DEVICE-ENGENDERED COGNITIVE ABILITY SCORE DIFFERENCES ON UNPROCTORED INTERNET-BASED ASSESSMENTS: THE ROLE OF SELECTIVE ATTENTION

A Thesis

by

ZACHARY KYLE TRAYLOR

Submitted to the Office of Graduate and Professional Studies 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
	S. Camille Peres
Head of Department,	Heather C. Lench

December 2018

Major Subject: Industrial Organizational Psychology

ABSTRACT

Although a volume of literature suggests that the device used to complete unproctored Internet-based tests (UIT) affects observed test scores, there have been a limited number of attempts to provide a psychological explanation for why this occurs. One such exception is Arthur, Keiser, and Doverspike's (2018) Structural Characteristics/Information Processing (SCIP) model, which provides a psychological explanation regarding the conditions under which one would expect UIT device types (e.g., desktop computer, smartphone, tablet) to affect test scores. The model proposes that systematic error is introduced via construct-irrelevant cognitive load attributable to the additional information-processing demands elicited by the UIT device's structural characteristics. While conceptually sound, there has been only one empirical examination of the propositions advanced by the model to date. Consequently, the primary objective of the present study was to test the SCIP model's propositions regarding selective attention, the information-processing demand elicited by permissibility, the associated structural characteristic of UIT devices.

Two hundred sixty-one participants completed measures of general mental ability (GMA), personality, and selective attention. Participants were randomly assigned to one of two conditions differing in terms of the testing (1) environment and (2) device used to complete the GMA test and personality assessment (i.e., a busy, outdoor location [smartphone condition] or a quiet, indoor location [desktop condition]). Scores on the GMA test did not differ as a function of the testing device and environment, however, in accordance with the tenets of the SCIP model, it appears that test takers in the smartphone condition experienced a greater degree of selective-attention demands while completing the GMA test. All of the observed results are interpreted within the context of using an undergraduate student sample, low testing stakes, and random assignment instead of the self-selection of participants into conditions. Implications and limitations of the present study as well as recommendations for future research are discussed.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by the student's thesis committee chair, Professor Winfred Arthur, Jr., with Professor Stephanie C. Payne of the Department of Psychological and Brain Sciences, and Professor S. Camille Peres of the Department of Environmental and Occupational Health as committee members.

Funding Sources

There are no outside funding contributions to acknowledge related to the research and compilation of this document.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION AND LITERATURE REVIEW

The proliferation of advancements in wireless Internet technology has unquestionably impacted the manner in which employment-related tests and assessments are completed and data are collected within the context of personnel selection. For organizations, these new technologies have enabled the use and adoption of remote, unproctored Internet-based tests and assessments (UIT; Tippins et al., 2006; Tippins & Alder, 2011). As a result, applicants are afforded the opportunity to complete UITs in virtually any location and while using any device of their choosing. Consistent with the persistent rise in mobile device ownership among the general population (Pew Research Center, 2015a, 2015b, 2018), the use of mobile devices for employment-related testing (i.e., UIT) has become increasingly commonplace in the past decade. Although several advantages associated with UIT have been advanced, the seemingly unrestricted Internet access permitted by new technologies raises a few concerns regarding potential device-type effects on the observed scores of UITs.

A young, but growing volume of literature has indicated observed differences in test scores as a function of UIT device type, the most common demarcation being mobile or nonmobile. In an effort to provide a psychologically sound explanation for these UIT device-type effects and guide subsequent research, Arthur, Keiser, and Doverspike (2018) developed a conceptual framework featuring a device-engendered, construct-irrelevant cognitive-load continuum whereby UIT devices are arranged in terms of the extent to which the *structural characteristics* of the device in question elicit *construct-irrelevant information-processing demands* on the test taker. Accordingly, to the extent that cognitive resources are finite, any additional cognitive load that is engendered (i.e., via construct-irrelevant information-processing demands) should impede one's ability to adequately complete the test (or task) at hand. Although conceptually sound, there has been only one empirical test of the propositions advanced by the SCIP model at the present time (i.e., those pertaining to working memory; Arthur, Keiser, Hagen, & Traylor, 2018). Consequently, the objective of the present study is to test the propositions associated with selective attention, the information-processing demand elicited by permissibility, the associated UIT-device structural characteristic.

UIT Device-Type Effects

There presently exists relatively few empirical investigations of UIT device-type effects, and the majority of these are conference presentations. Arthur, Keiser, and Doverspike (2018) present a comprehensive review of this literature, and the following section highlights what has been found in terms of observed score differences as a function of the UIT device-type used to complete measures of cognitive and noncognitive constructs.

Score Differences Between UIT Device Types

The majority of the mobile testing literature differentiates between devices in terms of whether a device is "mobile" or "nonmobile," or, rather, untethered to the wall or not (i.e., unplugged versus plugged, respectively; Arthur & Traylor, in press). According to this classification, then, smartphones, tablets, and laptops are examples of mobile devices and desktop computers are an example of a nonmobile device. Despite arguing against the preceding designation in subsequent sections (per Arthur, Keiser, & Doverspike, 2018), the following review and discussion of the literature maintains the original language used within the cited sources. Additionally, the empirical evidence regarding UIT device-type score differences is delineated in terms of whether the construct being assessed was cognitive or noncognitive.

Cognitive versus noncognitive constructs. With any comparative evaluation of predictor methods, in this case predictor modes, it is imperative that one (a) identifies the construct being assessed by each method and (b) holds the constructs constant (Arthur & Villado, 2008). Constructs in personnel selection and assessment can be classified as cognitive or noncognitive. To that end, the focal interest of the present study is whether the construct being assessed is cognitive or noncognitive, not on the *specific* cognitive or noncognitive construct. Namely, whether the observed scores on cognitive and noncognitive tests and assessments differ as a function of the UIT device used to complete said tests or assessments.

Cognitive constructs refer to those that measure aspects of individuals' cognitive functioning. Concomitant with this definition, cognitive constructs are ordinarily characterized by the fact that the measures of such consist of items that have prespecified incorrect and correct (or best) answers (Arthur & Glaze, 2011). Some examples of cognitive constructs are verbal, quantitative, and spatial reasoning, general mental ability (GMA) or intellectual capability, working memory, and perceptual speed.

In contrast, noncognitive constructs refer to those that measure individual differences in domains such as personality characteristics and emotional and volitional processes, and are reflected by measures for which the constituent items have no correct or incorrect answers (Arthur & Glaze, 2011). Some examples of noncognitive constructs are motivation, personality traits, and need for achievement. Per decades of research within personnel psychology, GMA is regarded as the single best predictor of work performance (e.g., Schmitt, 2014), and personality traits, particularly agreeableness and conscientiousness, are among the most valid noncognitive predictors of work performance (Sackett & Walmsley, 2014).

Cognitive constructs. UIT device-type effects on the observed scores of tests measuring cognitive constructs have been demonstrated by several studies within this relatively nascent research stream. For example, Arthur, Doverspike, Muñoz, Taylor, and Carr (2014) reported a considerable difference between mobile- and nonmobile-device users' scores ($d = 0.90$; nonmobile > mobile) using a large, operational dataset consisting of job applicants who completed a UIT. Similarly, Wood, Stephens, and Slither (2015) found the same pattern of score differences on two cognitive ability ($d = 0.46$; $d = 0.35$), and two mechanical aptitude tests ($d = 0.93$; $d = 0.26$) using a large, operational database. King, Ryan, Kantrowitz, Grelle, and Dainis (2015) used a within-subjects design to explore these effects and found a mean difference of $d = 0.16$ (nonmobile > mobile) between participants' scores on a cognitive ability test that was completed once using a desktop computer and once using a mobile device with a three-week interval between the two assessments.

Noncognitive constructs. In contrast to the measures of cognitive constructs, studies have consistently found no *meaningful* UIT device-type effects on observed noncognitive-assessment scores. Despite the nonzero effect sizes reported by the majority of studies, the magnitudes of these effect sizes are consistently small and not of much practical importance. For instance, McClure Johnson and Boyce (2015) used a large, operational dataset to examine scores that were obtained using an in-store kiosk, mobile device, or PC to complete an assessment tailored to either an entry-level or managerial position. For the entry-level assessment, the observed scores for individuals using the in-store kiosk, mobile device, and PC had corresponding average noncognitive *z*-scores of -0.107, -0.074, and 0.017, respectively, suggesting a negligible disadvantage to using both the kiosk or a mobile device. Similarly, for the managerial assessment, the observed scores for those using the in-store kiosk, mobile device, and PC had corresponding average noncognitive *z*-scores of -0.069, 0.026, and -0.002, respectively, suggesting a practically meaningless disadvantage to using the kiosk. Likewise, Arthur et al. (2014) used a large, operational dataset and found negligible differences between those who used mobile and those who used nonmobile devices to complete a personality assessment. For agreeableness, conscientiousness, emotional stability, extraversion, and openness, they reported corresponding *d*s of -0.01, -0.11, -0.08, -0.13, and 0.16, respectively.

Discrepancy in the literature. In contrast to the literature reviewed above, several studies have failed to obtain UIT device-type score differences on measures of cognitive constructs. For instance, Parker and Meade (2015) randomly assigned 692 individuals to complete assessments on a computer (did not specify whether desktop, laptop, etc.), a smartphone for which the assessment was "mobile optimized," or a smartphone for which the assessment was "mobile accessible." The omnibus test revealed no statistical difference in mean cognitive-test scores between the three conditions. Despite these and other contrary findings, there are several, consistent discrepancies in study type and context between those that have and have not obtained UIT device-type effects which might readily explain these mixed findings (Arthur et al., 2018; Arthur & Traylor, in press).

Unlike studies bounded within an operational context, test scores obtained by participants in nonoperational studies such as Parker and Meade's (2015) do not have any consequences associated with them. That is, in contrast to high-stakes, employment-related testing, participants' scores are rather inconsequential; the testing situations that characterize these studies are predominantly low stakes, hence, participants may not have a reason, or the motivation, to perform to the best of their ability. Additionally, nonoperational studies have generally randomly assigned participants to conditions (i.e., devices), which, by definition, eliminates test-takers' ability to choose (i.e., self-select) the device used to complete the administered tests and assessments. Brown and Grossenbacher (2017) purportedly tested this so-called self-selection hypothesis, however, a closer inspection revealed otherwise (Arthur & Traylor, in press). Akin to other nonoperational studies, Brown and Grossenbacher randomly assigned participants to conditions that differed by the device that was required to complete the study, necessarily removing the ability to self-select. Accordingly, test-takers' inability to self-select the device they use to complete the tests and assessments has been advanced as another viable explanation for why nonoperational studies have failed to observe score differences between those who use mobile and those who use nonmobile devices to complete cognitive tests (Arthur et al., 2014; Arthur et al., 2018; Arthur & Traylor, in press).

More germane to the present study, another plausible explanation for the mixed findings is that despite being framed as investigations of how mobile devices affect observed *UIT* scores, the vast majority of nonoperational studies have administered tests and assessments to participants in proctored, quiet, and controlled settings that are conducive to testing. To the extent that UITs grant the test taker vast degrees of freedom in terms of where (i.e., location) and how (i.e., device) to complete them, the environment wherein one chooses to complete the assessment may be less than favorable for optimal performance (Gray, Morelli, & McLane, 2015; Lawrence, Kinney, O'Connell, & Delgado, 2017). Although it is ultimately a decision on the part of the test taker, mobile devices such as smartphones and tablets readily allow one to complete tests in distracting environments such as a noisy bus ride, whereas nonmobile devices such as desktop computers are generally confined to quiet areas such as public libraries or one's office.

In summary, this stream of research suggests that one might expect UIT device-type effects *depending on the construct assessed*. Whereas scores for measures of cognitive constructs are consistently lower for those who used mobile devices compared to those who used nonmobile devices, there are generally no substantive score differences for measures of noncognitive constructs. In spite of the empirical evidence that has accumulated thus far, no attempts prior to Arthur, Keiser, and Doverspike's (2018) SCIP model had been made to provide a sound, psychological explanation as to (1) *why* one would expect UIT device-type effects, and (2) the conditions under which one would do so (e.g., when measuring cognitive versus noncognitive constructs).

The Structural Characteristics/Information Processing Model

As stated in the preceding review of the literature, score differences as a function of the UIT device type used (i.e., mobile versus nonmobile) have been found for measures of cognitive constructs, whereas only negligible differences have been found for measures of noncognitive constructs. Despite being consistent with how laypeople and the technology industry (both manufacturers and purveyors) differentiate among technological devices, the use of "mobile" and "nonmobile" to distinguish between UIT devices is problematic. From a psychological standpoint, simply referring to devices as one or the other is uninformative because doing so fails to provide a meaningful explanation for *why* one would expect the particular UIT device to affect scores on employment-related tests and assessments. To remedy this gap in the literature, Arthur, Keiser, and Doverspike (2018) developed a conceptual framework that provides a psychological explanation for why one would expect score differences as a function of UIT devices such as desktop computers, laptop computers, tablets, and smartphones. Novel in approach, Arthur et al.'s structural characteristics/information processing (SCIP) model is neither device-dependent nor an attempt to identify a hierarchy of UIT devices in terms of which devices are "better" for test-taking purposes. Rather, provided that the structural characteristics inherent to any particular UIT device can be assessed, the model allows one to appraise the extent to which the device engenders *additional* cognitive load, in the form of construct-irrelevant information-processing demands, which ultimately translates into observable device-type effects (or lack thereof).

The foundational tenet of the SCIP model is that systematic error is introduced via cognitive load attributable to the addition of construct-irrelevant information-processing demands elicited by the structural characteristics of UIT devices. Specifically, Arthur, Keiser, and Doverspike (2018) describe four structural characteristics that correspond to four information-processing demands: screen size and working memory, screen clutter and perceptual speed and visual acuity, response interface and psychomotor ability, and permissibility and selective attention. Accordingly, the model differentiates between UIT devices in terms of the extent to which construct-irrelevant cognitive load, in the form of information-processing demands, is engendered by devices' structural characteristics. The present study is primarily concerned with the propositions pertaining to the permissibility and selective attention component of the model; these are detailed below.

Permissibility and Selective Attention

Permissibility, one of four structural characteristics of UIT devices outlined by the SCIP model, refers to the flexibility that test takers have in terms of the environment wherein they choose to complete UITs. Gray et al. (2015) conceptualize the permissibility of a device as the extent to which the device allows test takers to complete tests in private versus public spaces, while being static versus moving, and indoors versus outdoors. For example, desktop computers would be less permissible than smartphones because they are often confined indoors (requiring electricity via traditional electrical outlets and sometimes hardwired to an Ethernet drop [no wireless card]) and require users to remain static, whereas smartphones may be used indoors or outdoors and while stationary or moving. It is precisely these device-dependent variations in permissibility that correspond to the extent to which distractions are present while the test taker completes an assessment. Whereas a desktop computer is most frequently used in a static, fixed location, a smartphone may be used in any location where one has sufficient cellular or wireless Internet (i.e., Wi-Fi) reception.

Selective attention, the information-processing demand associated with the structural characteristic of permissibility, refers to the ability to focus attention on a task while in the presence of distracting environmental variables (Arthur, Doverspike, & Bell, 2004). Selective attention has been demonstrated to require effortful cognitive control (Lavie, Hirst, de Fockert, & Viding, 2004) and is related to several outcomes including driving accident involvement and performance on complex tasks (Arthur & Doverspike, 1992; Arthur et al., 1995; Arthur, Strong, & Williamson, 1994).

Per the SCIP model, permissibility engenders construct-irrelevant cognitive load in the form of selective-attention demands such that a high degree of permissibility coincides with a high degree of selective-attention demands. Although user-specific, the greater degree of permissibility afforded by mobile devices coincides with a greater *potential* for test takers to complete UITs in high selective-attention demanding environments. That is, the extent to which a UIT device elicits additional selective-attention demands is *at the discretion of the test taker*, as it is the test taker who ultimately decides where to complete a UIT.

Distinguishing Between Devices

Arthur, Keiser, and Doverspike (2018) posit that conceptualizing UIT devices in terms of wired versus wireless, or "nonmobile" versus "mobile," respectively, does not provide any scientific insight as to why different devices (e.g., smartphones versus desktop computers) elicit differential outcomes in the form of test and assessment scores. Rather, it is argued that it would be more advantageous to conceptualize UIT devices in terms of the extent to which they elicit construct-irrelevant cognitive load. By differentiating UIT devices in terms of the additional cognitive load (i.e., construct-irrelevant information-processing demands) that any given device engenders, any existing or future device can be integrated within the framework instead of merely being labeled as "mobile" or "nonmobile."

To illustrate the advantages of differentiating between UIT devices using the SCIP framework instead of the typical mobile versus nonmobile conceptualization, consider the differences between laptop computers and smartphones. Whereas both laptop computers and smartphones are mobile (or wireless) devices, according to the SCIP model, one would expect the two to engender differential amounts of additional cognitive load according to each device's structural characteristics. That is, the extent to which each device elicits additional cognitive load in the form of construct-irrelevant information-processing demands depends on its structural characteristics. For example, it is apparent that smartphones provide test takers with more degrees of freedom in terms of where the test can be completed. So, if a test taker chooses to complete an employment-related test with a smartphone *and* does so in an environment that demands a great degree of *construct-irrelevant* selective attention (e.g., to concentrate amid distractions), the additional cognitive load engendered by such demands should compete for one's cognitive resources and, thus, adversely affect test scores.

In summary, the SCIP model is a conceptual framework that seeks to explain the observed UIT device-type effects on cognitive measures, and lack thereof on noncognitive measures, in a manner consistent with the vast psychological literature regarding information processing. However, many of the propositions derived from the model have yet to be empirically tested. Consequently, the objective of the present study is to test the propositions pertaining to selective attention (one of four information-processing demands) and permissibility (one of four structural characteristics).

In accordance with the selective attention and permissibility component of the SCIP model, in the present study, smartphones are considered as representative of UIT devices that are high in permissibility and characterized by greater selective-attention demands due to the extent to which they are able to be used in a variety of contexts, which directly corresponds to a greater *potential* for distractions. In contrast, desktop computers represent the low end of UIT device-type permissibility and are characterized by lower selective-attention demands due to the restricted context in which the test taker uses them. By differentiating between UIT devices with respect to the extent the device permits greater degrees of freedom in terms of the test-taking environment, any additional construct-irrelevant selective-attention demands engendered by each device can be subsequently compared and contrasted.

The Present Study

The present study tested two of the propositions derived from the tenets of the SCIP model provided by Arthur, Keiser, and Doverspike (2018). These propositions pertain to (1) UIT device-type effects in the form of observed test-score differences, and (2) the relationships between test scores and selective attention, the information-processing variable associated with permissibility, as outlined by the SCIP model. In addition to the focal study variables, subjective device usability was also measured in order to determine whether participants in both conditions similarly perceived the usability of the randomly assigned device used to complete the cognitive test and personality assessment. Those in the desktop condition completed the assessments on a prototypical desktop computer, whereas those in the smartphone condition completed the assessments on their personal smartphone.

Test Performance

The present study predominantly concerned the testing environment explanation for the aforementioned discrepancy between operational and nonoperational studies. Given the nature of wireless devices, test takers are able to remotely access the Internet in virtually any location, hence, it is quite possible that the observed UIT device-type effects are attributable to those using such devices simultaneously choosing to complete employment-related assessments in a distracting environment. For example, when completing a cognitively demanding test (e.g., GMA), to the extent that the UIT device (e.g., smartphone) permits the test taker more options in terms of where to complete the test *and* the test taker chooses to complete the test in a suboptimal testing environment (e.g., loud, crowded), additional construct-irrelevant cognitive demands (here, selective attention) should result in lower test scores compared to if the same test is completed on a device (e.g., desktop computer) that restricts the choice of environment, thereby restricting the presence of additional construct-irrelevant selective-attention demands.

Being a similarly situated, nonoperational study, it is important to note UIT device-type effects are not necessarily expected to be found by simply randomly assigning individuals to use a desktop computer or smartphone to complete a cognitive test. Rather, atypical for such studies, the present study administered the cognitive and noncognitive assessments in two distinctly dissimilar environments in order to provide a rigorous test of whether the test-taking environment accounts for the observed UIT device-type score differences, which, per the SCIP framework, would be due to additional construct-irrelevant selective-attention demands placed on the test taker. Specifically, those in the desktop condition completed the cognitive test while stationary in an indoor research laboratory akin to the aforementioned nonoperational studies that have failed to find UIT device-type effects on cognitive-test scores; in contrast, those in the smartphone condition completed the cognitive test *while moving* (i.e., walking) in a *distracting, outdoor environment*. These differences in test-taking environments directly correspond to each of the three factors proposed by Gray et al. (2015) to reflect the degree of permissibility: private versus public space, indoors versus outdoors, and static versus moving.

In contrast to cognitive measures, noncognitive measures (e.g., personality inventories) have been found to be influenced by construct-irrelevant cognitive demands to a lesser extent. By nature, cognitive measures engender cognitive load, so any additional construct-irrelevant cognitive demands (i.e., selective-attention demands) that are present compete for the test taker's limited cognitive resources. On the other hand, noncognitive measures do not require cognitive effort to the extent that cognitive measures do, so additional construct-irrelevant selective-attention demands should interfere with individuals' ability to complete such measures to a much lesser degree. Therefore, differences in noncognitive-assessment scores between the two conditions are not expected to be found, even when the assessment is completed in a distracting environment. In accordance with the propositions derived from the SCIP model, it is hypothesized that:

> ***Hypothesis 1.*** When a smartphone (a highly permissible device) is used to complete a GMA test in a distracting environment, mean scores will be lower than scores obtained

on the same test completed using a desktop computer (a device at the low end of permissibility) in an environment conducive to testing.

Hypothesis 2. When a smartphone is used to complete a noncognitive assessment in a distracting environment, mean scores will *not* be different from scores obtained on the same test completed using a desktop computer in an environment conducive to testing.

Selective Attention Relationships

To the extent that cognitive resources are finite, any (construct-irrelevant) cognitive demands beyond those elicited by a cognitively loaded test should compete for cognitive resources, thus affecting one's ability to optimally perform on a cognitive test. Per the SCIP model, the permissibility of a UIT device coincides with the presence of construct-irrelevant selective-attention demands such that test takers using highly permissible devices (i.e., smartphones) are afforded greater degrees of freedom in terms of potential testing environments, which translates into a greater likelihood to complete cognitively loaded UITs in distracting (i.e., selective-attention demanding) environments. Accordingly, the ability to adequately perform in such environments should depend on one's selective attention to a much greater extent than when completing tests in environments that are conducive to testing. Therefore, among those who complete cognitive tests in distracting environments, selective-attention test scores should positively covary with scores on the GMA test such that higher GMA scores coincide with higher selective-attention scores.

In contrast, because noncognitive measures require fewer cognitive resources compared to their cognitively loaded counterparts, construct-irrelevant selective-attention demands should not impede one's ability to adequately complete such assessments, even when the assessment is completed in a distracting environment. That is, selective-attention scores should be largely unrelated to personality-assessment scores regardless of the test-taking environment, so a similar relationship between selective attention and personality-assessment scores should be observed for those using smartphones and those using desktops because the selective-attention demands are not competing for cognitive resources.

Hypothesis 3. When smartphones are used to complete a GMA test (i.e., a cognitively loaded construct) in a distracting environment, there will be a stronger observed relationship between selective-attention and GMA test scores compared to the same relationship observed for those completing the test on a desktop computer in an environment conducive to testing.

Hypothesis 4. When smartphones are used to complete a personality assessment (i.e., a noncognitively loaded construct) in a distracting environment, the observed relationship between selective-attention and personality-assessment scores will be weak and similar to that observed for those using desktop computers to complete the assessment in an environment conducive to testing.

2. METHOD

Participants

Prior to data collection, a sensitivity power analysis was conducted in G*Power 3.1 (Erdfelder, Faul, & Buchner, 1996; Faul, Erdfelder, Buchnar, & Lang, 2009) in order to estimate the requisite sample size for detecting the hypothesized effects. The most conservative hypothesis (Hypothesis 3) was used for the statistical power calculation. Hence, effect sizes (Cohen's q; the difference between two Fisher z-values following Fisher's z' transformation) drawn from the literature that has examined differences in the relationship between information-processing variables and cognitively loaded test performance as a function of the UIT device used to complete said test were used (e.g., Arthur et al., 2018). Alpha and power parameters for the analysis were set at .05 and .80, respectively. Results necessitated a sample size of 2,480 in order to detect a small effect ($q = .10$) and a sample size of 626 for a small-to-medium effect ($q = .20$). Given that obtaining a sample of this magnitude was not feasible, a sample of 500 was proposed, which translated into a power of .30 to detect a small effect ($q = .10$) and .72 to detect a small-to-medium effect ($q = .20$). The study was able to obtain only 261 participants, however, which necessitated an effect size of $q = .31$ to achieve .80 statistical power (again, using the most conservative hypothesis). The implications of this particular sample size are addressed in the Discussion section.

Participants were individuals aged 18 years or older and recruited from the Texas A&M psychology subject pool. The sample was comprised of 166 females (64%), predominately White (63%), and the average age was 18.90 (SD = 1.20). Of the 261 participants, 149 were randomly assigned to the smartphone condition and 112 to the desktop condition.

Measures

All measures with the exception of the GMA test[1] (including supplemental measures) are presented in Appendix A.

General mental ability. GMA was operationalized as scores on a timed, 60-item (36 verbal and 24 quantitative), 4-alternative, multiple choice test (Arthur, 2017). Participants were allotted 10 minutes to complete the test. Convergent validities have been reported with standardized test scores (ACT and SAT; r = .42-.55), and criterion-related validities (rs) of .24-.29 for GPA, and .32 for supervisory ratings of job performance have been obtained. Naber, Arthur, Edwards, and Franco-Watkins (2016) reported 7-to-10 day, test-retest reliabilities of .76 and .70 for the scores obtained from two alternative forms of the test. Scores were calculated as the number of items that participants correctly responded to, so the maximum score one could receive was 60.

Selective attention. Selective attention was operationalized as scores on the 12-item, computer-administered visual-attention test (CA-VAT; Arthur, 1991; Arthur et al., 1995). Each item consisted of two blocks of multiple trials, and each trial consisted of one stimulus pair (number and letter or two letters; e.g., "2 L," "K 4," "H P"), which were presented at a rate of one stimuli pair per two seconds. Prior to the presentation of each block, a cue word (i.e., "COFFEE" or "APPLE") was presented on the screen for two seconds. Test takers were required to remember the prespecified rules associated with each cue word and respond to each stimuli pair accordingly using prespecified keyboard keys. For example, if "COFFEE" appeared prior to a block of stimuli pairs, test takers were instructed to respond with the left arrow key if the stimuli pair was an odd number displayed to the left of a letter (e.g., "3 H"). Strong (1992) reported a test-retest reliability of .83, and internal consistency estimates ranging .84-.98 have

[1] The GMA test is proprietary and cannot be reproduced.

been obtained (Arthur & Day, 2009; Arthur, Strong, & Williamson, 1994). Scores for each item were calculated as the number of trials (i.e., stimuli pairs) that participants correctly responded to for that particular item. The final score was calculated by summing the 12 items such that the maximum score one could receive was 242 (which necessarily equals the total number of trials). The estimated internal consistency reliability at the level of the 12 items was .99.

Personality. Personality was operationalized as scores on a subset of the 50-item International Personality Item Pool (IPIP) measure of the five-factor personality dimensions (Goldberg, 1999). Specifically, only the agreeableness and conscientiousness dimensions of the measure (10 items each) were administered. Participants responded to the items on a 5-point Likert scale (1 = very inaccurate; 5 = very accurate) to indicate the extent to which each item statement was descriptive of themselves. Cronbach's alpha was .79 and .65 for the conscientiousness and agreeableness scores, respectively. Scores were obtained by summing all of the responses for each dimension's item set, so the maximum score one could receive for each dimension was 50.

Device usability. Device usability was operationalized as scores on the 10-item System Usability Scale (SUS; Brooke, 1996) and served as a potential control variable, in that it was used to examine whether there were differences in the perceived usability of the devices to which participants had been randomly assigned (i.e., a desktop computer or one's personal smartphone). An example item is "I felt very confident using this smartphone/desktop computer." Either "smartphone" or "desktop computer" was used depending on one's randomly assigned condition. Participants responded on a 5-point Likert scale (1 = strongly disagree; 5 = strongly agree). The estimated Cronbach's alpha for the scores was .75. Higher scores reflected more favorable reactions regarding the usability of the randomly assigned device, and the maximum score one could receive was 50.

Design and Procedure

The present study employed a between-subjects, experimental research design. Participants first read and subsequently signed an informed consent form stating the purpose and instructions for the study. Upon consenting to participate in the study, individuals were randomly assigned to one of two conditions: desktop computer (n = 112) or smartphone (n = 149). The two device types (i.e., conditions) represent the extreme ends of UIT device-type permissibility posited by the SCIP model (Arthur, Keiser, & Doverspike, 2018). Participants in the smartphone condition were required to bring their own smartphone. For both conditions, all study measures except for the cognitive ability test and personality assessments were administered via desktop computer. The conditions only differed with respect to the location and device type used to complete the cognitive ability test and personality measures (i.e., agreeableness and conscientiousness).

The measures were administered in the following order: (1) selective-attention task, (2) GMA test, (3) personality assessment, (4) system usability measure, (5) attitudes toward mobile testing measure, and (6) demographics questionnaire. Participants in the smartphone condition were directed outside to a busy location on campus to complete the GMA test and personality assessment, whereas participants in the desktop computer condition completed the aforementioned measures in an indoor research laboratory. This difference in testing environments served as a manipulation of each of the factors that reflect the degree of permissibility (Gray et al., 2015): private versus public space, indoors versus outdoors, and static versus moving. Following completion of all of the study measures, participants were debriefed as to the purpose of the study and received subject-pool credit. Each session took approximately two hours to complete, and up to six participants were in any given session.

3. RESULTS

There were 44 participants (17%) who incorrectly responded to a quality check item that was programmed into the battery of measures. Descriptive statistics for and subsequent analyses using the variables of interest were computed for each subgroup (i.e., those who responded correctly and those who responded incorrectly to one [or both]) and revealed no meaningful differences between the subgroups, so all of the participants were retained.

The null hypothesis statistical tests were one-tailed for the analyses concerned with directional hypotheses (i.e., Hypotheses 1 and 3) and two-tailed for every other analysis, including the correlation matrices and supplemental analyses. For the reported effect sizes (*d*s), positive values indicate that the desktop condition scored higher and negative values indicate that the smartphone condition scored higher. The exact number of participants used for each analysis can be extrapolated from their respective degrees of freedom and are made explicit where appropriate. Descriptive statistics and correlations for all of the study's focal variables are presented in Table 1 and are depicted by condition in Table 2.

Table 1

Descriptive Statistics for and Correlations Between Focal Variables

Measure	*N*	*M*	*SD*	1	2	3	4	5	6	7
1. Sex	261	--	--	--						
2. Age	261	18.90	1.20	-.09	--					
3. GMA	261	35.67	7.07	-.02	-.06	--				
4. CA-VAT	260	136.35	73.13	.02	.15*	.12	(.99)			
5. Agreeableness	247	37.87	4.33	.22*	.06	-.03	.04	(.65)		
6. Conscientiousness	247	34.45	6.05	.14*	-.06	-.04	.00	.16*	(.79)	
7. SUS	254	29.22	2.93	-.04	-.07	-.03	-.13*	-.04	-.03	(.75)

Note. Where applicable, internal consistency reliability estimates are provided in parentheses on the diagonal. GMA = general mental ability; CA-VAT = computer-administered visual attention test; SUS = system usability scale. Sex is coded as 0 = Male, 1 = Female.
* $p < .05$ (two-tailed).

Table 2

Descriptive Statistics for and Correlations Between Focal Variables by Condition

Smartphone	Desktop									
	N	*M*	*SD*	1	2	3	4	5	6	7
1. Sex	112	--	--		-.14	.00	.01	.21*	.26*	.03
2. Age	112	19.03	1.33	-.08		.03	.09	.00	-.11	-.09
3. GMA	112	35.26	6.64	-.03	-.17		.18*	.00	-.04	-.10
4. CA-VAT	112	140.89	70.52	.03	.22*	.03		.02	.02	-.16
5. Agreeableness	108	38.30	4.29	.21*	.10	-.08	.06		.12	.07
6. Conscientiousness	108	34.72	5.95	-.06	-.02	-.03	-.02	.20*		.04
7. SUS	111	28.61	2.68	-.07	.00	.06	-.06	-.13	-.09	
N	--	--	--	149	149	149	148	139	139	143
M	--	--	--	--	18.81	35.97	132.92	37.55	34.24	29.70
SD	--	--	--	--	1.08	7.38	75.09	4.35	6.13	3.04

Note. Descriptive statistics for the smartphone (n = 149) and desktop (n = 112) conditions are horizontally and vertically presented, respectively; Correlations for the smartphone and desktop conditions are above and below the diagonal, respectively. GMA = general mental ability; CA-VAT = computer-administered visual attention test; SUS = system usability scale. Sex is coded as 0 = Male, 1 = Female. * $p < .05$ (two-tailed).

Prior to testing the primary hypotheses of interest, an independent samples *t*-test with condition (i.e., smartphone or desktop computer) as the independent variable was used in order to test whether there was a difference in perceived device usability between the two conditions. Levene's test for the equality of variances assumption indicated that the assumption held, $F(1, 252) = 1.31$, $p > .05$. The *t*-test revealed a statistical difference in device usability between the desktop ($M = 28.61$, $SD = 2.68$) and smartphone ($M = 29.70$, $SD = 3.04$) conditions, $t(252) = -2.98$, $p < .05$, $d = -0.38$ (see Figure 1), indicating that those in the smartphone condition perceived the usability of the device they were using (i.e., their own smartphone) slightly more favorably than those in the desktop condition (i.e., prototypical desktop computer).

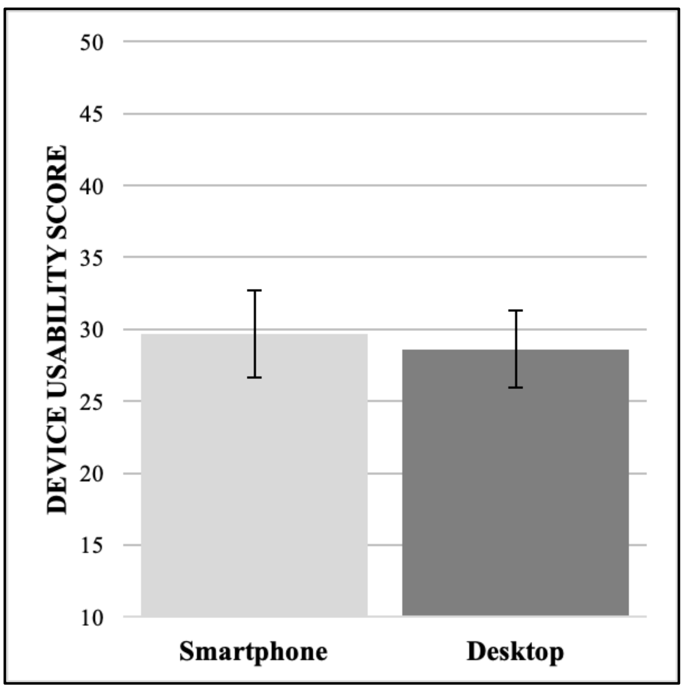

Figure 1. System usability by condition. Error bars represent the standard deviation of each condition's scores.

Hypotheses 1 and 2 were subjected to an independent samples *t*-test with condition as the independent variable. The dependent variable was GMA test scores for Hypothesis 1, which stated that those using a smartphone outdoors to complete the GMA test would perform worse than those using a desktop computer indoors. Levene's test for the equality of variances assumption indicated that the variances for GMA test scores were statistically equal, $F(1, 259) = 1.00$, $p > .05$. The results from the subsequent *t*-test indicated that the means for the desktop condition ($M = 35.26$, $SD = 6.64$) and the smartphone condition ($M = 35.97$, $SD = 7.38$) were not statistically different from each other, $t(259) = -0.81$, $p > .05$, $d = -0.10$ (see Figure 2). Thus, Hypothesis 1 was not supported; it appears that varying both the device one used and the environment wherein one completed the cognitive test did not substantively affect the observed scores.

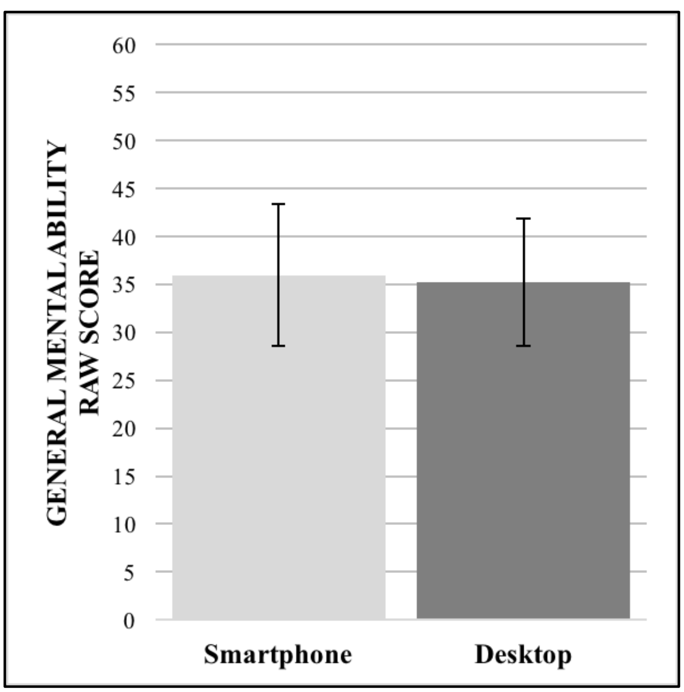

Figure 2. General mental ability by condition. Error bars represent the standard deviation of each condition's scores.

Agreeableness and conscientiousness scores were used as the dependent variables for Hypothesis 2, which stated that scores on the two noncognitive assessments would not differ between the two conditions. Levene's test for the equality of variances assumption indicated that the conditions' variances for both conscientiousness, $F(1, 245) = 0.01$, $p > .05$, and agreeableness assessment test scores were equal, $F(1, 245) = 0.02$, $p > .05$. The subsequent independent samples *t*-tests indicated no statistical difference between the desktop ($M_C = 34.72$, $SD_C = 5.95$; $M_A = 38.30$, $SD_A = 4.29$) and smartphone ($M_C = 34.24$, $SD_C = 6.13$; $M_A = 37.55$, $SD_A = 4.35$) conditions on conscientiousness, $t(245) = 0.62$, $p > .05$, $d = 0.08$, and agreeableness, $t(245) = 1.35$, $p > .05$, $d = 0.17$ (see Figure 3). As hypothesized, no meaningful evidence of a difference in noncognitive (here, both agreeableness and conscientiousness) assessment scores was found, even when varying the test-taking environment.

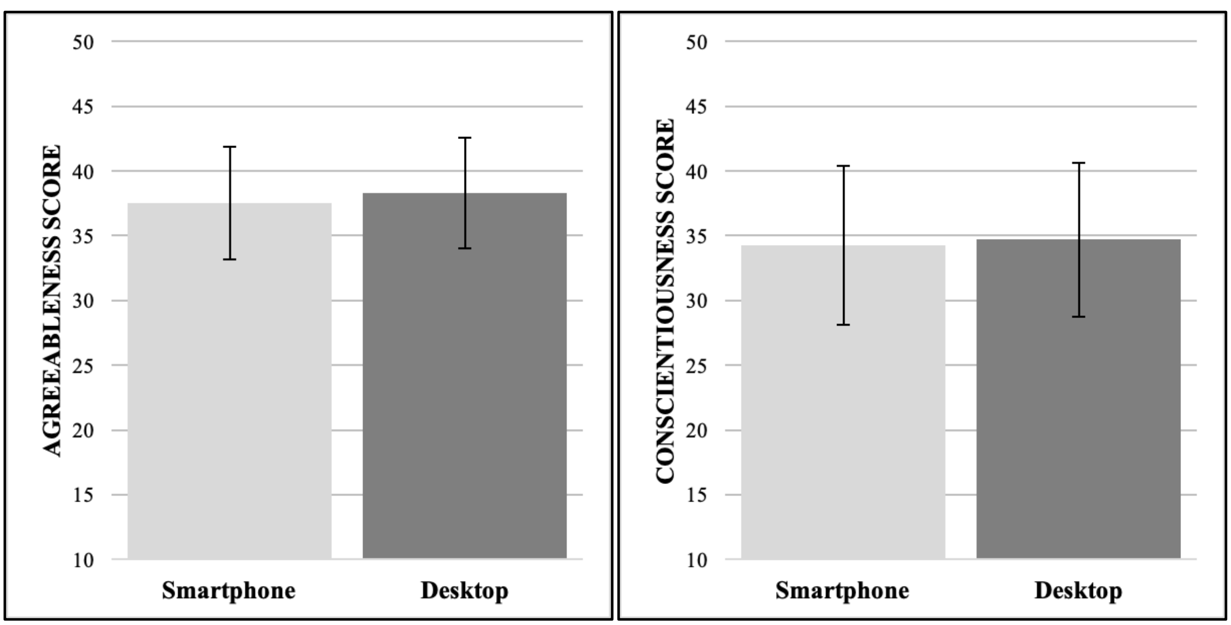

Figure 3. Agreeableness (left) and conscientiousness (right) by condition. Error bars represent the standard deviation of each condition's scores.

To test Hypotheses 3 and 4, correlations of interest were transformed using Fisher's *z*' transformation and tested for a statistically significant difference. To test the former, which stated that the relationship between selective-attention and GMA test scores would differ according to the device used to complete the GMA test (i.e., smartphone condition's relationship > desktop condition's relationship), the correlations between scores on the CA-VAT and scores on the GMA test for each condition (smartphone, $r(146) = .18$, $p < .05$; desktop, $r(110) = .03$, $p > .05$) were compared using an asymptotic z-test and found to not statistically differ, $z = 1.20$, $p > .05$, $q = .15$. Although the hypothesis did not receive statistical support, the difference in the magnitude of the correlations was in the hypothesized direction, namely, the smartphone condition's GMA/CA-VAT correlation was larger than that observed for the desktop condition.

For Hypothesis 4, which stated that the conditions' relationship between selective-attention and personality assessment scores would be comparable in magnitude and strength, the correlations between scores on the CA-VAT and scores on the agreeableness and conscientiousness dimensions of the IPIP for each condition were compared using an asymptotic z-test. The correlations for agreeableness (smartphone, $r[136] = .02$, $p > .05$; desktop, $r[106] = .06$, $p > .05$) did not statistically differ, $z = -0.31$, $p > .05$, $q = .04$, nor did the correlations for conscientiousness (smartphone, $r[136] = .02$, $p > .05$; desktop, $r[106] = -.02$, $p > .05$), $z = 0.31$, $p > .05$, $q = .04$. In accordance with what was hypothesized, these results suggest that noncognitive assessments are not particularly affected by selective-attention demands.

Supplemental Analyses

In addition to the preceding tests of the focal hypotheses, additional data were available and provided the ability to make additional comparisons between the two conditions. These supplemental data correspond to scores on (1) King et al.'s (2015) attitudes toward mobile testing measure, and (2) three items regarding distractibility while completing the GMA test. The entire sample completed the former measure, but only a subset completed the latter (smartphone n =

128, desktop n = 81). Akin to the preceding analyses, positive *d*s reflect higher scores for the desktop condition, and the exact number of participants used for each analysis can be extrapolated from their respective degrees of freedom.

Attitudes toward testing with smartphones. Attitudes toward the use of smartphones for employment-related testing purposes was operationalized as scores on King et al.'s (2015) 7-item attitudes toward mobile testing measure. Participants were instructed to respond to items using the particular device to which they were randomly assigned to complete the GMA test and personality assessment as a frame of reference. The first item is different from the rest and simply asks "How would you compare completing employment-related tests on a smartphone versus a desktop computer?" For this item, participants responded by selecting either "better on smartphone," "as good on a smartphone," or "worse on a smartphone." Participants responded to the remaining six items on a 5-point Likert scale (1 = strongly disagree; 5 = strongly agree), which were summed to form a scale score. An example item is "I would prefer to complete tests on a smartphone versus completing them on a desktop computer." The maximum score one could obtain was 30, and higher scores indicate more favorable attitudes toward using smartphones for employment-related testing purposes. Cronbach's alpha was .75 for the scores. Descriptive statistics for the attitudes toward mobile testing scores (i.e., items 2-7) as well as the correlations between them and the study's focal variables are displayed in Table 3.

Table 3

Descriptive Statistics for and Correlations with Attitudes Toward Testing with Smartphones

Measure		Total (N = 258)	Smartphone (n = 147)	Desktop (n = 111)
1. Sex		-.01	-.03	.06
2. Age		.00	.06	-.06
3. GMA		.12	.19*	-.02
4. CA-VAT		-.04	.04	-.15
5. Agreeableness		.09	.12	.05
6. Conscientiousness		-.10	-.14	-.02
7. SUS		.12	.22*	-.10
	M	16.69	17.01	16.26
	SD	4.24	4.63	3.65

Note. GMA = general mental ability; CA-VAT = computer-administered visual attention test; SUS = system usability scale. Sex is coded as 0 = Male, 1 = Female. * $p < .05$ (two-tailed).

Fisher's exact test revealed a statistical difference between the conditions' endorsement of the first item's three response options, $X^2(2) = 22.08$, $p < .05$, $V = .45$. Whereas about only half of those randomly assigned to the smartphone condition indicated that they believed using smartphones was worse than desktop computers for completing employment-related tests, approximately three-quarters of those in the desktop condition did. Additionally, in contrast to those in the smartphone condition, fewer participants in the desktop condition thought using a smartphone to complete such tests would be (a) as good as or (b) better than using a desktop computer (see Figure 4).

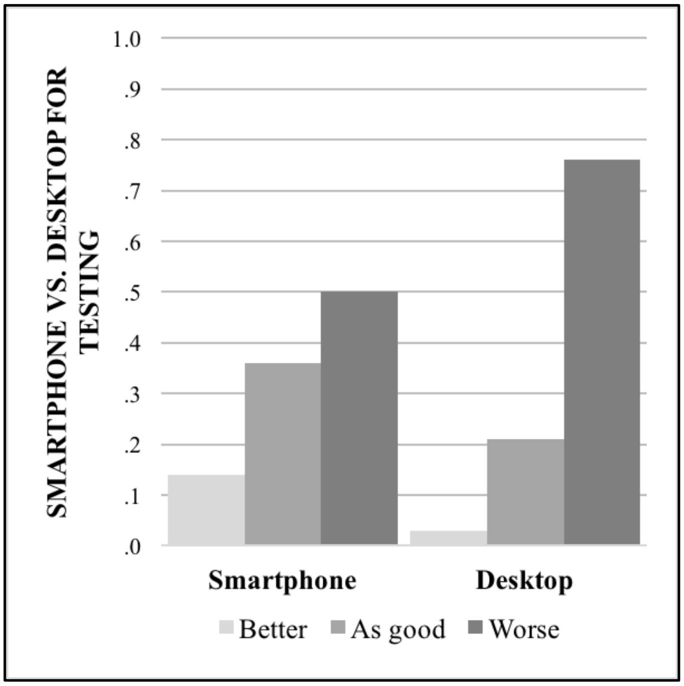

Figure 4. Testing device preference by condition. Figure reflects the proportion of endorsement to response options on an item that asked participants to compare completing employment-related tests on a smartphone versus a desktop computer. Error bars represent the standard deviation of each condition's scores.

Scores (aggregate of the latter six items) on the attitudes toward testing with smartphones measure were subjected to an independent samples *t*-test with condition as the independent variable. Levene's test for the equality of variances assumption indicated that the variances statistically differed, $F(1, 256) = 6.21$, $p < .05$. Welch's *t*-test for unequal variances revealed that the desktop ($M = 16.26$, $SD = 3.65$) and smartphone ($M = 17.01$, $SD = 4.63$) conditions did not statistically differ with respect to attitudes toward testing with smartphones, $t(255.48) = -1.45$, $p > .05$, $d = -0.18$ (see Figures 5 and 6). So, although those in the smartphone condition reported more favorable attitudes toward the use of smartphones for completing employment-related tests, as illustrated by the two conditions' proportions of responses to the first item, the two conditions had relatively equal attitudes toward the use of smartphones for employment-related testing purposes (i.e., approximately neutral) per the composite of the latter six items (see Figure 5).

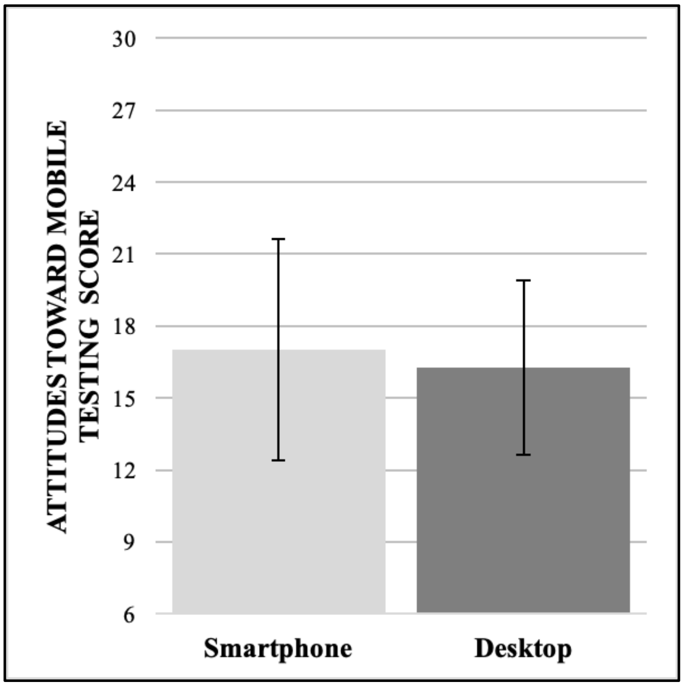

Figure 5. Attitudes toward mobile testing by condition. Error bars represent the standard deviation of each condition's scores.

Distractions while testing. Distractibility while completing the GMA test was operationalized as scores on three items developed for the purposes of the present study. Participants responded to the first item—"Did you experience any distractions while taking the test that you just completed?"—by simply selecting "yes" or "no." The second and third items concerned the extent to which the testing environment was distracting and the extent to which it was difficult to concentrate, respectively, and were responded to using a 5-point Likert scale (item two: 1 = not at all distracting; 5 = very distracting; item three: 1 = strongly disagree; 5 =

strongly agree). Descriptive statistics for each of the items as well as their correlations with the study's focal variables are displayed in Table 4.

Table 4

Descriptive Statistics for and Correlations with Three Distraction Items

Measure	*N*	Distractions	Extent of Distractions	Difficulty Concentrating
1. Distractions	208	--		
2. Extent of Distractions	208	.65*	--	
3. Difficulty Concentrating	208	.60*	.82	--
4. Condition	208	.45*	.48*	.52*
5. Sex	208	-.09	-.04	.02
6. Age	208	-.09	-.08	-.11
7. GMA	208	.08	-.03	-.04
8. CA-VAT	207	.01	-.06	-.06
9. Agreeableness	197	.09	.04	.03
10. Conscientiousness	194	-.08	-.10	-.08
11. SUS	203	.04	.00	.03
12. Attitudes	205	.11	-.06	-.03

Condition	*No*	*Yes*	*M*	*SD*	*M*	*SD*
SP (*n* = 128)	45 (35%)	83 (65%)	2.47	1.08	1.40	0.72
DT (*n* = 81)	65 (81%)	15 (19%)	2.69	1.20	1.38	0.72

Note. GMA = general mental ability; CA-VAT = computer-administered visual attention test; SUS = system usability scale. Distractions (Measure 1) coded as 0 = No, 1 = Yes; Sex is coded as 0 = Male, 1 = Female; Condition is coded as 0 = Desktop, 1 = Smartphone. * $p < .05$ (two-tailed).

Each of the three distraction items were examined in isolation. For the first item, a chi-square test using Yates' correction for continuity was conducted and indicated that the desktop and smartphone conditions statistically differed with respect to their proportions of endorsement of the two response options, $X^2(1) = 40.15$, $p < .05$, $V = .29$. Those in the smartphone condition were more likely to report experiencing distractions while completing the GMA test compared to those in the desktop condition. Specifically, roughly 65% of those in the smartphone condition reported experiencing distractions, whereas only approximately 19% of those in the desktop condition did so (see Figure 6).

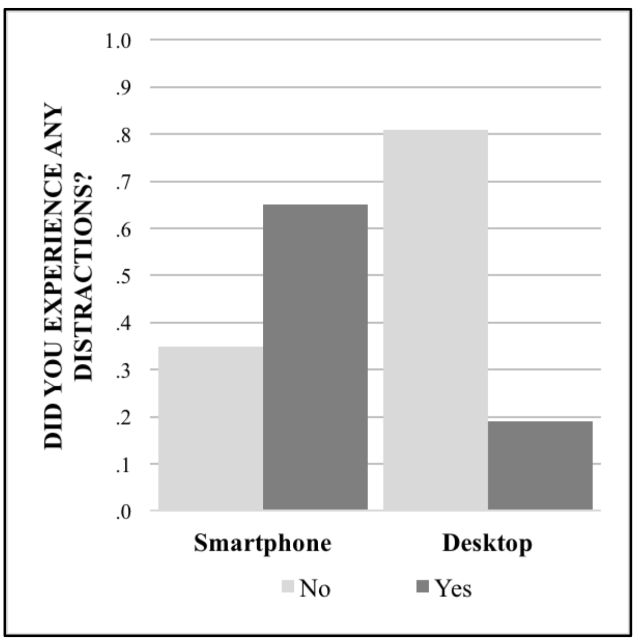

Figure 6. Distractions while testing by condition. Figure reflects each conditions' proportion of endorsement to an item that asked whether distractions were experienced while completing the GMA test. Error bars represent the standard deviation of each condition's scores.

For the second distraction item, Levene's test for the assumption of equal variances indicated that the variances statistically differed, $F(1, 207) = 19.82$, $p < .05$. Welch's *t*-test for unequal variances indicated that the desktop ($M = 1.40$, $SD = 0.72$) and smartphone ($M = 2.47$, $SD = 1.08$) conditions statistically differed with respect to the extent to which participants were distracted while completing the GMA test, $t(206.42) = -8.63$, $p < .05$, $d = -1.17$. On average, those in the smartphone condition reported a greater degree of distraction while completing the GMA test compared to the desktop condition (see Figure 7).

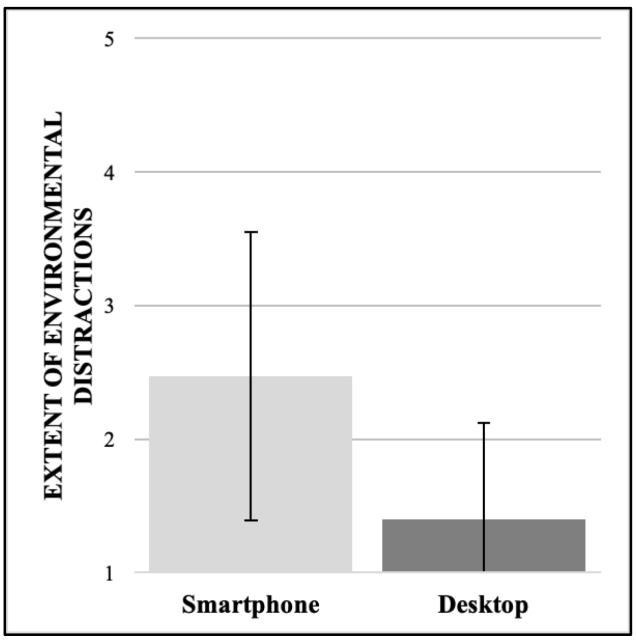

Figure 7. Environmental distractions by condition. Error bars represent the standard deviation of each condition's scores.

For the final distraction item, Levene's test for the equality of variances assumption suggested that the variances did statistically differ, $F(1, 207) = 47.46$, $p < .05$. Using Welch's *t*-test for unequal variances, it was found that the desktop ($M = 1.38$, $SD = 0.72$) and smartphone ($M = 2.69$, $SD = 1.20$) conditions statistically differed regarding the extent to which participants had difficulty concentrating while completing the GMA test, $t(206.50) = -9.86$, $p < .05$, $d = -1.32$ (see Figure 8). As expected, those in the smartphone condition reported greater difficulty concentrating compared to those in the desktop condition (see Figure 8).

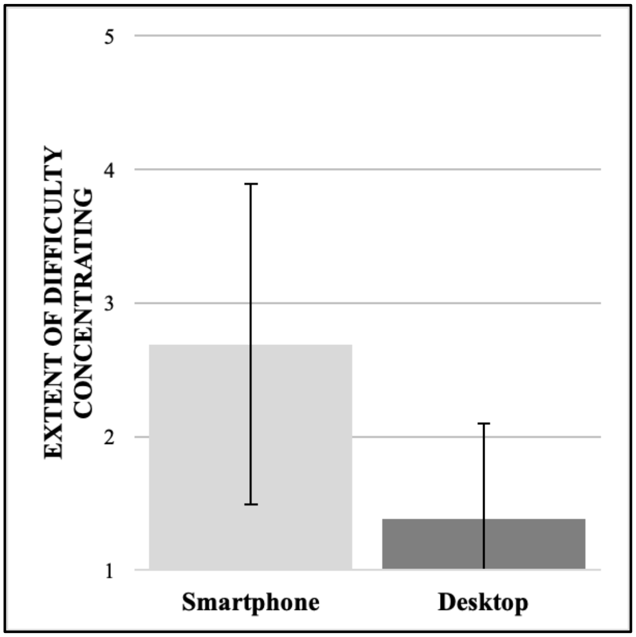

Figure 8. Difficulty concentrating by condition. Error bars represent the standard deviation of each condition's scores.

4. DISCUSSION AND CONCLUSION

The present study sought to test the propositions pertaining to selective attention—the information-processing demand elicited by permissibility, the associated UIT-device structural characteristic—explicated by Arthur, Keiser, and Doverspike's (2018) SCIP model. It was hypothesized that the GMA test scores observed among those using smartphones (i.e., high permissibility) to complete the test would be (1) lower than the scores observed among those using desktop computers (i.e., low permissibility) to complete the same test, and (2) more strongly related (i.e., correlated) to selective-attention test scores compared to the same relationship observed for those using desktop computers to complete the GMA test. In contrast, with respect to noncognitive constructs, it was hypothesized that the observed personality-assessment scores would (1) not differ between the two conditions, and (2) relate weakly to selective-attention test scores for both conditions.

The hypotheses regarding cognitive measures were generally not statistically supported. GMA test score means between the conditions only differed by a fraction of an item (Hypothesis 1), and the correlations between GMA and selective-attention test scores were not statistically significantly different from one another (Hypothesis 3). The estimated coefficients examined in isolation did, however, lend tentative support to the third hypothesis. GMA and selective-attention test scores were (1) positively correlated and statistically differed from zero for the smartphone condition (r = .18), and (2) basically uncorrelated for the desktop condition (r = .03). Consonant with what was expected, this pattern of results suggests that (1) the presence of selective-attention demands while completing the GMA test differed between the two conditions, and (2) a greater degree of selective-attention demands were experienced by those who used smartphones to complete the test. Despite the encouraging pattern of results, the preceding should be interpreted cautiously considering the correlation coefficient estimates did not statistically differ between the two conditions. In summary, no mean differences in cognitive-test scores and limited selective-attention effects were found.

For the set of hypotheses concerned with noncognitive assessment scores, the results coincided with what was anticipated. There was no statistical difference in conscientiousness and agreeableness assessment scores between the two conditions (Hypothesis 2), similar to what has been consistently found by prior research on UIT device-type effects on employment-related test and assessment scores (e.g., Arthur et al., 2014; Arthur et al., 2018; McClure Johnson & Boyce, 2015). In addition to the device that was used to complete the assessments, the results suggest that the environment wherein the assessments were completed did not seem to affect the observed scores. With respect to the relationship between selective attention and (a) conscientiousness and (b) agreeableness assessment scores (Hypothesis 4), there was no meaningful difference between the two conditions. The estimated correlation coefficients between selective attention and (a) conscientiousness (r_{SP} = .02; r_{DT} = -.02) and (b) agreeableness (r_{SP} = .02; r_{DT} = .06) were trivial and similar in magnitude. In sum, consistent with the only other empirical examination of propositions derived from the SCIP model to date (i.e., those pertaining to working memory [instead of selective attention]; Arthur et al., 2018), the present study found support for the hypotheses regarding noncognitive measures, but only found tentative support for those concerning cognitive measures.

The absence of a difference in observed GMA test scores between the two device-type conditions is particularly surprising considering the environment wherein those in the smartphone condition completed the test was in sharp contrast to prior nonoperational studies where *all* test takers completed the test in a proctored environment conducive to testing. Despite the failure to observe a difference in GMA test scores, several supplemental analyses were conducted and revealed an interesting pattern of results. For example, when asked whether any distractions were experienced while completing the cognitive test, the majority (65%) of those in the smartphone condition responded in the affirmative, whereas only a minority (19%) of those in the desktop condition reported experiencing distractions. Furthermore, for both the extent to which one (a) found the environment distracting, and (b) had difficulty concentrating while completing the cognitive test, those in the smartphone condition responded less favorably (i.e., greater degree of being distracted and experiencing difficulty concentrating) than those in the desktop condition (correlations with condition [0 = DT; 1 = SP] were .48 and .52, respectively). Despite these apparent differences, however, the correlations between each of the three distraction items and each of the focal variables were negligible and ranged from -.11 to .11. Indeed, upon closer examination of the smartphone condition's GMA test scores, no statistical difference between those who did not ($M = 35.42$; $SD = 7.36$) and those who did ($M = 36.78$; $SD = 7.66$) self-report experiencing distractions was found, $t(126) = -0.97$, $p > .05$, $d = -0.18$.

Consonant with Arthur et al. (2018), further exploratory, post-hoc analyses found that the two conditions did statistically differ ($M_{DT} = 168.19$s, $SD_{DT} = 35.22$s; $M_{SP} = 180.50$s, $SD_{SP} = 41.70$s) with respect to the amount of time it took to complete the personality assessment, $t(249) = -2.49$, $p < .05$, $d = -0.32$. So, whereas the expected difference in GMA scores was not found, those in the smartphone condition disproportionately reported (a) experiencing distractions, (b) a greater extent of distractions, (c) having difficulty concentrating compared to those in the desktop condition, and (d) took longer to complete the personality assessment. In other words, despite the apparent discrepancy between the two conditions on each of the three self-reported distraction items, the distractions did not translate into the expected observable differences in GMA test scores; however, replicating Arthur et al. (2018), there were differences in the amount of time it took to complete the personality assessment.

The pattern of results observed in the present study highlight the continued need to investigate the circumstances under which UIT device-type effects on cognitive-test scores occur. For instance, perhaps the conflicting literature is due to a generational effect whereby younger individuals are better equipped to complete cognitively demanding tests in distracting environments on their smartphone (or "mobile" device). The majority of those who participated in the nonoperational studies that have failed to find UIT device-type effects on cognitive-test scores, including the present study, belong to a generation that has never witnessed a world in which these technologies were not available and ubiquitous. Given that information-processing abilities such as selective attention decline with age (Plude, Enns, & Brodeur, 1994), the age-restricted range of the present study's sample (i.e., 18-26) may be another potential explanation for the failure to obtain the selective attention effects. Thus, the tendency for younger individuals to gravitate towards the use of smartphones is reflected in Arthur et al.'s (2014) data in which on the basis of applicants' self-reported demographic information, the mean age among applicants who used a mobile device to complete a pre-employment assessment was several years younger than those who did not. In addition, a greater proportion of young applicants used mobile devices compared to the proportion of older applicants who did. More recently, Pew Research Center (2018) revealed that as of early January 2018, 18-to-29 year olds had the greatest proportion of smartphone ownership among Pew's five, age-based categories. They also found that approximately 20% of U.S. adults identified as "smartphone-only"—those who own a smartphone, but do not have wireless Internet connections at home—among which the greatest proportion, again, were those aged 18 to 29. Regardless of their intimate familiarity with such devices, however, whether younger generations are better acclimated to using smartphones in distracting environments is an empirical question that has yet to be investigated.

Implications for Science and Practice

The present findings have important implications for both the scientific and applied communities. Broadly speaking, with respect to academic personnel psychology and consonant with calls to further explore how employment-related testing is affected by technology (Morelli, Potosky, Arthur, & Tippins, 2017), the present study sought to contribute to the literature by empirically testing a psychological explanation (i.e., the SCIP model; Arthur, Keiser, & Doverspike, 2018) for when, how, and why one would expect device-type effects on UITs. To reiterate, the SCIP model provides a way for organizational scholars to posit psychologically sound hypotheses regarding whether or not particular UIT devices are expected to affect observed employment-related test and assessment scores. As with any other empirical investigation, testing the viability of the SCIP model as an explanatory framework has the potential to increase its acceptance and use in subsequent research endeavors involving mobile device use within the context of personnel psychology, in general, and personnel selection, in particular. Furthermore, in accordance with the broader scientific enterprise, subjecting the propositions generated by the model to empirical tests provides organizational scholars the ability to refine (or even reject) the model if and where warranted.

The present study also has several important implications for applied psychology and human resource management. First, the results suggest that the testing environment may not be the source of the mixed findings regarding UIT device-type effects on cognitive-test scores, which consequently lends greater credence to other plausible explanations for these effects (e.g., generational and self-selection hypotheses). Second, although not one of the focal objectives, the present study replicated Arthur et al. (2018) in that those who used smartphones to complete an untimed, noncognitive assessment spent more time completing said assessment than those who used desktop computers, even when the assessment was "optimized for mobile use." Despite the small magnitude of the observed [raw] effect size (i.e., approximately 12 seconds), it is important to consider that the assessments that are administered to applicants for employment-related purposes such as personnel selection are typically quite longer than a handful of self-report items with a Likert rating scale. A much lengthier composite of a host of tests and assessments spanning a variety of constructs is more likely to be what is administered to applicants. Consequently, one would expect a much larger difference in completion times for such full-scale assessments. Moreover, it has been demonstrated that those who choose to use "mobile" devices are more likely to switch devices mid-assessment (e.g., to a desktop computer; Arthur et al., 2014; Dages & Jones, 2015), which further illustrates how the use of such devices for UITs might be troublesome for applicants. All considered, it may be worthwhile for organizations to reevaluate their current UIT procedures, especially with respect to informing test takers (i.e., applicants) about the potential deleterious effects of using such devices.

Limitations and Suggestions for Future Research

There are two primary substantive limitations with the present study. First, the statistical tests used were likely underpowered. Collecting a sample of the originally posited size turned out to be infeasible given the time constraints, hence the discrepancy between recommendations via power analysis and the final sample that was used for the present study. Constraining alpha and power parameters at .05 and .80, respectively, the sample sizes for each condition would have needed to be approximately quadruple the present size (i.e., 149 to 645 [smartphone]; 112

to 484 [desktop]) in order for the most conservative statistical test (i.e., the asymptotic z-test for Hypothesis 3) to reach statistical significance given the observed q of .15 (e.g., $r_{DT} = .00$, $r_{SP} = .15$; $z = 1.64$). Although it is acknowledged that statistical significance is not the be-all/end-all criterion of scientific research, a larger sample size would have led to greater precision in parameter estimation, which should be the goal of any scientific endeavor using statistical analyses to test hypotheses under any inferential methodology.

The remaining limitations concern threats to both internal and ecological validity. The predominant internal validity threat pertains to the confounding of device and environment with respect to the effects of distractibility on test performance. All of those in the desktop condition completed the measures in a quiet, indoor research laboratory, whereas all of those in the smartphone condition completed them while walking around a public, outdoor area with substantial foot traffic. Thus, each condition was restricted in terms of the environment wherein the cognitive and noncognitive measures were completed. Fully crossing the two independent variables (i.e., device permissibility and environmental distractibility) would have afforded the ability to better isolate any effects attributable to the testing environment and distractions therein.

The main ecological validity concern is the evident mismatch between operational contexts and that of the present study. Namely, UITs predominantly (1) occur in high-stakes contexts, and (2) afford test takers (e.g., applicants) the ability to choose the UIT device on which UITs are completed. To the extent that the testing stakes were low for those participating in the present study, and alongside the fact that random assignment to conditions was used, the ecological validity of the observed results is open to critique. With the preceding limitations in mind, the subsequent section elaborates on how organizational scholars might design and conduct successive research that addresses and conceivably avoids concerns such as those in the present study and other similarly situated studies.

The nascence of this stream of research affords a variety of trajectories for prospective scientific endeavors. Per the aforementioned limitations of and conclusions drawn from the present study, one such avenue is a replication and extension of the present study in order to provide further evidence regarding the extent to which environmental distractions might account for the observed UIT device-type effects demonstrated in the literature. Although not exhaustive by any means, the incorporation of more age-diverse samples and eliminating the device-environment confound (e.g., via experimental design) would provide stronger tests of the hypotheses and questions examined here. Accordingly, the underlying cause of UIT device-type effects remains inconclusive at best and warrants further investigations.

As previously mentioned, whether or not UIT device-type effects occur appears to coincide with the context of the particular study—whereas operational studies have consistently demonstrated UIT device-type effects on cognitive-test scores, nonoperational studies have consistently failed to do so. Although several tenable hypotheses for this discrepancy have been raised, most have received scant, if any, empirical attention. In no particular order, others have identified (1) the testing environment, (2) testing stakes, and (3) self-selection of device types as critical study characteristics that would explain the mixed findings between operational and nonoperational studies (e.g., Arthur et al., 2014; Arthur et al., 2018; Arthur & Traylor, in press). The present study is the first to explicitly investigate one of these plausible explanations—the test-taking environment (or distractions therein)—which, per the present findings, appears to not affect cognitive-test performance. So, to the extent that no major flaws render the present findings inconclusive and ungeneralizable to the operational context, greater attention toward testing stakes, self-selection, or perhaps all three as the explanatory mechanism(s) underlying the purported UIT device-type effects is warranted.

The difference in testing stakes between research conducted in operational and nonoperational contexts is readily apparent with respect to the literature concerning UIT device-type effects. Operational data are predominantly obtained from organizations that have administered employment-related tests and assessments to job applicants for personnel selection purposes. Consequently, these testing situations, characterized by applicants' vying for a limited number of job openings, are unequivocally high stakes, and are in stark contrast to the testing situations that characterize nonoperational studies. For those participating in nonoperational studies, most of whom are undergraduate students completing the study for course credit, scores on the assessments *neither help nor impede* their achievement of any particularly important objectives, hence, they are low-stakes testing situations. Stated differently, performance on the administered tests and assessments is *inconsequential* for the test takers, and all receive the same reward (e.g., via subject pool credit) *irrespective* of their test performance. Taking this discrepancy into consideration, there are a number of ways subsequent research, nonoperational or otherwise, might investigate the testing-stakes explanation for the conflicting UIT device-type effect findings. For example, one could incorporate a reward system whereby a prespecified percentage of the top performers are provided with some sought-after prize (e.g., $300 Amazon gift card), which starts to approximate operational applications where, depending on several factors, the top handful or so of applicants receive job offers. Hence, if UIT device-type effects are found, then perhaps the disparity in findings between the two study contexts might be attributable to differences in test-taking motivation between operational and nonoperational samples.

Initially advanced by Arthur et al. (2014), in addition to the divergent testing stakes, the so-called self-selection hypothesis is another viable explanation for the mixed findings. When organizations administer UITs for pre-employment testing purposes, it is frequently the case that applicants are afforded the opportunity to choose their own test-taking device. Consequently, device-type groups within the operational studies are *naturally* formed via applicants self-selecting their test-taking device. In contrast, nonoperational studies generally use random assignment to experimental and control conditions (i.e., "mobile" and "nonmobile," respectively). Under these circumstances, participants are directed to complete the study's tests and assessments on a specified, randomly assigned device instead of selecting their own. Ultimately, then, this particular discrepancy between the two study contexts pertains to how the device-type groups or conditions are formed—random assignment or self-selection. So, to the extent that the preceding is a sound explanation for the conflicting findings, it is unclear whether the observed differences in cognitive-test scores are a product of the device itself (e.g., the SCIP framework) or a self-selection phenomenon. If the latter, it would suggest that differences in cognitive-test scores might not arise from the device, per se, but the characteristics of the individuals (i.e., low cognitive ability) choosing to use particular devices. Put differently, the observed effects might be due to those on the lower end of the cognitive ability spectrum tending to complete assessments on "mobile" devices at a rate incommensurate with those on the higher end. All considered, it is plausible that the combination of self-selection and high testing stakes in addition to the construct-irrelevant information-processing demands engendered by the structural characteristics of the given testing device outlined by the SCIP framework (Arthur, Keiser, & Doverspike, 2018) is the root of the cognitive-test score differences observed within operational studies. If successive nonoperational studies were to incorporate (1) higher testing stakes, (2) self-selection of the test-taking device, or (3) a combination thereof, perhaps the presumed device-type effects consistently demonstrated by operational studies would be observed, corroborating the hypothesis that the effect is actually due to either or both of these study characteristics.

Conclusion

In conclusion, the present study is a response to recent calls for research examining UIT device-type effects on common personnel selection procedures (Morelli et al., 2017). The primary objective was to empirically investigate the propositions pertaining to selective attention that were derived from the SCIP model (Arthur, Keiser, & Doverspike, 2018). Although the observed GMA test scores did not statistically differ between those using smartphones and desktop computers, the results suggest that a greater degree of selective-attention demands may be present when highly permissible devices (e.g., smartphone) are used to complete cognitive tests in distracting environments. Indeed, Arthur et al. (2018) observed the same pattern with working memory (instead of selective attention). That is, despite being in the hypothesized direction, the observed GMA/selective attention (GMA/working memory [Arthur et al.]) relationship did not statistically differ between the conditions. It should be noted, however, that both of these findings are likely attributable to insufficient statistical power coupled with an age-restricted sample. So, with respect to the SCIP model as an explanatory framework, future efforts need be made to examine the propositions pertaining to (1) perceptual speed and visual acuity, and (2) psychomotor ability.

The present findings have implications for both academics and practitioners who use UIT in their research and practice. Concerning the former, although they did not translate into differences in GMA test performance for the present sample, the results suggest that using smartphones to complete a cognitive test in a distracting environment engenders construct-irrelevant selective-attention demands. Arthur et al. (2018) observed a similar pattern of results with respect to working memory demands. Specifically, cognitive test scores and working memory scores correlated more strongly among the smartphone condition compared to the desktop condition (even when both conditions completed the cognitive test in the same environment [i.e., indoor lab setting]). Akin to the present study, however, the apparent presence of additional construct-irrelevant cognitive demands did not translate into observable cognitive test score differences between the two conditions. To the extent that there are, indeed, UIT device-type effects on cognitive-test scores, further explorations of the particular mechanism(s) of action are warranted, including, but not limited to, explanatory frameworks such as the SCIP model. In accordance with prior research (e.g., Arthur et al., 2014; Arthur et al., 2018; Dages & Jones, 2015), the findings also suggest that using smartphones to complete [non-speeded] assessments translates into longer completion times and a greater degree of distractedness compared to those using desktops. Consequently, it is recommended that organizations and testing professionals inform applicants about preferable test-taking practices, especially under circumstances that allow applicants to choose their test-taking device. The "mobile testing" (Arthur & Traylor, in press) literature is still in its infancy and ripe for further empirical contributions, especially regarding the presence (or lack thereof) of UIT device-type effects on cognitive-test scores.

REFERENCES

Arthur, W., Jr. (1991). *Individual differences in the prediction and training of complex perceptual-motor skill tasks: The development and validation of the computer-administered visual attention test* (Report No. 8). College Station, TX: Texas A&M University, Psychology Department.

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

Arthur, W., Jr. (2015). *Development and validation of a situational judgment test-based measure of conscientiousness and agreeableness.* College Station, TX: Author.

Arthur, W., Jr., & Day, E. A. (2009). Information processing, personality, and demographic variables as predictors of crashes and moving violations: Some implications for prevention. In H. Bédard, & G. Delashmit (Eds.), *Accidents: Causes, analysis, and prevention* (pp.125-148). NY: Novinka/Nova Science Publishers.

Arthur, W., Jr., & Doverspike, D. (1992). Locus of control and auditory selective attention as predictors of driving accident involvement: A comparative longitudinal investigation. *Journal of Safety Research*, *23*, 73-80.

Arthur, W., Jr., Doverspike, D., & Bell, S. T. (2004). Information processing tests. In M. Hersen, & J. C. Thomas (Eds.), *Comprehensive handbook of psychological assessment: Volume 4, Industrial and organizational assessment* (pp. 56-74). New York: John Wiley & Sons, Inc.

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*, 113-123.

Arthur, W., Jr., & Glaze, R. M. (2011). Cheating and response distortion on remotely delivered assessments. In N. T. Tippins, & S. Adler (Eds.), *Technology-enhanced assessment of talent* (pp. 99-152). San Francisco, CA: Jossey-Bass.

Arthur, W., Jr., Keiser, N., & Doverspike, D. (2018). An information processing-based conceptual framework of the effects of unproctored internet-based testing devices on scores on employment-related assessment and tests. *Human Performance*, *31*, 1-32.

Arthur, W., Jr., Keiser, N., Hagen, E., & Traylor, Z. (2018). Unproctored Internet-based device-type effects on test scores: The role of working memory. *Intelligence*, *67*, 67-75.

Arthur, W., Jr., Strong, M. H., Jordan, J. A., Williamson, J. E., Shebilske, W. L., & Regian, J. W. (1995). Visual attention: Individual differences in training and predicting complex task performance. *Acta Psychologica*, *88*, 3-23.

Arthur, W., Jr., Strong, M. H., & Williamson, J. (1994). Validation of a visual attention test as a predictor of driving accident involvement. *Journal of Occupational and Organizational Psychology*, *67*, 173-182.

Arthur, W., Jr., & Traylor, Z. (in press). Mobile assessment in personnel testing: Theoretical and practical implications. In R. N. Landers (Ed.), *The Cambridge handbook of technology and employee behavior*. New York: Cambridge University Press.

Arthur, W., Jr., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology*, *93*, 435-442.

Bobko, P., & Roth, P. L. (2013). Reviewing, categorizing, and analyzing the literature on Black-White mean differences for predictors of job performance: Verifying some perceptions and updating/correcting others. *Personnel Psychology*, *66*, 91-126.

Brooke, J. (1996). SUS - A quick and dirty usability scale. *Usability Evaluation in Industry*, *189*, 4-7.

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*, 61-71.

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, Philadelphia, PA.

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

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*, 1149-1160.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (pp. 7-28). Tilburg, The Netherlands: Tilburg University Press.

Gray, C., Morelli, N. A., & McLane, W. (2015). Does use context affect selection assessments via mobile devices? In N. A. 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.

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*, 382-394.

Lavie, N., Hirst, A., de Fockert, J. W., & Viding, E. (2004). Load theory of selective attention and cognitive control. *Journal of Experimental Psychology: General*, *133*, 339-354.

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*, 15-24.

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.

Morelli, N. A., Potosky, P., Arthur, W., Jr., & Tippins, N. (2017). A call for conceptual models of technology in I-O psychology: An example for technology-based talent assessment. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *10*, 634-653.

Naber, A. M., Arthur, W., Jr., Edwards, B. D., & Franco-Watkins, A. (2018). *Increased retest scores on cognitive tests: Memory or learning effects.* Manuscript submitted for publication.

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.

Pew Research Center. (2015a). Device ownership. Retrieved from http://www.pewresearch.org/data-trend/media-and-technology/device-ownership/

Pew Research Center. (2015b). Technology device ownership: 2015. Retrieved from http://www.pewinternet.org/2015/10/29/technology-device-ownership-2015/

Pew Research Center. (2018). Mobile fact sheet. Retrieved from http://www.pewinternet.org/fact-sheet/mobile/

Plude, D. J., Enns, J. T., & Brodeur, D. (1994). *The development of selective attention: A life-span overview. Acta Psychologica*, *86*, 227-272.

Sackett, P. R., & Walmsley, P. T. (2014). Which personality attributes are most important in the workplace? *Perspectives on Psychological Science*, *9*, 538-551.

Schmitt, N. (2014). Personality and cognitive ability as predictors of effective performance at work. *Annual Review of Organizational Psychology and Organizational Behavior, 1*, 45-65.

Strong, M. H. (1992). *An assessment of the criterion-related validity and improvements in the utility of the computer administrated visual selective attention test* (Unpublished Master's thesis). Texas A&M University, College Station, TX.

Tippins, N. T., & Alder, S. (Eds.). (2011). *Technology-enhanced assessment of talent*. San Francisco, CA: Jossey-Bass.

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*, 189-225.

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, Atlanta, GA.

APPENDIX A

Agreeableness and Conscientiousness

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.

System Usability Scale

DIRECTIONS
Using the assessments you just took as your frame of reference, please use the scale provided below to rate the extent to which you agree with each statement.

①	②	③	④	⑤
Strongly disagree				Strongly agree

1.	I think that I would like to use this [desktop computer/smartphone] frequently	① ② ③ ④ ⑤
2.	I found the [desktop computer/smartphone] unnecessarily complex	① ② ③ ④ ⑤
3.	I thought the [desktop computer/smartphone] was easy to use	① ② ③ ④ ⑤
4.	I think that I would need the support of a technical person to be able to use this [desktop computer/smartphone]	① ② ③ ④ ⑤
5.	I found the various functions of this [desktop computer/smartphone] were well integrated	① ② ③ ④ ⑤
6.	I thought there was too much inconsistency in this [desktop computer/smartphone]	① ② ③ ④ ⑤
7.	I would imagine that most people would learn to use this [desktop computer/smartphone] very quickly	① ② ③ ④ ⑤
8.	I found the [desktop computer/smartphone] very cumbersome to use	① ② ③ ④ ⑤
9.	I felt very confident using the [desktop computer/smartphone]	① ② ③ ④ ⑤
10.	I needed to learn a lot of things before I could get going with this [desktop computer/smartphone]	① ② ③ ④ ⑤

Note. Items 2, 4, 6, 8 and 10 are reverse-coded. "[desktop computer/smartphone]" corresponds to the device used to complete the GMA test and personality assessment (i.e., those in the smartphone condition saw "smartphone," and those in the desktop condition saw "desktop computer").

Attitudes Toward Mobile Testing

1. How would you compare completing employment-related tests on a smartphone versus a desktop computer?
   - o Better on a smartphone than on a computer
   - o As good on a smartphone as on a computer
   - o Worse on a smartphone than on a computer

DIRECTIONS
Using the assessments you just took as your frame of reference, please use the scale provided below to rate the extent to which you agree with each statement.

①	②	③	④	⑤
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

2.	I would prefer to complete employment-related tests on a smartphone versus completing them on a desktop computer.	① ② ③ ④ ⑤
3.	I would be more likely to apply for a job at a company that allowed me to complete its employment-related tests on my smartphone versus a company that allowed taking the same test only on a desktop computer.	① ② ③ ④ ⑤
4.	It is equally fair to use an employment-related test given on a smartphone as it is to use the same one given on a desktop computer to make a hiring or promotion decision for a job.	① ② ③ ④ ⑤
5.	I believe a company that allows me to take its employment-related tests on my smartphone would be a better place to work compared to a company that only allows its employment-related tests to be taken on a desktop computer.	① ② ③ ④ ⑤
6.	I would be more likely to accept a job offer from a company that allows me to take its employment-related tests on my smartphone versus a company that only allows its employment-related tests to be taken on a desktop computer.	① ② ③ ④ ⑤
7.	Having the option to complete employment-related tests on a smartphone positively represents a company's brand image.	① ② ③ ④ ⑤

Note. The original measure used "mobile device" instead of "smartphone," and "computer" instead of "desktop computer."

Distractions While Testing

1. Did you experience any distractions while taking the test that you just completed?
   - No
   - Yes

2. To what extent did you find the environment distracting while taking the test that you just completed?

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

3. The testing environment made it difficult for me to concentrate on taking the test that I just completed.

①	②	③	④	⑤
Strongly disagree		Neutral		Strongly agree

Note. These items were drafted for the purposes of the present study and administered directly following the GMA test (which was completed inside [desktop condition] or outside [smartphone condition]).