatible devices have often been classified as mobile and nonmobile devices based on their ease of portability or “connectedness” to the wall (Arthur, Keiser, & Doverspike, 2018). Under this distinction, mobile devices often refer to smartphones and phablets, and sometimes tablets and laptops as well, whereas nonmobile devices usually refer to desktop computers. For the purpose of the discussion of relevant work, the terms used in the literature will be used to describe referenced works. However, in relation to the present study, these devices will be referred to as high (e.g., smartphones or phablets) and low (e.g., tablets, laptops, and desktops) cognitively demanding devices based on Arthur, Keiser, and Doverspike's (2018) structured characteristics/information processing-based framework. This framework is discussed in more detail in later sections.

Device Usage Effects

The overall percentage of job applicants using mobile devices to complete UITs is low compared to nonmobile device users, however these numbers have continued to increase from year to year (Arthur, Keiser, & Doverspike, 2018). Lawrence and Kinney (2017) reported an increase in smartphone use of 11%, from 6% in 2012 to 17% in 2015. Similarly, Golubovich and Boyce (2013) reported an increase from 3.1% to 14.3% between 2009 and 2013. Consistent with this increasing trend, data obtained from a major industrial-organizational psychology assessment consulting firm indicates that in 2020, 56.4% of job applicants for entry-level positions completed UITs using smartphones, phablets, and tablets, with a smaller percentage, 17.6%, for professional-level positions (S. Jarrett,[1] personal communication, December 31, 2020). These percentages have steadily increased from 7.8% for entry level applicants and 1.4% for professional-level positions in 2012. With growing access to various mobile internet compatible alternatives, these trends are likely to continue.

[1] Dr. Steven Jarrett holds the title of Director, Manufacturing COE at PSI Services, a major internet testing and assessment firm. PSI services delivers more than 30 million assessments per year (PSI Services, 2021).

From 2011 to 2021, the number of Americans owning smartphones increased from 35% to 85% (Pew Research Center, 2021). For many of these smartphone users, their smartphone is their primary means of internet access in the absence of a broadband connection at home (Pew Research Center, 2021). Thus, permitted use of mobile devices for selection assessment presents an opportunity to broaden the pool of job applicants for many organizations. It may also serve to increase minority representation as members of Latinx and African American communities represent a large percentage of those who are smartphone dependent (own a smartphone without a broadband connection at home; Pew Research Center, 2021). Because 27% of U.S. adults earning less than $30,000 a year are also among the most smartphone-dependent (Pew Research Center, 2021), mobile assessment may serve to increase the number of low socioeconomic status (SES) job applicants in the applicant pool.

In summary, as evidenced by a number of statistics, mobile assessment may be beneficial in a number of ways for both job applicants and employers, however the variance in the types of devices that can be used to complete tests has led researchers to question and examine the reliability and validity of test scores across testing devices (Tippins, 2015). The results of studies examining the effects of device type on UIT scores have been mixed with some studies obtaining cognitive mean score differences as a function of the device used, whereas others have not (Arthur, Keiser, & Doverspike, 2018). Specifically, operational studies have consistently found that participants who complete cognitively demanding assessments on high cognitively demanding devices score lower than individuals who complete cognitively demanding assessments on low cognitively demanding devices. However, no device-type effects have been observed for non-operational studies. Although several explanations have been advanced to account for these observed effects, they have received limited empirical examination (Arthur et al., 2014). Consequently, the present study sought to empirically examine one of these explanations, specifically, the self-selection hypothesis, as a viable explanation for device-type differences on cognitive UIT assessments.

Proposed Explanations

As previously noted, a cursory review or reading of the literature would suggest that cognitive mean score differences for UITs occur as a function of the type of device used. However, a more thorough examination of this literature indicates that the resulting device-type differences covary with the setting of the study (Arthur, Keiser, & Doverspike, 2018). Of the studies conducted, differences have only been observed for operational studies—studies conducted for the purpose of organizational decision making like personnel selection (e.g., Arthur et al., 2014; Impelman, 2013; Wood et al., 2015). These differences have not been observed in non-operational (lab) studies (e.g., Arthur Keiser, Hagen, & Traylor, 2018; Brown & Grossenbacher, 2017; Grelle & Gutierrez, 2019; Parker & Meade, 2015; Traylor et al., 2021). To better assess the magnitude of cognitive and noncognitive mean score differences in operational and non-operational settings, we conducted a small-scale bare bones meta-analysis to quantitatively summarize the results of Arthur, Keiser, and Doverspike's (2018) qualitative review.[2] This meta-analysis also expanded on the reports included in Arthur, Keiser, and Doverspike's qualitative review by updating the literature search to include reports that had been published or presented at conferences since their review was conducted. All studies included in the analysis and sample-weighted mean *ds* by research setting are reported in Table 1. A complete summary of the meta-analysis results is presented in Appendix A.

[2] This meta-analysis was concurrently published in Traylor et al. (2021).

Echoing the results of Arthur, Keiser, and Doverspike (2018), the meta-analysis revealed a relatively large effect (*d* = 0.79) for cognitive mean score differences in operational settings with participants using desktops, laptops and tablets scoring higher on cognitively demanding assessments than participants using phablets or smartphone devices. Also, in accordance with Arthur, Keiser, and Doverspike, the *d*s for noncognitive test scores across study settings and cognitive test scores in non-operational settings approached zero. The pattern of results would suggest that there is some differentiating feature pertaining to operational and non-operational studies that may explain or account for the differences in findings.

Table 1

List of Studies and Sample-Weighted Mean Differences (ds) by Study Type for Studies Examining the Effects of Device Type on Cognitive and Noncognitive UIT Scores

Study Type			
Operational		**Nonoperational**	
Arthur et al. (2014)		Arthur et al. (2018)	
Dages et al. (2017)		Brown & Grossenbacher (2017)	
Illingworth et al. (2015)		Chang et al. (2016)	
Impelman (2013)		Fursman (2016)	
LaPort (2016)		Grelle & Gutierrez (2019)	
Lawrence et al. (2017)		King et al. (2015)	
Lawrence et al. (2013)		Lawrence et al. (2016)	
Wood et al. (2015)		Martin et al. (2020)	
		Parker & Meade (2015)	
		Smeltzer (2013)	
Sample-Weighted Mean Differences (*d*s)			
Cognitive	0.79 (15)	Cognitive	0.05 (14)
Noncognitive	-0.02 (64)	Noncognitive	-0.02 (20)

Note. Positive *d*s reflect higher scores for low cognitively demanding devices. *k*s are in parentheses (Reprinted from Traylor et al., 2021).

Arthur et al. (2014; Arthur, Keiser, & Doverspike, 2018) draw attention to three characteristics differentiating the two study settings that may provide a plausible explanation for the observed effects. Each of the study conditions differ in the type of environment in which testing occurs, the testing stakes or consequences of the test, and the means by which participants are assigned to device-type conditions (assigned by the researcher or participant self-selection).

Non-operational studies for which device-type differences have not been observed are characterized by a relatively high degree of environmental control. They have generally been low stakes with little if any consequences associated with participants' test performance, and participant self-selection was not a factor in the assignment to device-type conditions for these studies.

In contrast, operational studies for which device-type differences have been observed are characterized by a relatively low degree of environmental control, and thus have allowed for variance in the types of environments in which UIT tests could be completed. Participants in these studies were drawn from samples of job applicants undergoing real selection processes. Thus, these studies can be further characterized as high stakes, as there are valued employment outcomes associated with the outcomes of the assessments. Finally, operational studies have allowed participants to self-select the device on which they chose to complete the test. Within this context, participants' choice of device represents another possible source of variance.

Each of these differences has led to the development of three explanations for the differential occurrence of mean score differences on cognitive UITs. The following sections discuss these three explanations.

Testing Environment

The use of UIT in conjunction with the introduction of different types of internet compatible devices has led to variance in the types of environments in which UITs can be completed. Since the introduction of UITs, the testing environment has been a general area of concern, primarily as it relates to applicant reactions, standardization, and malfeasant behavior

(Tippins et al., 2006; Wasko et al., 2015). Although research suggests that the testing environment does not affect a test's psychometric properties, it nevertheless can affect performance on the test (Petor et al., 2017). Some explanations for observed device-type differences on UITs have centered on these environmental differences and the possibility for distractions as a function of the device and the environment.

In their study specifically examining differences in cognitive test scores for mobile and non-mobile devices, Arthur et al. (2014) mentioned the environment as a possible explanation for the observed differences in performance, suggesting that individuals completing UITs on mobile devices are subject to more distractions in the environment. With mobile devices, there is both the opportunity to encounter distractions in the overall testing environment, as well as the possibility of encountering distractions due to the various, and often used functions of mobile internet compatible devices. In addition to serving as a medium for connecting to the internet, smartphones and phablets are also common methods of voice, video, and text communication. Therefore, it is possible that test takers could encounter distractions such as phone calls, text messages, or other media notifications while completing UITs. Gutierrez et al. (2015) found that distractions and interruptions were an issue with mobile device testing in unproctored environments due to phone calls and pop-up notifications occurring during the test. Likewise, Lawrence et al. (2017) found that receiving phone calls and having technical issues were among the most frequent in terms of disruptive events occurring during UITs. Although, some if not many of these occurrences are possible with a variety of different types of UIT devices, it is not unreasonable to expect that such occurrences are greater when using mobile as opposed to non-mobile devices (Chang et al., 2016). Subsequently, in line with this reasoning, increased distractions in the environment would lead to decreased test performance. In support of this proposition, Sinar and Wasko (2008) found a positive relationship between the favorability of one's environment and performance on a selection assessment. Likewise, they found that the most favorable unproctored environment was one in which test takers could complete the assessment alone on a computer with a high speed internet connection. Thus, the evidence would suggest that unfavorable environments with more opportunities for distractions may be negatively related to test performance.

In addition to the aforementioned concerns, some scholars suggest that there are many ways in which the design and function of various UIT devices may effect one's ability to use the device in a given testing environment (Arthur, Keiser, & Doverspike, 2018; Arthur, Keiser, Hagen, & Traylor, 2018). Essentially, certain aspects of the device serve to either increase or decrease the extent to which a device readily lends itself to completing assessments. The ways in which the design of UIT devices impact their use for cognitive tasks is outlined in the structural characteristics/information processing (SCIP) framework (Arthur, Keiser, & Doverspike, 2018).

The Structural Characteristics and Information Processing Framework

The SCIP framework was developed to provide a theoretical model to aid in the psychological classification of devices (Arthur, Keiser, & Doverspike, 2018). The framework conceptualizes UIT devices in terms of structural characteristics contributing to construct irrelevant information processing demands. Contrary to previous studies that appear to have conceptualized devices in terms of their "connectedness" to the wall using labels such as "mobile" and "non-mobile", the SCIP framework conceptualizes the mobile/non-mobile distinction by creating a continuum on which all UIT devices fall. Demands imposed by variance in characteristics such as (a) screen size, (b) screen clutter, (c) the response interface, and (d) permissibility of the device contribute to construct irrelevant cognitive load for the test taker.

Devices are thus characterized by the amount of construct-irrelevant cognitive load that they engender. The framework further groups devices according to the theorized similarity in outcomes as a function of using the devices (see Figure 1). Using this framework, devices such as desktops, laptops, and tablets would fall on the lower end of the spectrum inducing the least amount of cognitive load with similar outcomes. Likewise, devices such as smartphones and phablets would appear on the higher end.

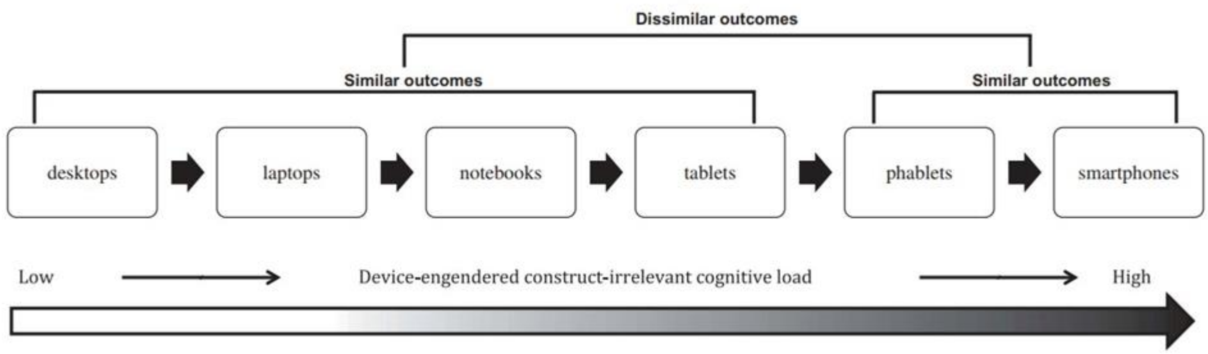

Figure 1. Device-engendered construct irrelevant cognitive load diagram. This figure illustrates how unproctored internet-based testing devices might be classified based on device-engendered construct irrelevant cognitive load and hypothesized outcome similarity (Reprinted from Arthur, Keiser, & Doverspike, 2018).

Because the SCIP framework provides a more device inclusive and study relevant method for classifying UIT devices, the present study classifies UIT devices as a function of the cognitive load that each device is likely to engender. So, henceforth, desktops, laptops, notebooks, and tablets will be referred to as low cognitive demand devices, and smartphones and phablets will be referred to as high cognitive demand devices. Furthermore, in the context of the SCIP framework, the role of the testing environment in device-type effects resides in the proposition that, using smartphones as an exemplar of a high cognitive demand device, smartphones provide the test taker with the greatest degree of freedom in terms of where assessments can be completed. So, if test takers who choose to use smartphones to complete employment-related tests choose to do so in environments that pose additional construct-irrelevant attentional demands (e.g., high levels of distractions), these additional demands should compete for cognitive resources (by making it more difficult to concentrate or focus on the assessment) and subsequently attenuate their test performance (Arthur, Keiser, & Doverspike, 2018; Traylor et al., 2021).

Although the testing environment hypothesis provides a reasonable explanation for observed device-type effects, it has received limited empirical examination. Traylor et al. (2021), the only study of which we are aware that has sought to test this hypothesis in a non-operational setting, did not find an effect of the environment on test performance after having participants complete cognitive UITs via smartphones and desktops in distracting and non-distracting environments. These results may be in part due to the age restricted sample used in this study (Traylor et al., 2021). Participants were college students (ages 18-26) of a generation accustomed to using smartphones in a variety of environments. Thus, the device familiarity and frequency of use that characterizes this generation, coupled with the absence of age-related declines in selective attention (and thus susceptibility to distraction) that would have been observed with an older sample may have attenuated the effects of environmental distractions (Traylor et al., 2021).

Testing Stakes

UIT studies classified as operational or non-operational have differed not only in terms of the degree of control over the testing environment, but also in the consequences that could result as a function of performance on the tests. In operational studies, participants have consisted of job applicants who stood to gain or lose something of value as a function of their performance on UITs. To the contrary, non-operational study participants have usually consisted of student volunteers for whom test consequences have been trivial at best. Highlighting these differences, the testing stakes hypothesis posits that the ability to detect device-type differences for UITs varies as a function of the testing stakes. The underlying premise suggests that differences in testing stakes or consequences will lead to differences in testing motivation. Likewise, differences in testing motivation will lead to test results which misrepresent test taker knowledge and ability. Thus, given that device-type effects have only been observed under high stakes testing conditions, the testing stakes hypothesis suggests that cognitive mean score differences are likely to be observed when the stakes are high enough to increase test taker motivation and effort.

Within the testing stakes literature, there is much support for the relationship between testing stakes and test performance. Numerous studies to date have examined this relationship primarily in educational settings with a focus on achievement testing (tests used to estimate a test taker's level of mastery of some domain [Wise & Smith, 2011]). There is an overwhelming consensus that low stakes tests suffer from issues with testing motivation. From their review of the literature, Wise and Demars (2005) concluded that most studies are in support of motivation as a positive predictor of test performance. These findings have led researchers and practitioners to question the accuracy of test scores generated under low-stakes conditions. Due to a lack of test taker motivation, test scores from low-stakes tests may not accurately represent a test taker’s ability in a given area (Wise & Demars, 2005). Consequently, resulting inferences made based on test scores from low stakes tests run the risk of not being valid (Wise & Demars, 2005).

The relationship between testing stakes, motivation, and performance is best explained via the expectancy value model of achievement motivation (Eccles, 1983; Wigfield & Eccles, 2000; Wise & Demars, 2005). The expectancy value model describes test taker effort in terms of test taker expectancies and test taker values. Expectancies refer to an individual's appraisal of their potential success in performing a task as well as their beliefs regarding their ability to perform the task. Values consist of four components: (a) attainment value, (b) intrinsic value, (c) utility value, and (d) cost. These four components, respectively, describe an individual's perception of the importance of the task, their enjoyment in performing the task, the usefulness of the task in relation to future endeavors, and the sacrifices one must make in order to complete the task. Testing with low stakes creates a situation in which expectancies may be comparable to those of high-stakes tests, with much lower values. Therefore, as the theory would suggest, effort applied to one's performance under low stakes conditions would be inferior to effort under high stakes conditions, resulting in lower performance on the assessment overall. Although, expectancies and value beliefs will vary from test taker to test taker, the idea is that a high stakes situation would be of much higher value to most and would lead to higher effort and maximal performance (Barry et al., 2010). Thus, with all participants performing at a high capacity in terms of effort exertion, high stakes environments would lead to results which better represent participants' ability on cognitive assessments.

In summary, although there is research that speaks to the relationship between testing stakes and test performance, this literature does not provide much, if any guidance as to a testing stakes and device-type effect on cognitive assessments. The present study also does not examine this relationship because it is a lab study with relatively low stakes. Nevertheless, the potential effects of testing stakes were controlled by holding it constant. In addition, because the viability of the self-selection hypothesis is predicated on the premise that test takers are motivated to perform well, a high stakes assessment was simulated in the present study by paying performance-based rewards based on selection ratios that one would typically encounter in operational contexts.

Self-selection

In addition to possible environmental differences and differences in testing stakes, self-selection has been proposed as the third explanation for observed cognitive mean score differences in UIT (Arthur et al., 2014; Arthur, Keiser, & Doverspike, 2018; Brown & Grossenbacher, 2017). When compared to operational studies, non-operational studies differ in the method of participant assignment to device-type conditions. As is the nature of non-operational studies, participants in lab studies for which no device-type differences have been observed, have been assigned to, rather than self-selected the type of device on which to complete the assessment. For example, in their study examining differences in scores for individuals completing cognitive and noncognitive assessments on mobile and nonmobile devices, Arthur et al. (2014) found that test scores were significantly lower for individuals who elected to complete the assessment on mobile as opposed to nonmobile devices. This study used an operational sample in which participants chose the device on which to complete the assessment. Other studies (Arthur, Keiser, Hagen, & Traylor, 2018; Brown & Grossenbacher, 2017; Traylor et al., 2020) examined this same relationship in non-operational settings in which participants were assigned to device-type conditions by the researcher and did not find significant differences in cognitive ability scores between the device-type conditions. By randomly assigning individuals to device-type conditions, several construct irrelevant effects, including possible selection effects were diminished.

Because the method of assignment to device-type conditions covaries with the observance of UIT cognitive mean score differences, the self-selection hypothesis posits that cognitive ability differences between those who choose to complete high-stakes cognitively-demanding assessments on smartphones (a poor decision) versus those who use devices more conducive to testing (e.g., laptops, desktops) may account for the observed differences. Thus, this explanation simply posits that individual differences in cognitive ability may account for the device-type effects observed for cognitive assessments in operational settings. This is because in operational settings participants not only get to choose where and when they complete assessments but also the device on which said assessments are completed. This is not the case in non-operational studies, because in these settings participants have been randomly assigned to device conditions. Consequently, the hypothesis advances the proposition that there is a relationship between cognitive ability and device-type selection such that individuals lower on the cognitive ability spectrum may be more likely to complete a high stakes cognitive ability assessment using a device higher in cognitive demand than individuals higher on the cognitive ability spectrum. This choice may be reflective of a poor decision, which may have resulted from the test taker being less cognizant of the potential consequences of using the device. In turn, the hypothesis also advances that individuals higher in cognitive ability would make better decisions in terms of assessment device—choosing to complete a high stakes cognitive assessment on a device more conducive to testing (i.e., a low cognitively demanding device).

As outlined in the SCIP framework, there may be some advantage from a cognitive load perspective to using devices such as desktops and laptops to complete cognitive assessments (Arthur, Keiser, & Doverspike, 2018). In contrast, given the additional construct-irrelevant cognitive load that they engender, devices such as smartphones may be less conducive to test taking in a high-stakes testing environment. Therefore, choosing to complete an assessment on a device that is less suitable for testing would represent a bad decision on behalf of the test taker.

Consonant with the self-selection hypothesis, the decision-making literature indicates that there is a relationship between cognitive ability and decision making such that individuals higher in cognitive ability are likely to engage in better decision making (Bruine de Bruin et al., 2007; Bruin de Bruin et al., 2020; Frederick, 2005; Jackson et al., 2017). Because we are unaware of any studies that have attempted to isolate the relationship between test taker cognitive ability, device choice, and performance, the present study seeks to advance the understanding of how differences in UIT devices result in differences in cognitive ability scores by isolating the effects of device self-selection.

The Present Study

To recapitulate, the variety of devices that can be used to complete UITs present a great opportunity in terms of accessibility for job applicants and diversification of the job applicant pool for employers. However, the present study recognizes some discrepancies in previous findings regarding test score means across testing devices and seeks to examine whether observed differences are the result of the type of device or a self-selection effect whereby device selection is influenced by test-taker individual differences in cognitive ability.

Brown and Grossenbacher (2017) claimed to have addressed the issue of selection effects with their study examining the effects of device type on general mental ability (GMA) scores. Brown and Grossenbacher controlled for selection effects by randomly assigning participants to device-type conditions to complete the parallel form of a baseline GMA assessment. Counter to their claim, by design, controlling for selection effects negates the ability to make causal inferences pertaining to the effects of device selection. The present study addressed these design issues and the aforementioned objectives by using a mixed design to assess test taker performance on both cognitive and noncognitive constructs. At Time 1 (device self-selection), the first repeated measures condition, participants completed a series of unproctored cognitive and noncognitive assessments on a UIT device of their choosing. At Time 2 (controlled paper-and-pencil), participants completed the cognitive and noncognitive assessments once more in a controlled lab setting via paper-and-pencil. To isolate the effect of device self-selection, testing stakes were held constant across each of the within-subjects testing conditions at the high stakes level by offering performance-based rewards. The uniform testing environment across the between-subjects, self-selected device type conditions at Time 2 (controlled paper-and-pencil) served as a control for the effects of the testing environment.

A strength of this design is that it allowed participants to self-select into device type groups and provided a standard with which to compare GMA performance across groups. This design permitted isolation of the relationship between GMA and device choice by comparing performance within subjects under a high discretion, UIT device condition and a standardized paper-and-pencil (controlled, no discretion) condition. Although beyond the focus of the present study, the design also permitted inferences about the effects of device-type/device characteristics on cognitive ability outcomes.

The self-selection hypothesis posits that individual differences in cognitive ability translate into differences in device choice that result in performance differences on cognitive ability assessments. This is predicated on the relationship between cognitive ability and decision making such that individuals higher in cognitive ability are likely to make better decisions than individuals lower in cognitive ability (Frederick, 2005; Jackson et al., 2017). Therefore, the self-selection hypothesis asserts that differences in cognitive ability are related to an individual's decision to complete a cognitive ability assessment on a device more (a low cognitively demanding device) or less (a device higher in cognitive demand) conducive to testing. These differences relate to device-type selection such that participants on the lower end of the cognitive ability spectrum would be more likely to elect to complete UITs on a high cognitively demanding device than individuals on the higher end of the cognitive ability spectrum. Choosing to use a device on the higher end of the cognitive demand spectrum would be reflective of poor decision making or a poor choice because whereas on the surface the use of a smartphone, for example, would appear to be more convenient presenting the test taker with greater degrees of freedom on several dimensions (e.g., where and when), this advantage masks its gross disadvantage as a suboptimal test taking platform (Arthur, Keiser, & Doverspike, 2018; Arthur & Traylor, 2019). Therefore, to the extent that cognitive mean score differences are the result of a self-selection effect, it was hypothesized that:

> ***Hypothesis 1.*** Participants who elect to complete the cognitive ability UIT via a high cognitively demanding device will score lower at both Time 1 (device self-selection) and Time 2 (controlled paper-and-pencil) compared to participants who elect to complete the cognitive ability UIT via a low cognitively demanding device.

Previous studies examining mean score differences between devices have consistently observed differences for cognitive assessments and no differences for noncognitive assessments (Arthur et al., 2014). Noncognitive assessments are less cognitively demanding than cognitive assessments, therefore any excess construct irrelevant cognitive demand imposed by the device should not pose a threat to noncognitive test scores (Arthur et al., 2014). To the extent that the results of previous studies are replicated (previous empirical examinations have not found any device type effects in both operational and non-operational contexts for noncognitive constructs), it was hypothesized that:

***Hypothesis 2*.** The pattern of results that is supportive of a self-selection explanation will not be observed for the noncognitive constructs.

In summary, to the extent that the self-selection hypothesis is supported, the expected results will be as illustrated in Figure 2. Likewise, to the extent that the results of the present study reflect those of past studies, the expected results for noncognitive constructs will be as illustrated in Figure 3.

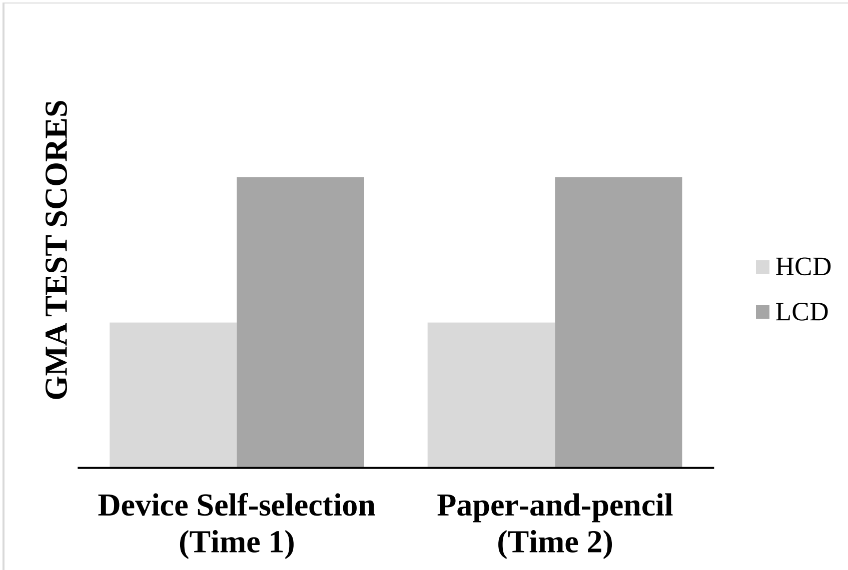

Figure 2. Expected pattern of results for Hypothesis 1 (self-selection effect). HCD = Participants who completed the UIT on a high cognitively demanding device under the device self-selection condition; LCD = Participants who completed the UIT on a low cognitively demanding device under the device self-selection condition

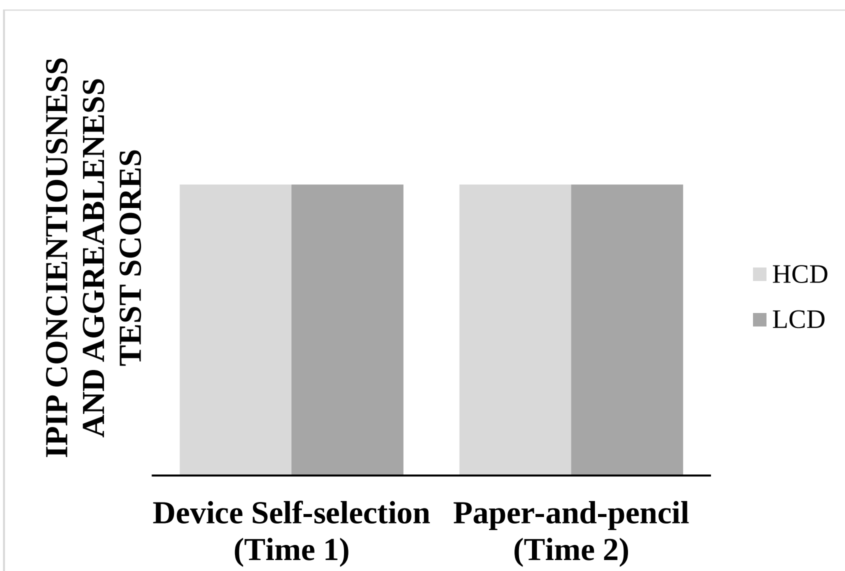

Figure 3. Expected pattern of results for Hypothesis 2. HCD = Participants who completed the UIT on a high cognitively demanding device under the device self-selection condition; LCD = Participants who completed the UIT on a low cognitively demanding device under the device self-selection condition

2. METHOD

Participants

A power analysis was conducted to estimate the sample size needed to achieve a small to medium effect with an alpha of .05 and power of .80 (Erdfelder et al., 1996; Faul et al., 2009). Using the most conservative hypothesis (Hypothesis 1), an a priori power analysis for a 2 (device type) × 2 (device selection discretion) mixed analysis of variance (ANOVA) was conducted. Effect sizes were based on those drawn from literature examining cognitive mean score differences on UITs (Arthur, Keiser, Hagen, & Traylor, 2018; Traylor et al., 2021). Results indicated that a sample of 788 participants would be necessary to detect a d of 0.10 as statistically significant ($p < .05$) at a power of .80. A slightly larger effect of $d = 0.20$ would require a sample of 200 participants, and a medium effect of $d = 0.40$ would require a sample of 52 participants.

Given these boundary conditions, coupled with the time and resources available for data collection, data were collected for a sample of 496 participants (which translates into a power of .80 to detect a d of 0.13). A total of 496 participants completed both the device self-selection (Time 1) and paper-and-pencil (Time 2) repeated measures condition assessments, however 8 participants were removed for failing to report the type of device used to complete the UIT in the device self-selection condition, resulting in a sample of 488 participants.

Documented in the extant literature are observed relationships between race/ethnicity and device use, device ownership, and performance on measures of cognitive ability (Arthur et al., 2014; Pew Research Center, 2021; Roth et al., 2001). Consequently, race/ethnicity was included as a statistical control for analyses pertaining to GMA. Ethnic/racial subgroup differences in device usage, device ownership, and cognitive ability performance reported in the literature most often reference differences between White/non-Hispanic and Black/African American and Hispanic racial/ethnic groups. Consequently, the sample was further limited to only participants belonging to one of these three racial/ethnic groups, resulting in a final sample of 383 participants—accounting for 78% of the initial sample of 488 participants. This final sample size translated into a power of .72 to detect a small effect of $d = 0.13$ using the most conservative hypothesis (i.e., Hypothesis 1).

Participants were students 18 years or older who participated in the study for course-related research credit and were recruited via the psychology subject pool at a large southwestern U.S. university. The sample was predominantly White/non-Hispanic (70%) and female (65%) with a mean age of 18.9 ($SD = 1.21$). Data were collected over the course of two semesters. Participants were classified as belonging to one of two cohorts based on the semester in which they participated in the study. The percentage of participants in each cohort was practically equal with 50.1% of the participants belonging to the second cohort.

Measures

Cognitive Construct–General Mental Ability

Cognitive ability was assessed via scores on a 60-item (36 verbal and 24 quantitative items), 4-alternative, speeded assessment of generally mental ability (GMA_{60}, Arthur, 2017). Scores were calculated as the number of items answered correctly (max = 60). Arthur (2107) reported convergent validities from .42 to .55 with ACT and SAT scores. Criterion-related validities of .24-.29 were reported for GPA and .32 for supervisor ratings. Test-retest reliabilities of .76 and .70 for two alternate forms of the test with 7 to 10 days between administrations, were also reported (Naber et al., 2021). Sample items for this test, along with the other measures used in the study are presented in Appendix B.

Noncognitive Constructs

Conscientiousness and agreeableness were selected as the noncognitive constructs of interest given their recognition as the two most important personality attributes in the workplace (Sackett & Walmsey, 2014; Schmitt, 2014). Both constructs were assessed using the specified items (10 items each) from the 50-item International Personality Item Pool (IPIP; Goldberg, 1999) which is representative of Goldberg's (1992) Big-Five factor markers. Each item was rated on a five-point Likert scale (1 = very inaccurate; 5 = very accurate) to reflect the extent to which it was descriptive of the test taker. Cronbach's alphas were .82 and .81 for conscientiousness and .80 and .84 for agreeableness for the device self-selection UIT and paper-and-pencil administrations of the conscientiousness and agreeableness measures respectively.

Demographics and Device Characteristics

At the end of the online assessment (device self-selection condition), participants indicated the type and brand of the device used to complete the assessment and the type of environment in which the assessment was completed. Items assessing demographic information including sex, race/ethnicity, and age were included at the end of the in-person paper-and-pencil assessment completed one week later. High school GPA, current GPA, and standardized test scores (i.e., SAT & ACT) were collected as part of the Time 2 paper-and-pencil assessment, and three items assessing the distractibility of the testing environment and motivation to perform well on the assessments were included as well[3]. Specifically, two items were included at the end of the device self-selection UIT and paper-and-pencil assessments to ascertain (1) the extent to which participants found the testing environment to be distracting and (2) the extent to which

[3] Data were collected over the course of two semesters. Due to an oversight, in-person, paper-and-pencil assessments administered during the first semester (*n* = 247) did not include any items assessing the distractibility of the testing environment and motivation of the test taker.

participants found it difficult to concentrate while completing the assessment. An additional item was included at the end of the Time 2 paper-and-pencil assessment to assess the extent to which participants were motivated to perform well on the assessment.

Manipulation Check

Two items assessing participants' knowledge of the performance rewards were included at the beginning of the Time 1 device self-selection assessment, and again at the end of the device self-selection and Time 2 (paper-and-pencil) assessments. Participants were asked to indicate (1) whether there were performance rewards associated with the study, and (2) the amount of the performance rewards if so.

Design and Procedure

The present study used a two-wave design with a one-week interval. During the first wave of the study (i.e., device self-selection condition), participants completed a 15-minute online battery consisting of the GMA_{60} and the conscientiousness and agreeableness subscales of the IPIP. Participants were given the discretion to use any internet compatible device of their choosing in any location of their choosing to complete the online assessments. During the second wave one week later (paper-and-pencil condition), participants completed an in-person proctored assessment via paper-and-pencil[4]. The assessment lasted approximately 1 hour and 45 minutes and included the GMA_{60} measure and the IPIP conscientiousness and agreeableness subscales. The time for completion included demographic measures, items measuring environmental

[4] A paper-and-pencil assessment format was used to administer the Time 2 assessment to provide a standard for comparison which did not utilize an electronic device. Although devices like desktop computers are hypothesized to require less cognitive resources to use, there is still the possibility that characteristics of desktop devices may influence the outcomes of the assessment. Paper-and-pencil assessments are the gold standard in selection assessment and are free from internet device-type influences. Consequently, the present study chose to use a paper-and-pencil format as the standard for comparison.

distraction and motivation, and other measures not germane to the present study that were administered as part of a larger data collection effort. The presentation of the GMA_{60} measure and an additional measure of general mental ability, the Raven Advanced Progressive Matrices short form[5] (APM; Arthur & Day, 1994; Arthur et al., 1999; Raven et al., 1985), were counterbalanced by study session for the paper-and-pencil assessment to control for potential order effects.

Upon registering to participate in the study, participants selected a day and time to complete the online assessment and were automatically registered for the in-person portion of the study. Participants were informed during registration that they would be required to complete the in-person portion of the study one week from the start date and time of the online portion of the study. After signing up to participate in the study, participants received an email with the link to complete the online assessment at the specified date and time that they selected. Participants had 48 hours to complete the online assessment before the link became inactive. Participants who did not complete the online assessment within 48 hours were not permitted to participate in the in-person portion of the study.

Large classrooms were used to proctor the in-person assessment, which was administered by two proctors. Each session included from 4 to 45 participants seated with at least one empty desk between them to prevent cheating.

Based on the SCIP framework as illustrated in Figure 1, participants who chose to complete the online assessment using a desktop, laptop, or tablet were classified under the low

[5] The Advanced Progressive Matrices short form (APM) was included in the study protocol as part of a larger data collection effort. So, although it was used in part as a determinant of participant performance-based payment, it was not a variable of interest for the present study. This measure was only administered during the Time 2 portion of the study.

cognitively demanding device-type condition. Conversely, participants who chose to complete the online assessment using a smartphone or phablet were classified under the high cognitively demanding device-type condition.

Two factors of interest in examining the effects of device self-selection on cognitive and noncognitive ability performance were the potential effects of the testing stakes and race/ethnicity of the participants—given the extant literature demonstrates a relationship between these two variables and cognitive ability test performance. To increase participant motivation and better approximate the high-stakes nature of operational study conditions, participants were informed of the opportunity to earn a performance-based reward. Reward decisions were based on the sum of participants' cognitive ability test scores for the online and in-person assessments. Raw scores for the GMA_{60} and the APM were *z*-transformed and summed to create a composite cognitive ability score for each participant. Participants scoring in the top 10% received $120 dollars (selection ratio = .10), and participants scoring in the following 6% received $50 dollars (selection ratio = .06).

To verify that participants were aware of the opportunity to earn a performance-based reward as well as the amount of the reward, manipulation check items were included at the beginning of the online assessment and at the end of both the online and paper-and-pencil assessments. Specifically, participants were asked to indicate whether there were performance rewards and if so, the maximum amount of the reward.

Halfway through the data collection process, it was clear that many of the participants were unaware of the opportunity to earn performance-based rewards. Consequently, changes were made prior to continuing with the second cohort of data collection to increase the saliency of the reward opportunity. As one of the changes, participants were encouraged at the start of the online assessment (device self-selection condition) to use the information on the preceding information screen to assist them in responding to the manipulation check item relating to performance rewards. Additionally, a highlight was added to the section of the information page that mentioned the opportunity to earn performance rewards to make the text more noticeable. As a result, the number of participants in the overall sample (including all racial groups) who passed the manipulation check item (i.e., correctly indicating $120 performance reward) increased from 35% to 90% (see Table 2).

Table 2
Pass Rate for Knowledge of Performance Rewards by Cohort

Sample Cohort	Total Sample	Percentage Passed	Sample for Analysis	Percentage Passed
Cohort 1	248	35%	191	73%
Cohort 2	240	90%	192	97%

Note. Cohort 1 = participants who participated in the study prior to making changes to increase the saliency of the opportunity to earn performance rewards. Cohort 2 = participants who participated in the study after changes were made to increase the saliency of the performance rewards opportunity.

3. RESULTS

Descriptive statistics and correlations for all variables of interest (including variables used in supplemental analyses) are presented in Table 3. Sample sizes for the two self-selected device-type conditions were very unbalanced, with the preponderance of participants choosing devices on the low end of the construct-irrelevant cognitive-demand continuum. Of the 383 participants included in the study analyses, 333 (86.9%) chose to use a low cognitively demanding device, and 50 participants (13.1%) chose to use a high cognitively demanding device.

Although the order in which the two GMA assessments (GMA_{60} & APM) were completed during the in-person assessment was counterbalanced by study session, there were no meaningful mean differences in GMA_{60} scores based on the order in which the assessments were completed, $t(381) = 0.86$, $p > .05$, $d = 0.09$.

In acknowledgement of the relationship between testing stakes and test performance on cognitive ability assessments observed in the extant literature (Wise & Demars, 2005; Wise & Smith, 2011), analyses were undertaken to examine the relationship between scores for two different items (embedded in the device self-selection and paper-and-pencil assessments) used to assess knowledge of performance-based rewards and motivation to perform well on the device self-selection (UIT) and paper-and-pencil GMA_{60} assessments. As is shown in Table 3, there was a significant relationship between passing the manipulation check item and scores on the GMA_{60} assessment for both repeated measures conditions (i.e., those who passed the manipulation check tended to score better than those who did not) [$r = .12$ (device self-selection) and $r = .23$ (paper-and-pencil)]. Likewise, there was a significant positive relationship between motivation to perform well on the assessment and scores on the paper-and-pencil GMA_{60} assessment ($r = .09$).

Although both effects were relatively weak, there was a stronger relationship between knowledge of performance-based rewards and scores on both the device self-selection and paper-and-pencil GMA_{60} assessments. Consequently, knowledge of performance-based rewards was included in the analyses to statistically control for the effects of testing stakes.

Knowledge of performance-based rewards was measured both at the beginning of the device self-selection condition assessment and at the end of the paper-and-pencil assessment. The response data from the first assessment of knowledge of performance-based rewards were used in the following analyses because the correlations with scores on the GMA_{60} assessments [r = .12 (device self-selection) and r = .23 (paper-and-pencil)] were stronger for these data than that for performance rewards data collected at the end of the paper-and-pencil (Time 2) assessment.

In recognition of the relationship between race/ethnicity and cognitive ability test scores reported in the extant literature, similar analyses were conducted to examine the relationship between race/ethnicity and GMA_{60} scores for the device self-selection and paper-and-pencil conditions. There were significant, although relatively small, relationships between race/ethnicity and scores on both the device self-selection UIT (r = .24) and paper-and-pencil (r = .23) GMA_{60} measures such that White non-Hispanic participants tended to score higher on the GMA_{60} assessment than Black/Hispanic participants. Consequently, race/ethnicity was also included as a covariate for each set of analyses measuring the relationship between testing conditions and GMA_{60} performance. It is important to note however that although there were differences in GMA_{60} test scores based on race/ethnicity, the overall percentage of Black/African American and Hispanic participants (6%) who chose to use a high cognitively demanding device to complete the device self-selection assessment was lower than the percentage of White non-Hispanic (14%) participants who chose to do the same.

Table 3

Descriptive Statistics and Correlations Between All Study Variables

Variable	*N*	*M*	*SD*	1	2	3	4	5	6	7
1. Age	379	18.90	1.21							
2. Race/Ethnicity	383			**.10**						
3. Sex	383			**-.17**	-.03					
4. GMA (DSS)	383	36.83	7.15	.06	**.24**	-.08				
5. GMA (Paper)	383	41.59	6.60	.05	**.23**	**-.18**	**.57**			
6. AGREE (DSS)	383	3.99	0.53	.01	-.02	**.26**	.03	-.02	(.80)	
7. CONSC (DSS)	383	3.62	0.61	-.05	**.10**	.07	-.01	.05	.08	(.82)
8. AGREE (Paper)	383	4.00	0.60	-.01	-.08	**.30**	-.03	-.07	**.79**	.02
9. CONSC (Paper)	383	3.64	0.62	.00	**.10**	**.13**	.02	.08	.08	**.87**
10. GPA (high school)	341	4.05	0.60	**-.13**	**.10**	.05	.08	**.14**	-.01	**.13**
11. GPA (college)	223	3.18	0.60	.03	**.11**	.11	**.15**	**.26**	.04	**.30**
12. SAT/ACT	370	1289.27	160.53	.01	**.29**	**-.16**	**.42**	**.32**	.07	.07
13. Extent of distractions (DSS)	295	1.81	0.92	-.01	.05	-.06	-.06	.03	**-.12**	-.07
14. Difficulty concentrating (DSS)	295	1.79	1.00	-.04	.02	-.04	-.05	-.04	-.08	-.09
15. Extent of distractions (Paper)	276	1.42	0.73	.02	.04	**.15**	.10	-.01	.05	**.14**
16. Difficulty concentrating (Paper)	276	1.51	0.91	**.14**	.04	.09	.04	-.06	.01	.05
17. Knowledge of rewards (DSS)	383			-.04	.06	-.03	**.12**	**.23**	-.07	**.09**
18. Knowledge of rewards (DSS)	377			-.05	.01	-.02	**.10**	**.17**	**-.09**	.06
19. Motivation (DSS)	381	3.41	1.02	-.07	.03	.02	.08	**.09**	**.14**	.07
20. Device type	383			.00	.05	.02	-.01	-.07	.03	.00

Note. Bold correlations are statistically significant, $p < .05$ (one-tailed). Cronbach's alpha estimates are presented in parentheses on the diagonal. Race/Ethnicity variable coded as 2 = White, and 1 = Black/ Hispanic. Sex coded as 0 = male (%) and 1 = female (%). Device type coded as 1 = low cognitively demanding device and 2 = high cognitively demanding device. GMA = general mental ability; GPA = grade point average. DSS = device self-selection condition (Time 1); Paper = in-person paper-and-pencil condition (Time 2).

Table 3 (continued)

Descriptive Statistics and Correlations Between All Study Variables Continued

Variable	8	9	10	11	12	13	14	15	16	17	18	19	20
8. AGREE (Paper)	(.84)												
9. CONSC (Paper)	.01	(.81)											
10. GPA (high school)	-.01	**.11**											
11. GPA (college)	.07	**.30**	.08										
12. SAT/ACT	.03	.04	**.11**	**.14**									
13. Extent of distractions (DSS)	-.09	-.03	-.03	-.01	.03								
14. Difficulty concentrating (DSS)	-.06	-.07	.03	.03	.07	**.65**							
15. Extent of distractions (Paper)	.01	**.14**	-.06	.10	.02	-.05	.00						
16. Difficulty concentrating (Paper)	-.06	.09	-.04	.05	.05	.07	.04	**.60**					
17. Knowledge of rewards (DSS)	**-.13**	**.12**	**.13**	**.19**	.06	-.07	.03	.10	.09				
18. Knowledge of rewards (Paper)	**-.12**	**.09**	.04	**.19**	.03	-.01	.06	.05	.02	**.62**			
19. Motivation (Paper)	**.13**	**.09**	.01	.02	-.01	**-.20**	**-.21**	-.07	**-.13**	**.10**	**.09**		
20. Device type	.03	.00	-.04	-.06	.05	-.01	.02	-.05	-.04	-.03	-.03	**-.10**	

Note. Bold correlations are statistically significant, $p < .05$ (one-tailed). Cronbach's alpha estimates are presented in parentheses on the diagonal. Race/Ethnicity variable coded as 2 = White, and 1 = Black/ Hispanic. Sex coded as 0 = Male and 1 = Female. Device type coded as 1 = low cognitively demanding device and 2 = high cognitively demanding device. Abbreviations: GMA = general mental ability; GPA = grade point average. DSS = device self-selection condition (Time 1); Paper = in-person paper-and-pencil condition (Time 2).

Hypothesis Testing

Hypothesis 1 posited that participants who elected to complete the cognitive ability UIT via a high cognitively demanding device would score lower under the device self-selection and paper-and-pencil repeated measures conditions compared to participants who elected to complete the cognitive ability UIT via a low cognitively demanding device. A 2 (device selection discretion: device self-selection vs. controlled paper-and-pencil) × 2 (device type under device self-selection condition: high cognitive demand device vs. low cognitive demand device) mixed factorial ANOVA with repeated measures on the device selection discretion condition was used to test this hypothesis, incorporating race/ethnicity and knowledge of performance rewards as covariates[6]. A main effect of device type in the absence of an interaction would indicate that the results are more supportive of a self-selection explanation. As the results in Table 4 indicate, the main effect for device type was not significant, $F(1, 379) = 1.04$, $p > .05$, $\eta^2 = 0.00$[7], $d = 0.06$, and neither was the interaction, $F(1, 379) = 1.63$, $p > .05$, $\eta^2 = 0.00$. However, the main effect of device selection discretion, although very weak, was significant, $F(1, 379) = 5.35$, $p < .05$, $\eta^2 = 0.00$, $d = 0.07$ a finding that is supportive of the ubiquitous retest effect (Hausknecht et al., 2007; Kulik et al., 1984; Scharfen et al., 2018). Figure 4 presents illustrations of the hypothesized and observed results.

[6] Although race/ethnicity and knowledge of performance rewards were incorporated as covariates, analysis results varied little from those not incorporating race/ethnicity and knowledge of performance rewards as covariates (see Table 4). Most noteworthy was that incorporating race/ethnicity and knowledge of performance rewards as covariates reduced the magnitude of the main effect of device selection discretion.

[7] Eta-squared is non-zero at 5 decimal places.

Table 4
Analysis of Variance Results for the Effects of Device Selection Discretion and Device Type Selected under the Device Self-Selection Condition on GMA Scores

	Without controlling for Race and Performance Rewards			**Controlling for Race and Performance Rewards**		
Variable	***F***	**η^2**	***d***	***F***	**η^2**	***d***
Race/ethnicity	--	--		**$27.86_{(1, 379)}$**	**0.02**	**0.29**
Performance rewards	--	--		**$13.52_{(1, 379)}$**	**0.01**	**0.20**
Device selection discretion	**$76.72_{(1, 381)}$**	**0.04**	**0.41**	**$5.35_{(1, 379)}$**	**0.00**	**0.07**
Device type	$0.70_{(1, 381)}$	0.00	0.06	$1.04_{(1, 379)}$	0.00	0.06
Device selection discretion x Device type	$1.88_{(1, 381)}$	0.00		$1.63_{(1, 379)}$	0.00	

Note. Values in bold represent statistically significant effects, $p < .05$. Race/Ethnicity variable coded as 2 = White, and 1 = Black/ Hispanic.; Device type coded as 1 = low cognitively demanding device and 2 = high cognitively demanding device. Eta-squared is non-zero at 5 decimal places.

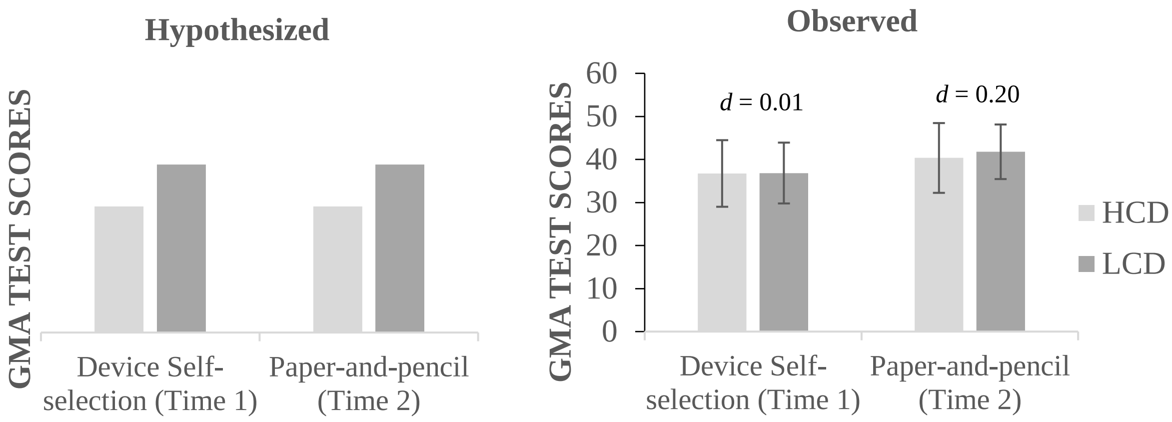

Figure 4. Hypothesized and observed results for general mental ability scores. The expected patterns of results for a self-selection effect (Hypothesis 1) are presented in the left pane. The observed results are presented in the right pane. LCD (low cognitively demanding devices) and HCD (high cognitively demanding devices) were self-selected at Time 1. Error bars represent the standard deviation of the GMA_{60} test scores for each condition. Positive *d* values indicate that scores were higher for the low

Hypothesis 2 stated that the pattern of results that is supportive of a self-selection explanation would not be observed for the noncognitive constructs because there has been no historic precedence to suggest a relationship between device self-selection and noncognitive test scores. Two 2 (device selection discretion: device self-selection vs. controlled paper-and-pencil) × 2 (device type under device self-selection condition: high cognitive demand device vs. low cognitive demand device) mixed factorial ANOVAs were used to test Hypothesis 2, with repeated measures on the device selection discretion variable and scores on the conscientiousness and agreeableness subscales of the IPIP as dependent variables. In support of Hypothesis 2, the device-type × device selection discretion interaction and main effects of device type on the agreeableness and conscientiousness assessment scores were not statistically significant (see Tables 5 and 6).

Table 5
Analysis of Variance Results for the Effects of Device Selection Discretion and Device Type Selected under the Device Self-selection Condition on Agreeableness Scores

Variable	***F***	**η^2**	***d***
Device selection discretion	$0.20_{(1, 381)}$	0.00	0.02
Device type	$0.47_{(1, 381)}$	0.00	0.06
Device selection discretion x Device type	$0.04_{(1, 381)}$	0.00	

Note. Values in bold represent statistically significant effects, $p < .05$.

Table 6
Analysis of Variance Results for the Effects of Device Selection Discretion and Device Type Selected under the Device Self-selection Condition on Conscientiousness Scores

Variable	***F***	**η^2**	***d***
Device selection discretion	$0.86_{(1, 381)}$	0.00	0.02
Device type	$0.00_{(1, 381)}$	0.00	0.00
Device selection discretion x Device type	$0.00_{(1, 381)}$	0.00	

Note. Values in bold represent statistically significant effects, $p < .05$.

Figure 5 presents a graph of the hypothesized results for both noncognitive assessments, and Figure 6 presents the observed results for the conscientiousness and agreeableness subscales of the IPIP.

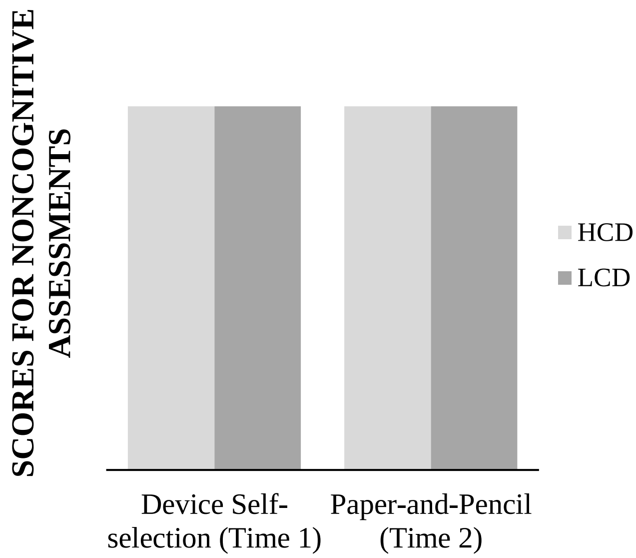

Figure 5. Expected pattern of results for noncognitive test scores (Hypothesis 2). HCD = High cognitively demanding device; LCD = Low cognitively demanding device.

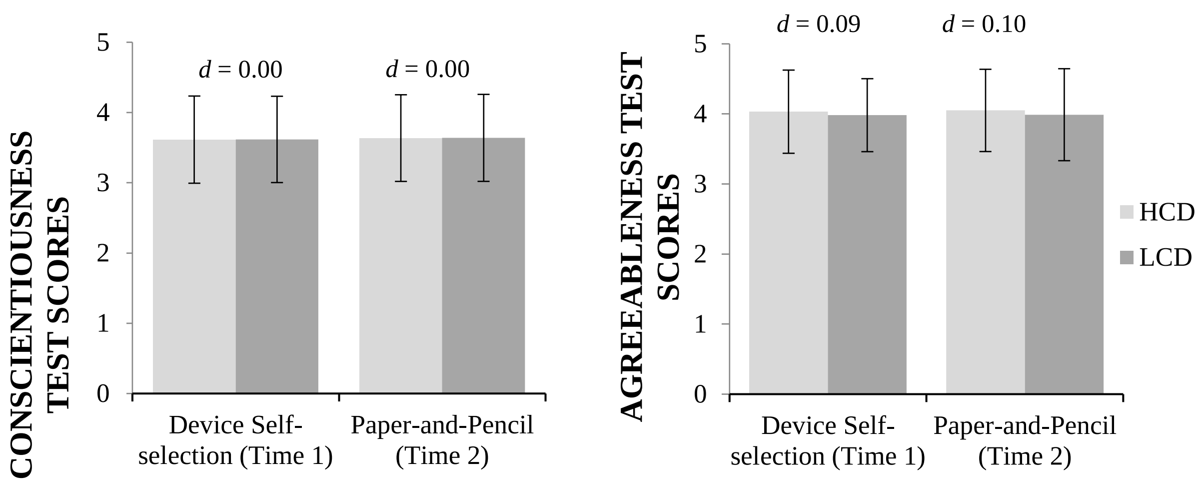

***Figure* 6.** Observed pattern of results for conscientiousness and agreeableness test scores (Hypothesis 2). Error bars represent the standard deviation of the IPIP conscientiousness and agreeableness test scores for each condition. Positive *d* values indicate that scores were higher for the low cognitively demanding device type condition. HCD = High cognitively demanding device; LCD = Low cognitively demanding device.

Although there was no evidence of a self-selection effect, the data lent itself to further exploration of a device-type influence on cognitive and noncognitive outcomes. Contrary to the self-selection effect, the pattern of results that would be supportive of a device-type effect would include an interaction between the type of device used to complete the UIT under the device self-selection condition and device selection discretion (testing via a device of choice or standardized via paper-and-pencil). Specifically, an interaction in which mean differences are observed for scores on the GMA_{60} assessment under the device self-selection condition in the absence of mean differences for scores on the paper-and-pencil GMA_{60} assessment would suggest that the observed mean score differences on the cognitive ability assessments were more likely the result of a device-type effect than an individual difference effect. The pattern of results suggestive of a device-type effect is illustrated in Figure 7. The results of the 2 (device selection discretion: device self-selection vs. paper-and-pencil) × 2 (device type under the device self-selection condition: high cognitive demand device vs. low cognitive demand device) mixed factorial ANOVA presented in Table 4 indicate that the data are not supportive of a device-type effect, as the device-type × device selection discretion interaction was not significant.

To further examine the effects of the testing device on GMA scores, additional analyses were undertaken to examine the relationships between GMA_{60} scores (device self-selection condition – Time 1) from the two device-type conditions and SAT/ACT, high school GPA, and college GPA scores. Descriptive statistics are presented in Table 7, and criterion-related validity estimates are presented in Table 8. As shown in Table 8, the correlations between GMA and high school GPA, college GPA, and SAT/ACT scores were relatively consistent across the two device-type conditions under the device self-selection condition (Time 1). The pattern of results reflects the absence of device-type effects in the predictor/criterion correlations.

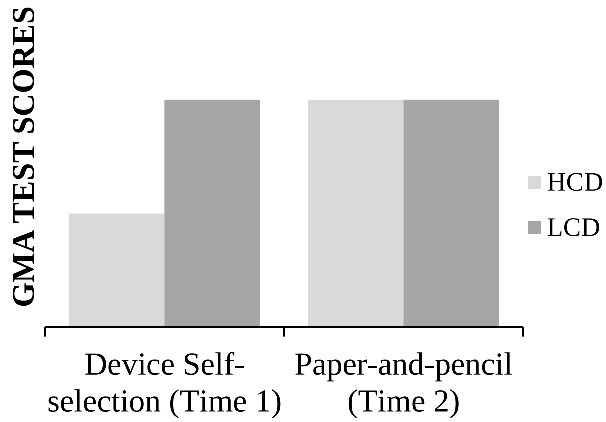

Figure 7. Expected pattern of results for a device-type effect.
HCD = High cognitively demanding device self-selected at Time 1;
LCD = Low cognitively demanding device self-selected at Time 1.

Supplementary Analyses

Supplementary analyses were undertaken to examine the effect of environmental distractions on test scores for the online assessment (device self-selection condition), where participants chose the time, place, and device used to complete the assessment. The results of these analyses echoed those of Traylor et al. (2021). Specifically, the results indicated that there were no relationships between either environmental distraction ratings, $r = -.06$, $p > .05$, or concentration difficulty ratings, $r = -.05$, $p > .05$, and scores on the GMA_{60} assessment under the device self-selection condition. However, there were significant mean differences in ratings of the distractibility of the testing environment, $t(272) = 5.83$, $p < .05$, $d = 0.35$, and difficulty concentrating, $t(272) = 3.94$, $p < .05$, $d = 0.24$, between the digital (device self-selection condition) and controlled paper-and-pencil testing environments such that the testing environment under the device self-selection condition was rated as more distracting and difficult to concentrate in. These differences, however, were not reflected between device-type groups for the device self-selection UIT. Specifically, there were no differences in ratings of environmental distractions, $t(293) = 0.21$, $p > .05$, $d = 0.04$, or difficulty concentrating, $t(293) = -0.31$, $p > .05$, $d = -0.05$, between the high and low cognitively demand device-type conditions. Unsurprisingly, similar to Traylor et al. (2021), there was a significant positive relationship between the extent to which participants perceived the environment as distracting during the device self-selection (Time 1) assessment and the extent to which participants had difficulty concentrating in their testing environment under the device self-selection condition, $r = .65$, $p < .05$. Participants who rated their testing environment as more distracting reported experiencing higher levels of difficulty in concentrating on the assessments under the device self-selection condition than participants who rated their testing environment as less distracting. The same was true for ratings of environmental distraction and difficulty concentrating in the paper-and-pencil testing environment, $r = .60$, $p < .05$.

Table 7
Descriptive Statistics for Criterion-Related Variables: GPA and SAT/ACT Scores

	Full Sample			High cognitive demand device			Low cognitive demand device		
Variable	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
GPA (High School)	341	4.05	0.60	47	3.99	0.60	294	4.05	0.60
GPA (College)	223	3.18	0.60	29	3.09	0.83	194	3.19	0.56
SAT/ACT	370	1289.27	160.53	49	1309.18	133.34	321	1286.23	164.26

Note: GPA = grade point average.

Table 8
Criterion-Related Validity Coefficients for GMA Scores under the Device Self-selection Condition

Variable	*n*	Full Sample	*n*	LCD Device	*n*	HCL Device
GPA (High School)	341	.08	379	.08	47	.08
GPA (College)	223	.15	247	**.12**	29	.25
SAT/ACT	370	**.42**	321	**.44**	49	**.37**

Note: Bold correlations are statistically significant, $p < .05$ (one-tailed). GPA = grade point average; LCD = low cognitive demand; HCL = high cognitive demand.

4. DISCUSSION

As Arthur et al. (2018) note, there are discrepancies in the observance of cognitive ability mean score differences for UITs completed on “mobile” and “non-mobile” devices, such that differences have only been observed in operational settings. Multiple explanations have been proposed to account for this inconsistency. The primary goal of the present study was to determine the extent to which one of those explanations, the device self-selection explanation, accounts for the inconsistency.

Like Arthur et al. (2014), Brown and Grossenbacher (2017) acknowledged the issue of self-selection, which they refer to as selection bias, in operational contexts and suggested based on the results of their study that device-type differences may result from individual differences in cognitive ability between groups. However, the use of random assignment of participants to device-type conditions as part of their study design precluded the ability to isolate the effects of device choice to determine the effects of self-selection on device selection and cognitive ability performance. The present study addressed the preceding issue by isolating the effects of device self-selection. Participants completed cognitive and noncognitive ability UITs under two device selection discretion conditions while controlling for testing stakes and environmental distractions.

Contrary to what was expected, the results were not supportive of a device self-selection effect. Specifically, the results failed to support Hypothesis 1 which stated that participants who elected to complete the cognitive ability UIT via a high cognitively demanding device would score lower under both the device self-selectin condition (Time 1) and the paper-and-pencil condition (Time 2) than participants who elected to complete the cognitive ability UIT via a low cognitively demanding device. Although participants chose the type of device used to complete the device self-selection assessment, the results echoed those of other non-operational studies (see Arthur et al., 2018, Traylor et al., 2021), such that no mean differences were observed for cognitive ability scores based on the type of device used to complete the assessment.

Even so, there was a main effect of device selection discretion such that cognitive ability test scores were higher under the paper-and-pencil condition in which participants completed the same form of the GMA_{60} assessment via paper-and-pencil one week following the initial administration of the assessment using any device of their choosing. However, as previously noted, the observed differences are representative of a retest effect. Researchers have shown that test performance increases for subsequent iterations of cognitive ability assessments, with larger gains observed when identical forms of the test are used for each administration of the assessment (Hausknecht et al., 2007; Kulik et al., 1984; Sharfen et al., 2018).

Contrary to cognitive constructs, no evidence of device-type differences has been reported for scores on measures of noncognitive constructs in operational and nonoperational contexts. As presented in Table 1, standardized mean difference estimates (*d*s) for scores on noncognitive measures approximated zero for scores obtained from operational (d = -0.02) and nonoperational studies (d = -0.02). Consonant with these findings, the pattern of results obtained for the noncognitive constructs used in the present study (i.e., conscientiousness and agreeableness) supported Hypothesis 2, which stated that the pattern of results that is supportive of a self-selection explanation would not be obtained for noncognitive constructs. Specifically, noncognitive mean scores did not vary based on the type of device used at Time 1 or the level of device selection discretion.

In summary, the data were not supportive of a self-selection effect for cognitive mean scores, however—as predicted—the data supported the literature pertaining to noncognitive constructs. A probable explanation for the observed effects for Hypothesis 1 centers on differences in operational study samples and the student sample used in the present study. This explanation suggests that there are individual difference factors that may lead to selection biases in operational contexts (Brown et al., 2021). Furthermore, this bias may lead to endogeneity, such that difference factors between the two groups correlate with both the independent and outcome study variables. The present study used a student sample which was not representative of operational study samples for which device-type differences have been observed, consequently attenuating the population validity of the study. Per the above reasoning, general differences in the demographic characteristics or variability of demographic characteristics between operational samples and the student sample used in the present study would lead to differences in the observance of said biases between the two study contexts.

In support of this explanation, Brown et al. (2021) tested the effects of selection bias using a sample of job applicants and found that device usage was predicted by educational attainment, such that job applicants with lower levels of educational attainment were more likely to use a mobile device to complete a selection assessment than job applicants with higher levels of educational attainment which translated into score differences. Participants in their study were not randomly assigned to device conditions, which resulted in treatment groups that were unbalanced in their levels of the underlying factors (e.g., educational attainment) that correlate with the variables of interest.

As reported in related studies, various dispositional factors of mobile and non-mobile test takers often covary with the device used within operational contexts. Specifically, based on race and sex, African American, Hispanic, and female test takers are more likely to complete a selection assessment via a mobile device than white males (Arthur et al., 2018; Arthur et al., 2014; Illingworth et al., 2015; McClure Johnson & Boyce, 2015; Rossini, 2016). Concurrently, the literature reports that, in comparison to White individuals, a larger percentage of African American and Hispanic individuals in the U.S. are smartphone dependent. Smartphone dependency in the U.S. is also negatively correlated with SES such that smartphone dependency decreases with increases in income. According to the Pew Research Center (2021), 27% of U.S. adults earning less than $30,000 per year indicated that they were smartphone dependent in 2021 compared to 6% of U.S. adults earning $75,000 or more. Unsurprisingly Pew Research Center (2021) also reported that, in comparison to U.S. adults with a college degree, a larger percentage of U.S. adults without a college degree indicated that they were smartphone dependent.

As the data illustrate, there is a relationship between SES, race, and educational attainment, such that traditionally marginalized communities in terms of race and ethnicity have and continue to suffer from a lack of monetary and educational resources. The data suggest that the likelihood of owning a device more conducive to testing such as a desktop or laptop computer is predicated on the availability of resources to obtain such devices.

Acknowledging the relationship between the variables of interest and underlying demographic variables, Brown et al. (2021) found that the magnitude of cognitive mean score differences between "mobile" and "non-mobile" devices decreased after controlling for selection bias (from d = 0.58 to d = 0.25). They then found that the magnitude of the difference decreased further after using post-stratification (sample weighting) to better estimate population parameters. These results suggest that individual difference characteristics largely contribute to device-type cognitive mean score differences in operational contexts.

Although the sample of student participants used in the present study was relatively homogenous in reference to educational attainment, a small subsample still chose to use a high cognitively demanding device to complete the device self-selection assessment. While college students use their smartphones quite frequently to perform a variety of tasks, they are generally afforded access to alternate assessment devices such as desktop and laptop computers either through device ownership or technology resources provided by the universities they attend. Therefore, students are likely to have access to a laptop or desktop computer on or off campus. Contrary to many job applicants without a desktop, laptop, or tablet, college students generally have more discretion in the types of internet devices they use. Unfortunately, for the present study, participants were not asked to disclose their reasons for using a high or low cognitively demanding device to complete the device self-selection UIT. Consequently, we can only speculate about potential motivations for participants choosing to use one type of device over another.

However, although we cannot speak to motivations for device choice pertaining to the sample used for the present study, an examination of qualitative responses from a small sample of students from the same university may serve to inform student motivations for using a particular device to complete a UIT.

As part of a course assignment in which college students completed a battery of assessments—including the GMA_{60} measure used here—under conditions similar to that for device self-selection condition, participants responded to an item asking them to elaborate on their reason for choosing to complete the measures via the device they chose (Traylor, 2019). Responses were obtained for 68 students. The short answer responses revealed that students who chose to use their laptop or desktop to complete the survey most often did so because (1) they perceived the device to be more efficient than their smartphone either due to screen size or functionality, (2) they mostly use their laptop or desktop for school related activities and assignments, or (3) their laptop or desktop was either readily available or was the only option available. Students who chose to use a smartphone device to complete the assessment generally did so because their smartphone was the most readily accessible/convenient, or—to a lesser extent—because their smartphone was the only option available either due to traveling or similar circumstances.

Although these data cannot directly speak to participants' motivations for using a particular type of device to complete assessments in the present study, the samples are similar enough to suggest they are representative of one another. The data suggest that students may have a tendency to use low cognitively demanding devices to complete university-related assessments because they associate these assessments with schoolwork and prefer to complete schoolwork on devices that are better designed for testing (devices with larger screens, etc.).

These results are consistent with Brown et al.'s (2021) educational attainment explanation for their observed results. In pursuit of higher education, low cognitively demanding devices such as desktop and laptop computers are necessary to complete schoolwork. Consequently, the students were likely to have access to one of these devices.

Environmental Distractions

Supplementary analyses were conducted (1) to determine the effects of environmental distractions and difficulty concentrating on cognitive ability mean scores, (2) to examine the relationships between characteristics of the testing environment and the types of devices used, and (3) to determine the relationship between environmental distractions and difficulty concentrating. The data show that environments considered more distracting were also environments in which participants had more difficulty concentrating. However, these environmental factors did not influence performance on the cognitive assessments irrespective of the type of devices used to complete them. These results mirror those of Traylor et al. (2021) in that there was no relationship between the environmental factors (distractions and difficulty concentrating) and cognitive ability performance. Additionally, there were no device-type differences in the extent to which participants found the testing environment to be distracting or the extent to which participants had difficulty concentrating. These results support those of Lawrence et al. (2017) which did not find any device-type differences in the extent to which participants reported interruptions while completing an assessment. Contrary to the environmental distractibility hypothesis, Lawrence et al. found that individuals used mobile and non-mobile internet compatible devices in similar environments when completing assessments.

Implications for Research and Practice

The present study speaks to the importance of using a representative sample to achieve population validity when studying phenomena unique to a specific context. The study attempted to closely approximate the conditions of an operational study context by including a performance-based reward to increase motivation and testing stakes and by allowing participants to use a device of their choosing to complete a UIT composed of cognitive and noncognitive measures. However, simulated operational conditions pertaining to the study environment, testing stakes, and device selection were not enough to produce device-based outcomes in a lab setting. The study was representative of an operational context in many ways except the sample used. These outcomes suggest that to successfully replicate the outcomes of a real-world context in a nonoperational context, researchers must work to achieve both ecological and population validity.

Although the present study was not successful in producing observed device-type effects unique to operational contexts, the results do not diminish the need for controlled laboratory studies of device-type effects. Such studies are needed to permit causal inferences and explanations pertaining to the specific individual differences that contribute to differences in device use. Brown et al. (2017) were able to reduce device-type effects by statistically controlling for selection biases, however experimental studies are still necessary to infer causal relationships in this domain.

Furthermore, as discussed by Brown et al. (2021), controlling for extraneous variables through random assignment generally eliminates the concern for selection bias, as is the case with true experimental study designs (Brown et al., 2021). Given this level of control generally is not feasible in an operational context, researchers studying UIT should consider the effects of such underlying variables as educational attainment, sex, race, and income when interpreting operational study results.

The present study combined with the results of Brown et al. (2021) also suggests that device-type cognitive means score differences observed in operational contexts likely are not the result of differences in device characteristics. To the contrary, various sources of device usage data (Arthur et al., 2018; Pew Research Center, 2021) and the results of Brown et al. (2021) suggest that device usage is a product of factors relating to device access (i.e., educational attainment, SES, etc.). Consequently, employers should tread with caution when deciding whether to impose device restrictions on selection assessments. Prematurely restricting the use of certain types of internet compatible devices could potentially limit the diversity of the applicant pool by removing an important source of internet connection for specific subgroups of job applicants.

Limitations

The present study attempted to approximate the conditions of an operational study context while isolating the effects of device self-selection. Nonetheless, it was unsuccessful in replicating the observed device-related cognitive outcomes of operational studies. As previously mentioned, the study did not account for demographic differences between samples of job applicants and the sample of students who—in addition to the potential performance-based reward—completed the study for course credit. To closely approximate the conditions of an operational study context and determine the effects of individual differences on device selection, the sample used must also be representative of the population of job seekers. Samples of job applicants from various fields include individuals from different walks of life. The variation in cognitive ability, and consequently factors relating to it (i.e., educational attainment), is likely greater in operational samples than in samples of students attending the same university. Participants in the present study were generally homogenous in age (M = 18.9, SD = 1.21) and educational attainment—as they were all in the undergraduate stage of their educational journey. Contrary to some operational study contexts in which the availability of resources such as desktop and laptop computers likely varies, samples of college students often have access to such resources because they are necessary for their college studies. Consequently, it is likely that the study sample did not closely approximate some of those used in operational study contexts.

In their study measuring the effects of environmental distractions and device type on cognitive ability performance, Traylor et al. (2021) also used a student sample and found no differences in cognitive ability performance resulting from the distractibility of the testing environment and the type of device used. In their interpretation of the results, Traylor et al. (2021) suggest that the results may have been indicative of a cohort effect, whereby the age cohort's familiarity with using high cognitively demanding devices may have resulted in their adjustment to using such devices in distracting environments. The hypothesized self-selection and environmental effects of the present study and Traylor et al. (2021) respectively have yet to be examined in a controlled lab environment using a non-student sample that is more representative of the general working population. Consequently, the outcomes could potentially differ from those observed if one were to use a more representative sample.

The results of the present study pertaining to environmental distractions and difficulty concentrating mirrored the results of Traylor et al. (2021) such that there was no observed relationship between the level of environmental distraction and cognitive ability outcomes. However, the sample used for analyses was greatly reduced by an oversight in data collection. No questions assessing environmental distractions or difficulty concentrating were included in the assessments administered to the first cohort of participants. Environmental distraction and concentration difficulty data were only obtained for the second cohort (half the starting sample) of participants—consequently reducing the statistical power and increasing the likelihood of a type one error.

In acknowledgement of the relationships between race/ethnicity and both cognitive ability outcomes and device use/ownership (Lambert, 1970; Parker, 2014; Pew Research Center, 2021), the present study statistically controlled for the effects of race/ethnicity. Studies and statistics also show that there is a relationship between socio-economic status (SES) and both cognitive ability (Lambert, 1970; Parker, 2014) and internet compatible device ownership (Pew Research Center, 2021), however, due to an oversight when developing the study assessments, SES was not measured as part of the study. Consequently, we were only able to control for the effects of race, as these data were provided, however we cannot speak to the relationships between SES and race, device ownership, or cognitive ability for the student sample used. A direct measure of SES such as family household income would have informed the homogeneity of the sample in regard to social economic status, allowing for further contemplation about the role of SES in explaining observed device-type effects on mean cognitive ability scores for UITs.

Using race as a covariate also posed an issue to the overall size of the sample used for analysis. The sample was limited to participants who identified as White/non-Hispanic, African American, or Hispanic (collectively representing the majority of the sample) to incorporate race as a covariate and maintain the same sample of participants for each set of primary analyses. This reduction in the sample size decreased the overall power of the study and further limited the size of the high cognitively demanding device-type condition. Furthermore, limiting the sample to three specific ethnic groups reduced the generalizability of the results to a much more diverse population of job applicants.

Future Research

As previously mentioned, the present study highlights the importance of using a representative sample when conducting controlled lab research. To increase population and ecological validity, the sample and other contextual elements of the study should well represent those found within the context of interest. The present study controlled for key elements of the operational environment yet used a sample characteristically different from the population of interest. Although random sampling from the population is not always feasible, researchers should consider the similarity of the sample to the population of interest in designing future studies.

The present study used the SCIP framework to inform the classification of internet compatible devices based on the construct irrelevant cognitive demand imposed by the device characteristics. The SCIP framework posits that differences in the physical characteristics of UIT devices contribute to differences in cognitive demand imposed by the device. Per the framework, all other things being equal, devices higher in cognitive demand would require more cognitive resources—consequently competing with an individual's pool of cognitive resources available for cognitive tasks. Although no differences that would be representative of a device-type explanation were observed in the present study, the potential cohort or generation effect that Traylor et al. (2021) describe as a potential issue in their study may have attenuated or masked any effects deriving from differences in construct irrelevant demands within the present study. Participants were college students accustomed to using various types of internet compatible devices regularly, thus increasing their familiarity with these devices and the ease of use of the devices. Acknowledging the specificity of these characteristics to the sample, future studies should seek to determine the extent to which construct irrelevant cognitive demands imposed by internet compatible devices influence cognitive ability performance, if at all, using a diverse sample of participants. It is plausible that the magnitude of any differences in cognitive demand imposed by each type of device may be too small to be of meaningful concern. Nonetheless, a representative sample is needed to ensure that the effects are not a product of sample specific factors.

Conclusion

In summary, the present study empirically examined the extent to which the self-selection hypothesis accounts for the discrepancy in the observance of cognitive UIT device-type effects in operational and nonoperational study contexts. The findings did not provide support for a self-selection explanation, however the results emphasized the importance of ecological and population validity when conducting organizational research.

REFERENCES

Arthur, W., Jr. (2017). *An unproctored internet-based test of general mental ability. A validation report.* College Station, TX: Author.

Arthur, W., Jr., & Day, D. V. (1994). Development of a short form for the Raven Advanced Progressive Matrices Test. *Educational and Psychological Measurement*, *54*(2), 394-403. https://doi.org/10.1177/0013164494054002013

*Arthur, W., Jr., Doverspike, D., Muñoz, G. J., Taylor, J. E., & Carr, A. E. (2014). The use of mobile devices in high-stakes remotely delivered assessments and testing. *International Journal of Selection and Assessment*, *22*(2), 113-123. https://doi.org/10.1111/ijsa.12062

Arthur, W., Jr., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2010). The magnitude and extent of cheating and response distortion effects on unproctored internet-based tests of cognitive ability and personality. *International Journal of Selection and Assessment*, *18*(1), 1-16. https://doi.org/10.1111/j.1468-2389.2010.00476.x

Arthur, W., Jr., Keiser, N. L., & Doverspike, D. (2018). An information-processing-based conceptual framework of the effects of unproctored internet-based testing devices on scores on employment-related assessments and tests. *Human Performance*, *31*(1), 1-32. https://doi.org/10.1080/08959285.2017.1403441

*Arthur, W., Jr., Keiser, N. L., Hagen, E., & Traylor, Z. (2018). Unproctored internet-based device-type effects on test scores: The role of working memory. *Intelligence*, *67*, 67-75. https://doi.org/10.1016/j.intell.2018.02.001

Arthur, W., Jr., & Traylor, Z. (2019). Mobile assessment in personnel testing: Theoretical and practical implications. In R. N. Landers (Ed.), *The Cambridge handbook of technology and employee behavior* (pp. 179-207). New York: Cambridge University Press. https://doi.org/10.1017/9781108649636.009

Arthur, W., Jr., Tubre, T. C., Paul, D. S., & Sanchez-Ku, M. L. (1999). College-sample psychometric and normative data on a short form of the Raven Advanced Progressive Matrices Test. *Journal of Psychoeducational Assessment*, *17*(4), 354-361. https://doi.org/10.1177/073428299901700405

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing*, *10*(4), 342-363. https://doi.org/10.1080/15305058.2010.508569

Beaty, J. C., Nye, C. D., Borneman, M. J., Kantrowitz, T. M., Drasgow, F., & Grauer, E. (2011). Proctored versus unproctored internet tests: Are unproctored noncognitive tests as predictive of job performance? *International Journal of Selection and Assessment*, *19*(1), 1-10. https://doi.org/10.1111/j.1468-2389.2011.00529.x

*Brown, M. I., & Grossenbacher, M. A. (2017). Can you test me now? Equivalence of GMA tests on mobile and non-mobile devices. *International Journal of Selection and Assessment*, *25*(1), 61-71. https://doi.org/10.1111/ijsa.12160

Brown, M., Grossenbacher, M., & Warman, Z. (2021). *Investigating the conflicting findings for mobile device effects on cognitive test scores: Evidence for selection bias in an operational, applicant sample*. PsyArXiv. https://doi.org/10.31234/osf.io/rkd8p

Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. *Journal of Personality and Social Psychology, 92*(5), 938-956. https://doi.org/10.1037/0022-3514.92.5.938

Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2020). Decision-making competence: More than intelligence? *Current Directions in Psychological Science*, *29*(2), 186-192. https://doi.org/10.1177/0963721420901592

*Chang, L., Lawrence, A. D., O'Connell, M. S., & Kinney, T. B. (2016). *What was that? Investigating distractions on test performance and reactions*. Paper presented at the 31st Annual Conference of the Society for Industrial and Organizational Psychology, April, Anaheim, CA.

*Dages, K. D., & Jones, J. W. (2015). Mobile device administration: Does length or level of assessment matter? In N. Morelli (Chair), *Mobile devices in talent assessment: The next chapter.* Symposium presented at the 30th Annual Conference of the Society for Industrial and Organizational Psychology, April, Philadelphia, PA.

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment*, *18*(2), 192-203. https://doi.org/10.1037/1040-3590.18.2.192

Eccles, J. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motives* (pp. 75-146). San Francisco: Freeman.

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers, 28*(1), 1-11. https://doi.org/10.3758/BF03203630

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analysis using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, Instruments, & Computers, 41*(4), 1149-1160. https://doi.org/10.3758/BRM.41.4.1149

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic perspectives*, *19*(4), 25-42. DOI: 10.1257/089533005775196732

Fry, R., Bennett, J., Barroso, A. (2021) *Racial and ethnic gaps in the U.S. persist on key demographic indicators*. Pew Research Center.

https://www.pewresearch.org/interactives/racial-and-ethnic-gaps-in-the-u-s-persist-on-key-demographic-indicators/

*Fursman, P. (2016). Warning message impact on assessment scores delivered on mobile devices. In T. D. McGlochin (Chair), *Mobile equivalence: Expanding research across assessment methods, levels, and devices*. Symposium presented at the 31st Annual Conference of the Society for Industrial and Organizational Psychology, April, Anaheim, CA.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*(1), 26-42. https://doi.org/10.1037/1040-3590.4.1.26

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe*, *7*(1), 7-28.

Golubovich, J., & Boyce, A. (2013). Hiring tests: Trends in mobile device usage. In N. A. Morelli (Chair), *Mobile devices in talent assessment: Where are we now?* Symposium presented at the 28th Annual Conference of the Society for Industrial and Organizational Psychology, April, Houston, TX.

*Grelle, D. M., & Gutierrez, S. L. (2019). Developing device-equivalent and effective measures of complex thinking with an information processing framework and mobile first design principles. *Personnel Assessment and Decisions*, *5*(3), 21-32. https://doi.org/10.25035/pad.2019.03.004

Gutierrez, S. L., Meyer, J. M., & Fursman, P. (2015). What exactly drives positive reactions to mobile device administration? In N. Morelli (Chair), *Mobile devices in talent assessment: The next chapter.* Symposium presented at the 30th Annual Conference of the Society for Industrial and Organizational Psychology, April, Philadelphia, PA

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology*, *92*(2), 373-385. doi:10.1037/0021-9010.92.2.373

*Illingworth, A. J., Morelli, N. A., Scott, J. C., & Boyd, S. L. (2015). Internet-based, unproctored assessments on mobile and non-mobile devices: Usage, measurement equivalence, and outcomes. *Journal of Business and Psychology*, *30*(2), 325-343. https://doi.org/10.1007/s10869-014-9363-8

*Impelman, K. (2013). Mobile assessment: Who’s doing it and how it impacts selection. In N. A. Morelli (Chair), *Mobile devices in talent assessment: Where are we now?* Symposium presented at the 28th Annual Conference of the Society for Industrial and Organizational Psychology, April, Houston, TX.

Jackson, S. A., Kleitman, S., Stankov, L., & Howie, P. (2017). Individual differences in decision making depend on cognitive abilities, monitoring and control. *Journal of Behavioral Decision Making*, *30*(2), 209-223. https://doi.org/10.1002/bdm.1939

*King, D. D., Ryan, A. M., Kantrowitz, T., Grelle, D., & Dainis, A. (2015). Mobile internet testing: An analysis of equivalence, individual differences, and reactions. *International Journal of Selection and Assessment*, *23*(4), 382-394. https://doi.org/10.1111/ijsa.12122

Kulik, J. A., Kulik, C. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal*, *21*(2), 435-447. https://doi.org/10.3102/00028312021002435

Lambert, N. M. (1970). Paired associate learning, social status and tests of logical concrete behavior as univariate and multivariate predictors of first grade reading achievement. *American Educational Research Journal*, *7*(4), 511-528. https://doi-org.srv-proxy1.library.tamu.edu/10.3102/00028312007004511

*LaPort, K. (2016). Mobile assessment: Comparing traditional cognitive, cognitive reasoning, and non-cognitive performance. In T. D. McGlochin (Chair), *Mobile equivalence: Expanding research across assessment methods, levels, and devices*. Symposium presented at the 31st Annual Conference of the Society for Industrial and Organizational Psychology, April, Anaheim, CA.

*Lawrence, A., Chang, L., O'Connell, M. S., & Kinney, T. B. (2016). Mobile simulations: Can you control for device? In T. D. McGlochin (Chair), *Mobile equivalence: Expanding research across assessment methods, levels, and devices*. Symposium presented at the 31st Annual Conference of the Society for Industrial and Organizational Psychology, April, Anaheim, CA.

Lawrence, A. D., & Kinney, T. B. (2017). *Mobile devices and selection*. [White Paper] Retrieved November 6, 2018, from SIOP White Paper Series. Visibility Committee Society for Industrial and Organizational Psychology: https://www.researchgate.net/profile/Amie_Lawrence/publication/320188262_Mobile_Devices_and_Selection_SIOP_White_Paper_Series/links/59d3a62b4585150177f96619/Mobile-Devices-and-Selection-SIOP-White-Paper-Series.pdf

*Lawrence, A. D., Kinney, T. B., O'Connell, M. S., & Delgado, K. M. (2017). Stop interrupting me! Examining the relationship between interruptions, test performance, and reactions. *Personnel Assessment and Decisions*, *3*(1), 15-24. https://doi.org/10.25035/pad.2017.002

*Lawrence, A. D., Wasko, L., Delgado, K., Kinney, T. B., & Wolf, D. (2013). Understanding the mobile experience: Data across device and industry. In J. Scott (Chair) and N. Morelli (Co-Chair), *Mobile devices in talent assessment: Where are we now*. Symposium presented at the 28th Annual Conference of the Society for Industrial and Organizational Psychology, April, Houston, TX.

Lim, B. C., & Ployhart, R. E. (2006). Assessing the convergent and discriminant validity of Goldberg's International Personality Item Pool: A multitrait-multimethod examination. *Organizational Research Methods*, *9*(1), 29-54. https://doi.org/10.1177/1094428105283193

*Martin, N., Capman, J., Boyce, A., Morgan, K., Gonzalez, M. F., & Adler, S. (2020). New frontiers in cognitive ability testing: working memory. *Journal of Managerial Psychology*. https://doi.org/10.1108/JMP-09-2018-0422

McClure Johnson, T. K., & Boyce, A. S. (2015). Selection testing: An updated look at trends in mobile device usage. In N. Morelli (Chair), *Mobile devices in talent assessment: The next chapter.* Symposium presented at the 30th Annual Conference of the Society for Industrial and Organizational Psychology, Philadelphia, PA.

Naber, A. M., Arthur, W., Jr., Edwards, B. D., & Franco-Watkins, A. (2021). *Increased retest scores on cognitive ability and knowledge tests: Retrieval or acquisition effects*. Manuscript submitted for publication.

Parker, B. N. (2014). Smartphones in selection: Exploring measurement invariance using item response theory. (Doctoral dissertation) Retrieved from ProQuest Dissertations & Theses Global. (1660761067).

*Parker, B. N., & Meade, A. W. (2015). Smartphones in selection: Exploring measurement invariance using item response theory. In N. Morelli (Chair), *Mobile devices in talent Assessment: The next chapter.* Symposium presented at the 30th Annual Conference of the Society for Industrial and Organizational Psychology, Philadelphia, PA.

Petor, J. R., Lawrence, A. D., & Kinney, T. B. (2017). *Painting the picture: What is the mobile test environment.* Poster Presented at the 32nd Annual Conference of the Society for Industrial Organizational Psychology, Orlando, FL.

Pew Research Center. (2021, April 7) *Mobile fact Sheet.* Retrieved from https://www.pewresearch.org/internet/fact-sheet/mobile/

PSI Services (2021) *About us*. https://www.psionline.com/company/about-us/

Raven, J. C., Court, J. H., & Raven, J. (1985) *A Manual for Raven's Progressive Matrices and Vocabulary Scales*. London: H. K. Lewis.

Rossini, J. (2016) Mobile device testing: A five-year look across job level. In T.D. McGlochlin (Chair), *Mobile equivalence: Expanding research across assessment methods, levels, and devices.* Symposium presented at the 31st Annual Conference of the Society for Industrial and Organizational Psychology, Anaheim, CA.

Roth, P. L., Bevier, C. A., Bobko, P., Switzer III, F. S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis.

Personnel Psychology, *54*(2), 297-330. https://doi.org/10.1111/j.1744-6570.2001.tb00094.x

Ryan, A. M., & Ployhart, R. E. (2014). A century of selection. *Annual Review of Psychology*, *65*, 693-717. https://doi.org/10.1146/annurev-psych-010213-115134

Sackett, P. R., & Walmsley, P. T. (2014). Which personality attributes are most important in the workplace? *Perspectives on Psychological Science*, *9*(5), 538-551. https://doi-org.srv-proxy1.library.tamu.edu/10.1177/1745691614543972

Scharfen, J., Peters, J. M., & Holling, H. (2018). Retest effects in cognitive ability tests: A meta-analysis. *Intelligence*, *67*, 44-66. https://doi.org/10.1016/j.intell.2018.01.003

Schmitt, N. (2014). Personality and cognitive ability as predictors of effective performance at work. *Annual review of organizational psychology and organizational behavior*, *1*(1), 45-65. https://doi.org/10.1146/annurev-orgpsych-031413-091255

Sinar, E., & Wasko, L. (2008*). Further Exploring the Nature and Impact of Differing Test Environments.* Symposium presented at the 23rd Annual Conference of the Society for Industrial and Organizational Psychology, San Francisco, CA.

*Smeltzer, S. (2013). *Mobile internet testing: Applicant reactions to mobile internet testing* (Master's thesis). Retrieved from ProQuest. 154893.

Steger, D., Schroeder, U., & Gnambs, T. (2018). A meta-analysis of test scores in proctored and unproctored ability assessments. *European Journal of Psychological Assessment*. https://doi.org/10.1027/1015-5759/a000494

Tippins, N. T. (2009). Internet alternatives to traditional proctored testing: Where are we now? *Industrial and Organizational Psychology*, *2*(1), 2-10. https://doi.org/10.1111/j.1754-9434.2008.01097.x

Tippins, N. T. (2015). Technology and assessment in selection. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 551-582. https://doi.org/10.1146/annurev-orgpsych-031413-091317

Tippins, N. T., Beaty, J., Drasgow, F., Gibson, W. M., Pearlman, K., Segall, D. O., & Shepherd, W. (2006). Unproctored internet testing in employment settings. *Personnel Psychology*, *59*(1), 189-225. https://doi.org/10.1111/j.1744-6570.2006.00909.x

Traylor, Z. (2019). [Student survey] [Unpublished raw data]. Texas A&M University.

Traylor, Z., Hagen, E., Williams, A., & Arthur Jr, W. (2021). The testing environment as an explanation for unproctored internet-based testing device-type effects. *International Journal of Selection and Assessment*, *29*(1), 65-80. https://doi.org/10.1111/ijsa.12315

Wasko, L., Lawrence, A. D., & O'Connell, M. S. (2015). *What matters in the test environment?* Symposium presented at the 30th Annual Conference of the Society for Industrial and Organizational Psychology, Philadelphia, PA.

Wigfield, A., & Eccles, J. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68-81. https://doi.org/10.1007/BF02209024

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational assessment*, *10*(1), 1-17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S. L., & Smith, L. F. (2011). A model of examinee test-taking effort. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendahl (Ed.), *High-stakes testing in education: Science and practice in K-12 settings* (pp. 139-153). Washington, DC: American Psychological Association.

*Wood, E., Stephens, K., & Slither, K. (2015). *Apples to oranges? Use and comparative scores for mobile and nonmobile selection assessments*. Paper presented at the 2015 Annual Conference of the International Personnel Assessment Council, July, Atlanta, GA.

APPENDIX A

Table A1

Meta-Analysis Results of the Difference in UIT Scores for Devices Classified as Mobile and Non-mobile in Operational and Nonoperational Contexts

						95% CI	
Studies included	*K*	*N*	*d*	*SD*	%Var	LL	UL
Overall	114	31,259,446	0.06	0.27	0.02	-0.46	0.58
Operational	79	31,236,593	0.06	0.27	0.01	-0.46	0.58
Cognitive	15	3,105,082	0.79	0.22	0.04	0.35	1.22
Noncognitive	64	28,131,511	-0.02	0.09	0.11	-0.20	0.16
Lab	35	2122,853	0.04	0.12	29.43	-0.21	0.28
Cognitive	14	17,176	0.05	0.14	14.97	-0.22	0.33
Noncognitive	20	5,534	-0.02	0.00	100.00	-0.02	-0.02

Note. *K* = number of independent studies; *N* = sample size; *d* = sample-weighted mean difference; *SD* = corrected standard deviation; %Var = percentage of variance accounted for by sampling error; CI = confidence interval; LL = lower limit; UL = upper limit.

APPENDIX B

MEASURES

Sample Items from GMA_{60} measure

This is a 10-minute timed test. There is a total of 60 items, but the test will probably be too long for you to finish. However, complete as many items as you can in the allotted time. Work quickly and accurately. Do not spend too much time on any one item. Your score will be the number of items that you answer correctly. Since some of the problems you will encounter may require some "figuring out", you may write in the test booklet in trying to solve them. Record your answer by legibly circling the letter (e.g., "A") corresponding to your choice.

1. What is 15% of 200?

   A. 15
   B. 20
   C. 30
   D. 45

2. 0 5 0 5 1 4 1

   A. 0
   B. 1
   C. 4
   D. 5

3. FAST is most similar in meaning to

   A. Light
   B. Quick
   C. Primary
   D. Attach

Conscientiousness and Agreeableness

Directions

Listed below are phrases describing people's behaviors. Please use the scale provided below to identify how accurately each statement describes you. Describe yourself as you generally are now, not as you wish to be in the future. Describe yourself as you honestly see yourself in relation to other people you know of the same sex and roughly the same age as you. Please read each statement carefully, and then rate the extent to which it accurately describes you.

①	②	③	④	⑤
Very Inaccurate	Inaccurate	Neither Inaccurate nor Accurate	Accurate	Very Accurate

1.	Have a soft heart.	① ② ③ ④ ⑤
2.	Am always prepared.	① ② ③ ④ ⑤
3.	Sympathize with others' feelings.	① ② ③ ④ ⑤
4.	Get chores done right away.	① ② ③ ④ ⑤
5.	Feel others' emotions.	① ② ③ ④ ⑤
6.	Make a mess of things.	① ② ③ ④ ⑤
7.	Am not really interested in others.	① ② ③ ④ ⑤
8.	Am exacting in my work.	① ② ③ ④ ⑤
9.	Feel little concern for others.	① ② ③ ④ ⑤
10.	Like order.	① ② ③ ④ ⑤
11.	Make people feel at ease.	① ② ③ ④ ⑤
12.	Leave my belongings around.	① ② ③ ④ ⑤
13.	Am not interested in other people's problems.	① ② ③ ④ ⑤
14.	Pay attention to details.	① ② ③ ④ ⑤
15.	Take time out for others.	① ② ③ ④ ⑤
16.	Shirk my duties.	① ② ③ ④ ⑤
17.	Insult people.	① ② ③ ④ ⑤
18.	Follow a schedule.	① ② ③ ④ ⑤
19.	Am interested in people.	① ② ③ ④ ⑤
20.	Often forget to put things back in their proper place.	① ② ③ ④ ⑤

Note. Odd-numbered items measure agreeableness, and even-numbered items measure conscientiousness. Items 1, 6, 7, 9, 12, 13, 16, 17, and 20 are reverse-coded.

Demographics

Sex:

- ☐ Male
- ☐ Female

Age in years: _______

Ethnicity:

- ☐ Hispanic or Latino
- ☐ White (Not Hispanic or Latino)
- ☐ Black or African American (Not Hispanic or Latino)
- ☐ Native Hawaiian or Other Pacific Islander (Not Hispanic or Latino)
- ☐ Asian (Not Hispanic or Latino)
- ☐ American Indian or Alaska Native (Not Hispanic or Latino)
- ☐ Two or More Races (Not Hispanic of Latino)
- ☐ Other (please fill in) ______________________________

Classification:

- ☐ Provisional Freshman
- ☐ Freshman
- ☐ Sophomore
- ☐ Junior
- ☐ Senior
- ☐ Non-Degree Seeking

Work Status:

- ☐ Employed full time
- ☐ Employed part time
- ☐ Not employed

If employed, how long have you held your current position (in years): __________

If employed, then please select your current Job Function in the organization in which you work. Please note, there may not be an exact match, so select the overarching function you think is best.

- ☐ Operations
- ☐ Sales
- ☐ Services
- ☐ Information Technology (IT)
- ☐ General and Administrative
- ☐ Executive/Upper Management
- ☐ Research and Development (R&D)
- ☐ Engineering
- ☐ Human Resources
- ☐ Accounting/Finance

If employed, then please provide your current job title: ____________________________

High School GPA: _________

Current Overall GPA: _________

What was your best overall SAT score? ___________

What was this score out of?

- ☐ 1600
- ☐ 2400

What was your best overall ACT score (if available)? ___________

Device Characteristics

Device Type

What kind of device did you use to complete the assessment?

- ☐ Desktop computer
- ☐ Laptop computer
- ☐ Notebook computer
- ☐ Tablet
- ☐ Phablet
- ☐ Smartphone

*Based on previous selection, participants would respond to one of the following questions.

What kind of smartphone?

	Make	Model
Android	☐	☐
Blackberry	☐	☐
iPhone	☐	☐
Other	☐	☐

What kind of phablet?

	Make	Model
Android	☐	☐
Blackberry	☐	☐
iPhone	☐	☐
Other	☐	☐

What kind of tablet?

	Make	Model
Android		
Blackberry		
iPhone		
Other		

What operating system is running on this device?

- ☐ Windows
- ☐ Mac
- ☐ Other (please fill in)

Distractibility and Motivation

To what extent did you find the environment in which you completed the assessments to be distracting?

① Not at all distracting
②
③ Somewhat distracting
④
⑤ Very distracting

The environment in which I completed the assessments made it difficult for me to concentrate.

① Strongly disagree
②
③ Neutral
④
⑤ Strongly agree

How motivated were you to do your best on the assessments in this study?

① Not at all motivated
② Somewhat motivated
③ Quite a bit motivated
④ Very motivated
⑤ Extremely motivated