An Information-Processing-Based Conceptual Framework of the Effects of Unproctored Internet-Based Testing Devices on Scores on Employment-Related Assessments and Tests

Winfred Arthur Jr., Nathanael L. Keiser & Dennis Doverspike

To cite this article: Winfred Arthur Jr., Nathanael L. Keiser & Dennis Doverspike (2017): An Information-Processing-Based Conceptual Framework of the Effects of Unproctored Internet-Based Testing Devices on Scores on Employment-Related Assessments and Tests, Human Performance, DOI: 10.1080/08959285.2017.1403441

To link to this article: https://doi.org/10.1080/08959285.2017.1403441

Published online: 15 Dec 2017.
An Information-Processing-Based Conceptual Framework of the Effects of Unproctored Internet-Based Testing Devices on Scores on Employment-Related Assessments and Tests

Winfred Arthur Jr., Nathanael L. Keiser, and Dennis Doverspike

Texas A&M University; University of Akron

ABSTRACT
The widespread use of unproctored Internet-based testing (UIT) in employment-related assessments has resulted in a burgeoning use of mobile devices to complete these assessments. Coupled with this is a concomitant interest in UIT-device-type effects, specifically, differences or lack thereof between assessments completed on “mobile” versus “nonmobile” devices. However, there is limited theoretical or conceptual work that seeks to explain the observed effects of UIT device type on test scores. Consequently, this article first presents a review of the extant empirical literature and then, on the basis of that, presents a framework—the structural characteristics/information processing framework—for psychologically conceptualizing the effect of UIT device types on test and assessment scores. The framework is used to explain previous findings and advance testable propositions for future research.

Advances in Internet technology have permitted the remote administration of personnel selection tests and assessments, increasing more via unproctored Internet-based testing (UIT; Tippins & Alder, 2011). Because UITs, by their very nature, allow test takers to decide when, where, and the device on which to complete their tests, it is not surprising that there has been a burgeoning increase in the use of mobile devices to complete these assessments (Arthur, Doverspike, Muñoz, Taylor, & Carr, 2014; Illingworth, Morelli, Scott, & Boyd, 2015; McClure Johnson & Boyce, 2015). Indeed, mobile assessments was the number one workplace trend in the Society for Industrial and Organizational Psychology’s (SIOP; 2015) top-10 workplace trends for 2015. This increasing use of mobile devices to complete employment-related selection tests and assessments can be attributed to two factors. First, as previously noted, mobile devices augment and magnify one of the primary advantages of UIT, that is, the ability of job applicants to test anywhere, and at any time, and the choice in deciding the device to use to complete the assessment. Thus, mobile devices untether test-takers from the wall in terms of Internet access, giving them more degrees of freedom in terms of where they can complete employment-related tests and assessments. Second, there is a continued increase in mobile device ownership in the general population as reflected in Pew Research Center’s annual surveys (http://www.pewinternet.org/2015/10/29/technology-device-ownership-2015/).

In the present work, we define a UIT device as any device that a test-taker can use to complete an unproctored Internet test or assessment where, by definition, the test-taker also decides when and where to complete the assessment or test. Thus, a “UIT device” is not synonymous with a smartphone or other mobile devices. A smartphone or other mobile devices are just one example of a UIT device; so is a desktop computer. That being said, in spite of the prevalent increase in the use of mobile devices to complete unproctored Internet tests and assessments, the
psychological and assessment research literature on the effects of UIT device types on test scores is very limited. Thus, for instance, whereas there are observed mean differences on cognitive but not noncognitive constructs as a function of UIT device type (see results of the literature review in the next section of this article), there are no theoretical or conceptual models in the literature that provide guidance to inform the psychological and conceptual classification of device types (i.e., What is a “mobile device”?)). Furthermore, there is very limited theoretical or conceptual work that seeks to explain the effects of UIT device type on test scores, that is, why and when their use should or should not affect test scores (cf. Potosky, 2008, which is discussed later in this article). Finally, detailed scholarly reviews of this literature are absent; hence both researchers and practitioners currently lack ready access to summaries of the pertinent literature to inform their scholarly and applied research and practice. To address these gaps, the present article first presents a detailed review of the empirical UIT device type employment-related assessment literature. This review is organized in terms of (a) the sources of the papers reviewed (e.g., published article, conference presentation), (b) the settings of the studies (operational vs. lab data), (c) the sample size and proportion of specified device-type users, (d) the constructs assessed (cognitive, noncognitive), (e) the test type/method (e.g., multiple-choice, Likert scales, situational judgement tests [SJTs]), and (f) whether the assessments were timed versus untimed. The literature is also reviewed in terms of the following outcomes: (g) the measurement equivalence of device-type scores, (h) score differences between UIT device types (taking the constructs assessed into account), (i) criterion-related and differential validity for device-type scores, and (j) test-taker reactions and preferences.

Building on the results of the literature review in terms of whether there are, and whether there are not, measurement outcome-related device-type effects, we next present a framework—the structural characteristics/information processing (SCIP) framework—for psychologically conceptualizing the effect of UIT device type on assessment scores. Thus, our interest and focus is on the psychological aspects of UIT devices that influence test scores and less so on their technological sophistication per se. Consequently, rather than simply defining UIT device types as being “mobile” (untethered from the wall) versus “nonmobile” (being tethered to the wall), we present two dimensions—the structural characteristics of UIT devices and the associated construct-irrelevant information processing variables that they engender—and then use these dimensions to develop a conceptual framework, the SCIP framework, for why different UIT devices should or should not have different effects on testing and assessment outcomes. Hence, the SCIP framework explains and accounts for the measurement outcome-related findings in the literature and, on the basis of this, subsequently permits the development of testable propositions pertaining to when one might and might not obtain UIT-device-type effects on scores on employment-related assessments and tests. We conclude the article with a brief comparison of the SCIP framework with Potosky’s (2008) conceptual framework. For instance, because it focuses on the role of administration medium on the personnel assessment process, Potosky’s (2008) framework also speaks to the effect of technology on assessment scores. However, because it is embedded in communication theory, it views assessment medium effects not in terms of individual differences in specified abilities but instead as communicative acts between the test-taker and the individual or anyone who wants to measure attributes of the test-taker. Consequently, the focus is on how the communication channel or the medium’s structural attributes affect the message quality and hence test scores. In contrast, the SCIP framework conceptualizes assessment device-type effects in terms of how individual differences of the test-taker on specified information processing attributes interact with the structural characteristics of the assessment device to generate construct-irrelevant cognitive load on the test-taker that subsequently influences the individual’s performance on the assessment. That is, because there is a finite pool of cognitive information-processing resources (Cowan, 1988; Norman, 1976) available to take the assessment, these sources of construct-irrelevant cognitive demands differentially draw on these resources (as a function of the construct assessed) resulting in observed device-type effects or lack thereof.
Review and summary of the empirical literature

It should be noted that in the review that follows, we defer to the authors’ designation of their UIT devices as being either “mobile” or “nonmobile”; that is, in the review of specified studies, we use the authors’ own terminology to describe the device types. We also acknowledge that in most instances it was impossible to determine exactly what the authors operationally defined as a mobile versus nonmobile device. Hence, it is our supposition that most of them used the tethered versus untethered-from-the-wall definition. As we later demonstrate, one advantage of the SCIP framework is that it provides guidance to future research on how to classify and describe UIT device types in comparative studies in a manner that is psychologically and conceptually meaningful.

Literature search and inclusion criteria

The terminal date for the search window for the literature review was December 2016 (inclusive), and it commenced with a search of the following electronic databases, PsycINFO, PsycARTICLES, ERIC, Web of Science, EBSCO Academic Search Complete, Science and Technology Collection, Applied Science and Technology Source, Business Source Complete, Psychology & Behavioral Sciences Collection, Science and Technology Collection, Vocational and Career Collection, and Vocational Studies Complete using the following search terms, mobile device, mobile, tablet, smartphone, and employment testing, employee selection, employee testing, selection, assessment, selection and assessment, and mobile web survey design. The choice of search terms was guided by the focus on comparative examinations of the effects of UIT device types in personnel testing and assessment. The conference programs of SIOP, Society of Human Resource Management, International Personnel Assessment Council, American Psychological Association, Association of Psychological Science, and Academy of Management were also searched. These efforts resulted in 2,917 hits. To be retained in the detailed review, a work had to (a) be an empirical study with a focus on testing in personnel assessment and/or selection; (b) be a UIT study, that is, the testing/assessment was unproctored; and (c) report a comparative examination of UIT device types (e.g., compare a mobile device [as stipulated by the authors] to a nonmobile device or compare different mobile devices [e.g., smartphone vs. tablet]). A review of each source in the context of the preceding inclusion criteria resulted in a retention of only 23 papers.

We next present a summary of these papers, which is organized in terms of (a) the sources of the papers reviewed (e.g., published article, conference presentation), (b) the settings of the studies (operational vs. lab data), (c) the sample size and proportion of specified device-type users, (d) the constructs assessed (cognitive, noncognitive), and (e) the test type/method (e.g., multiple-choice, Likert scales, SJT), and (f) whether the assessments were timed versus untimed. These studies are also reviewed in terms of the following outcomes: (g) the measurement equivalence of device-type scores, (h) score differences between UIT device types (taking into account the constructs assessed), (i) criterion-related and differential validity for device-type scores, and (j) test-taker reactions and preferences. A detailed summary of each of the 23 studies is presented in Appendix A. It should be noted that four studies (i.e., Gutierrez & Meyer, 2013; Huff, 2015; King, Ryan, Kantrowitz, Grelle, & Dainis, 2015; Smeltzer, 2013) examined some of the issues of interest here but were excluded because, although they examined device-type effects, the testing was proctored and thus did not meet the requirement that the testing or assessment be unproctored; these were all lab studies. However, for the sake of completeness, these studies are summarized in Appendix B.

Review of the empirical research

Source

Of the 23 studies, only three were peer-reviewed articles. The remaining 20 were conference presentations, all but one of which were presented at a SIOP conference. It is also noteworthy that
most of this work is fairly recent, with none older than 2012. Thus, in summary, the volume of empirical research on UIT device types and testing is quite limited, and very recent, with the preponderance of it being conference presentations. The latter is noteworthy because, whereas we do not provide a detailed review of the quality of each of these studies, to the extent that conference presentations do not undergo the same rigorous peer-review process as journal articles, then it would not be unreasonable to question the relative quality of the extant mobile device literature. (For instance, as previously noted in most instances, beyond the labels used by the authors, it was impossible to determine the authors’ operational definition of what was and was not considered to be a mobile device.) Consequently, as a cautionary note, the summaries of the extant findings that follow need to be considered in the context of this characterization. On the other hand, this state of affairs is reflective of the fact that this is a domain where practice has outpaced theory and research in the academy and peer-reviewed scholarly literature (Arthur, Doverspike, Kinney, & O’Connell, 2017; O’Connell, Arthur, & Doverspike, 2015), and the present work can be considered to be one attempt to play “catch up.”

**Operational versus lab data**

Sixteen of the studies used operational data, that is, data collected primarily for operational decision making (e.g., selection) purposes. The remaining seven were lab studies, six of which used Amazon Mechanical Turk (MTurk) samples and the seventh used potential job candidates who completed practice employment tests.

**Sample size and proportion of specified device-type users**

The sample sizes for the operational studies were generally very large, with the prototypical sample size being in the hundreds of thousands and sometimes millions (e.g., Arthur et al., 2014; Golubovich & Boyce, 2013; McClure Johnson & Boyce, 2015; Rupayana & Hedricks, 2013); the smallest studies were in the 6,000–10,000 range. In contrast, the lab studies are characterized by relatively small sample sizes, which is typical for these sorts of studies. Specifically, they ranged in sample sizes from 143 (Chang, Lawrence, O’Connell, & Kinney, 2016a) to 2,787 (Grossenbacher, Brown, & Nguyen, 2016) with a mean of 1,099 (\(Mdn = 692, SD = 925\)).

Concerning the proportion of mobile users, in relative terms, the percentages are fairly small. For the operational studies (where by definition participants self-selected their assessment device type), the mean was 11.2% (\(Mdn = 7.8\%, SD = 10.9\%\)) for the studies that reported this information. That being said, there is clearly a drastic, continued, seemingly exponential increase in the number of test-takers using mobile devices to complete assessments. For instance, McClure Johnson and Boyce (2015) presented data (spanning 2009–2014) which indicated levels of 2.1% in 2009, 9% in 2012, 21.5% in 2013, and 31% in 2014. Similar trends and patterns have been reported by other researchers as well (e.g., O’Connell et al., 2015). Concerning who is using these devices for employment-related assessments, the research to date suggests that African Americans, Hispanics, and females are more likely to take a test on a mobile device than are White males (Arthur et al., 2014; Illingworth et al., 2015; McClure Johnson & Boyce, 2015; Rossini, 2016). The data for age are limited and, when reported, are mixed. In summary, although a small percentage, because of the total number of test-takers, this translates into a meaningful number of test-takers who are using mobile devices to complete their high-stakes assessments, a pattern that is consistent with the national small screen device ownership data (Pew Research Center, 2015). It is also expected that the percentage of users will continue to increase.

**Constructs assessed**

Eight studies assessed both cognitive and noncognitive constructs, three assessed cognitive constructs only, eight noncognitive constructs only, three did not specify the constructs assessed, and one focused on the use of mobile devices to submit references. The cognitive constructs were prototypically general mental ability (GMA), but also on occasion included constructs like “learning”
(Morelli, Mahan, & Illingworth, 2014), and mechanical aptitude (Wood, Stephens, & Slither, 2015). There was much wider variability in the range of noncognitive constructs assessed. So, in addition to a fair representation of the five-factor model (FFM) dimensions, there were also constructs such as “customer service orientation” and “neatness” (Morelli et al., 2014), and “stress tolerance” and “service and sales potential” (Lawrence, Wasko, Delgado, Kinney, & Wolf, 2013).

**Test type/method**

There was a fairly wide range of test types and methods used across the 23 studies, with the most common method being multiple-choice item formats and Likert-type scales for the GMA and personality tests, respectively. Five studies reported using a simulation (e.g., high fidelity interactive simulation; Chang et al., 2016a) and seven an SJT item format. As many as eight studies used biodata, and three failed to report or did not report sufficient information to permit the determination of the test type/method.

**Timed versus untimed**

The majority of the studies did not provide any information pertaining to whether the assessments were timed. For those that did, noncognitive assessments were not timed whereas, as would be expected, cognitive assessments were.

**Measurement equivalence of UIT-device-type scores**

As with previous concerns about the equivalence of computer-administered and paper-and-pencil test scores (Mead & Drasgow, 1993; Ployhart, Weekley, Holtz, & Kemp, 2003), and later proctored versus unproctored assessment test scores (Arthur, Glaze, Villado, & Taylor, 2010; Davies & Wadlington, 2006; Delgado, Kung, & O’Connell, 2009; Do, Shepherd, & Drasgow, 2005), nine of the 23 studies examined the equivalence of mobile and nonmobile device test scores. These studies were typically very large-sample operational studies; the robust conclusion arising from this research is that the psychometric properties, such as factor structure, the reliability of scores, and differential item functioning (DIF), among others, for mobile and nonmobile device scores are quite similar. That is, for both cognitive and noncognitive constructs, UIT device types (i.e., mobile and non-mobile devices) display measurement equivalence.

**Score differences between UIT device-types**

Of the 23 studies, 13 examined mean differences between mobile and nonmobile devices on noncognitive constructs and eight for differences on cognitive constructs. Three examined device-type differences but did not specify the constructs assessed. Of these three, two (Chang et al., 2016a; Lawrence, Chang, O’Connell, & Kinney, 2016) examined device-type differences when the assessment was a “high-fidelity simulation,” and the third (Rossinni, 2016) failed to distinguish constructs from methods. All these three studies reported lower scores for mobile devices compared to nonmobile devices. In terms of the results for the 13 studies that assessed noncognitive constructs, only two—LaPort (2016) and McClure Johnson and Boyce (2015)—obtained device-type differences. However, LaPort (2016) described her differences as “negligible” ($d = 0.13$). Likewise, McClure Johnson and Boyce (2015) reported similar small but lower mobile device scores for entry-level jobs but less so for the manager assessment.

For studies that examined cognitive constructs, five obtained lower scores for mobile compared to nonmobile devices, and three did not obtain any device-type differences. Of the three that did not obtain any differences, two (i.e., Brown, Grossenbacher, & Nguyen, 2016; Parker & Meade, 2015) were lab studies that used MTurk nonapplicant samples in which participants were assigned to devices instead of via self-selection, and the testing was low stakes. In the third (Morelli et al., 2014), the cognitive construct was described as “learning.” Six studies also examined differences in completion times (on untimed noncognitive constructs), and all reported longer completion times for mobile device assessments. Finally, Parker and Meade (2015) reported a higher attrition rate for
mobile devices, that is, a higher failure to complete the assessments when test-takers used a mobile
device compared to a nonmobile device.

So, in summary, a general finding characterizes this literature: Whereas there are no mean
differences on noncognitive (e.g., personality) assessments taken on mobile and nonmobile devices
(and when present, they are very small), under high-stakes conditions where test-takers select their
assessment device, there are pronounced differences for cognitive constructs with scores on mobile
devices being substantially lower. For instance, Arthur et al. (2014) reported a $d$ of 0.90. Impelman
(2013) reported similar performance differences on cognitive measures across four organizational
samples. Wood et al. (2015) reported $d$s of 0.46 and 0.35 for two cognitive ability tests and 0.93
and 0.26 for two mechanical aptitude tests. Likewise, LaPort (2016) reported $d$s of 0.46 and 0.60 for tests
requiring minimal scrolling and higher scrolling, respectively.

**Criterion-related and differential validity for device-type scores**

Arthur and Villado (2008) identified and discussed three facets on which predictors in the personnel
selection literature could be compared: criterion-related validity, subgroup differences, and test-taker
reactions. Whereas the preceding and subsequent sections indicate that UIT device types have been
compared on the second and third of these dimensions, none of the 23 studies identified and
reviewed here reported any information on or examined the criterion-related validity of the specified
test scores. Lawrence, Chang, O’Connell, and Kinney (2016) reported criterion-related validities for
only their mobile device (tablet and smartphone) scores ($r = .11$–.22; avg. = .17). However, the
constructs assessed were not specified; instead, the focus was on high-fidelity interactive simulations.
In summary, the absence of any examinations of the comparative criterion-related validity of UIT
device types (with a focus on the constructs assessed as well) is a major gap in the literature.

**Test-taker reactions and preferences**

Seven of the 23 studies examined test-taker reactions and preferences. Six of these used a
between-subjects design, and all but two of these six were operational studies. Of the two lab
studies, one used potential job applicants and the other an MTurk sample. The one within-
subjects design study was a lab study that also used an MTurk sample. By their very nature,
within-subjects designs represent more robust comparisons because each test-taker completes the
assessment on both UIT device types and thus can make an intrasubject comparison of their
reactions to and preference for the devices. Consonant with this, less favorable reactions were
observed for the one within-subjects design (Chang et al., 2016a) with lower levels of satisfaction
for screen sizes less than 4 in. Three of the between-subjects designs (Fursman & Tuzinski, 2015;
Gutierrez, Meyer, & Fursman, 2015; Gutierrez & Sanderson, 2015) also found substantially lower
preferences and less favorable reactions to mobile devices; however, three (Fursman, 2016;
Kinney, Lawrence, & Chang, 2014; Rossini, 2016) found none or very limited differences in
test-takers’ preferences for mobile versus nonmobile devices as assessment platforms. Of interest,
even for these three between-subjects studies that did not obtain any device-type differences,
there were no instances in which the mobile device reactions and preferences were higher
than those for nonmobile devices. It is also noteworthy that like Chang et al. (2016a), Gutierrez et al.
(2015) obtained a positive relationship between test-takers’ satisfaction with the testing experi-
ence and the screen size of the UIT device on which they took the assessment.

Finally, Arthur et al.’s (2014) and Dages and Jones’s (2015) results indicated that there were large
numbers of test-takers who started their assessments on a mobile device and then switched to a
nonmobile device. For instance, Dages and Jones reported that for applicants who used multiple
devices, 89% for their 156-item measure and 74% for their 50-item measure started on a smartphone
and switched to a computer. In summary, whereas test-takers generally indicated that applicants
should be given the opportunity to complete assessments on mobile devices, they also generally did
not have more favorable reactions toward them, and in more instances than not had less positive
reactions.
Summary and conclusions

As previously noted, a noteworthy observation is that the preponderance of the extant literature comprises conference presentations, and to the extent that these are qualitatively different from peer-reviewed journal articles, then some caution is warranted in interpreting and drawing firm conclusions from these studies. So, with that as a backdrop, a number of measurement outcome-related conclusions can be drawn from the preceding review of the empirical literature. First, for both cognitive and noncognitive constructs, UIT device types (i.e., mobile and nonmobile devices) display measurement equivalence in terms of factor structure, DIF, and score reliability. Second, there is an absence of any comparative device-type criterion-related validity studies. Third, there are substantial UIT-device-type score differences (with lower scores for mobile devices) on cognitive constructs but no (or very limited) differences on noncognitive constructs. Fourth, assessment and test completion times are longer for mobile devices. Fifth, job applicants generally do not have more favorable reactions toward completing high-stakes assessments and tests on mobile devices, and indeed in most instances they report less positive reactions.

The third, fourth, and fifth conclusions warrant some explanation in that from a psychological perspective, it is reasonable to ask why UIT device types should or should not affect assessment and test scores, and the conditions under which they should and should not do so. That is, what psychological phenomena or variables are at play to cause or explain the observed UIT-device-type differences or lack thereof? We present the SCIP framework to explain the pattern of findings observed in the empirical literature. Specifically, to the extent that the four information-processing variables (i.e., working memory, perceptual speed and visual acuity, psychomotor ability, and selective attention) that correspond to four structural characteristics of UIT assessment devices (i.e., screen size, screen clutter, response interface, and permissibility) identified by the SCIP framework play a role in using the UIT device, they then result in additional construct-irrelevant cognitive load or demands that are likely to influence performance on the test or assessment when said cognitive demands are not the focal construct of interest. We next present the SCIP framework along with some testable propositions that follow from it.

The psychological basis of UIT-device-type effects on test and assessment scores: The SCIP framework

There are obviously a wide range of UIT device types, currently ranging from desktop computers, laptops, and notebooks to handheld small-screen devices like tablets, personal digital assistants, smartphones, and conceivably a host of other devices in the not-too-distant future. These various UIT devices serve as the basis for the mobile versus nonmobile distinction that characterizes the literature. However, in spite of the wide variability in the structural and design characteristics of these devices (e.g., screen size, screen clutter, response interface), the literature typically differentiates device types simply in terms of whether the device is tethered to the wall. Whereas this may be technologically accurate, “tetheredness-to-the-wall” fails to provide any conceptual basis or psychologically meaningful explanations for why the use of specified UIT devices should or should not have an effect on test scores. Consequently, building on two dimensions—the structural characteristics of UIT devices and the associated information processing variables that they engender—we present a framework that conceptualizes device types in terms of the construct-irrelevant information processing demands placed on the test-taker while taking the assessment. Specifically, differences in specified structural characteristics engender specified associated information-processing demands, resulting in additional construct-irrelevant cognitive load that interacts with the device type, resulting in differential outcomes as a function of the construct (cognitive vs. noncognitive) assessed. We refer to this conceptual framework as the SCIP framework.

It should be emphasized that the SCIP framework is not device specific; that is, it is not tied or yoked to any specific UIT device (e.g., smartphone) or set of devices (e.g., mobile devices). Instead,
as long as an assessment device engenders the construct-irrelevant information-processing cognitive load posited by the SCIP framework, it then subsequently allows one to appraise and make inferences as to whether one is likely to have device-type effects. So, although the sections that follow use the current prototypical UIT devices such as smartphones, notebooks, and desktop computers as examples, they are just that, examples.

**UIT-device-type structural characteristics**

As illustrated in Figure 1, we identify four structural characteristics that can be used to describe current prototypical UIT device types such as desktop computers, laptops, notebooks, tablets, phablets, and smartphones as exemplars. These characteristics are (a) screen size, (b) screen clutter, (c) response interface, and (d) permissibility. Before proceeding to describe the structural characteristics, it should be noted that although the examples used as illustrations are suggestive of devices that are “high” or “low” on all the characteristics, this is of course not necessarily the case; any configuration of structural characteristics is possible; it is a function of the specific device configuration (e.g., a desktop with a touch screen interface).

**Screen size** simply refers to the size of the viewable surface on which information is presented. Thus, at one end of this continuum are desktop computers, which can have big screens at 23 in. or larger (frequently with dual monitors). In contrast, at the other end of the continuum are smartphones, which can have screen sizes down to 3.5 in. or smaller. Depending on the specific make and model, laptops and tablets are characterized by intermediate screen sizes in the order of 5 to 12 in. Screen size is obviously associated with the level of **screen clutter**, which we conceptualize in terms of the “place[ment] of a large amount of information within a limited amount of display ‘real estate’” (Kroft & Wickens, 2002, p. 44). Screen clutter is related to the extent to which graphics and text are readily readable (Horrey & Wickens, 2004), with larger screens (displays) presenting information, specifically text and symbols, at a greater size, making them easier to read. In contrast, smaller screens represent higher levels of clutter in terms of the number of objects within the display, resulting in a subsequent increase in the density of information around a given point (Kroft & Wickens, 2002).

**Response interface** pertains to the means by which the user interacts with the UIT device. Thus, these range from a full-size physical keyboard and/or mouse (that typically characterize desktop computers), to smaller physical or virtual keyboards (that characterize laptops and tablets), to touch screens, and finger swipes ([and sometimes a stylus] that characterize phablets and smartphones).  

---

**UIT Device-Type SCIP Framework**

---

**Figure 1.** Illustration of unproctored Internet-based testing (UIT) device-type structural characteristics and associated information-processing (IP) demands. SCIP = structural characteristics/information processing.
Indeed, it even encompasses game console controllers as well. The fourth structural characteristic, *permissibility*, pertains to the degrees of freedom that one has in terms of the choice of locations and the conditions under which the UIT device can be used, and therefore the amount of potential distractions present. For example, Arthur et al. (2014) speculated that one plausible explanation for the low GMA scores observed for mobile devices might be because individuals using these UIT device types work in more distracting environments, with their scores reflecting lower levels of concentration and on-task focus. Consonant with this, Chang, Lawrence, O’Connell, and Kinney (2016b) reported that job applicants who completed their assessments on a mobile device self-reported higher levels of distractions in their testing environment compared to those who used a nonmobile device. Along similar lines, Gray, Morelli, and McLane (2015) identified and examined three contexts (privacy [public vs. private], movement [static vs. moving], and location [indoor vs. outdoor vs. on a transport]) in which UIT assessments were completed, further highlighting the high degree of permissibility that they provide in the completion of high-stakes assessments. Thus, generally speaking, at one extreme, nonmobile devices (e.g., desktops) are characterized by the lowest levels of permissibility with mobile devices (e.g., smartphones) engendering the highest levels of permissibility.

**General information-processing theory**

Information-processing theories view the human mind as an information processor that codes, stores, and retrieves inputs that are used to execute responses in response to a stimulus (Ackerman, 1988; Neisser, 1967; Newell & Simon, 1972; Sternberg, 1977). Although there are a number of information-processing models (e.g., Card, Moran, & Newell, 1980; Newell & Simon, 1972), they all include common structural elements, which are summarized in Arthur, Doverspike, and Bell (2004) as (a) a short-term sensory store, (b) a perceptual mechanism, (c) a mechanism for decision making or response selection, (d) memory (short term, working, and long term), (e) mechanisms for response execution, and (f) an attentional pool. In addition to the common structural components, information-processing models also share certain basic assumptions, and three that are most critical to the tenets of the SCIP framework are nonimmediacy, limited capacity, and individual differences.

Nonimmediacy pertains to the fact that there is always a time interval between the presentation of a stimulus and the generation of a response. This *time latency* corresponds to time spent in mental processing and corresponds to the amount of time required to accomplish various mental tasks or activities (Posner & Mitchell, 1967). In reference to *limited capacity*, individuals are viewed as having limited attentional and structural capabilities or capacities (Broadbent, 1957; Fisher, 1982; Navon & Gopher, 1979). This limited capacity is a result of the necessity to share attentional resources and is the cause of various bottlenecks in the information-processing system. Third, it is recognized that there are differences among individuals in information-processing capacity, speed, and accuracy and that these *individual differences* can be effectively measured to differentiate between individuals (Ackerman, 1987, 1988). These individual differences (i.e., information-processing abilities/variables) are conceptualized as specific (narrow) abilities in structural models of cognitive ability. Thus, for example, in Carroll’s (1993) three-stratum theory of cognitive abilities, which is considered to be the most widely accepted structural model in terms of general intelligence, g (Drasgow, 2013; Grubb, Whetzel, & McDaniel, 2004), they are the first stratum of abilities. Individual differences in information-processing abilities have been shown to be predictive of performance on a wide range of lab and field tasks and activities such as air traffic control, and driving (e.g., Ackerman, 1987, 1988; Ackerman & Kanfer, 1993; Arthur, Barrett, & Alexander, 1991; Arthur, Barrett, & Doverspike, 1990; Arthur et al., 1995). In summary, information-processing theories and models are based on the premise of a finite pool of cognitive resources that are used to process and respond quickly and accurately to stimuli. Furthermore, to the extent that there is competition for this finite pool of resources (i.e., limited capacity), there will be decrements in performance (both latency and accuracy). Finally, because there are individual differences in information-processing abilities,
there are concomitant differences in the extent to which individuals’ performance on the primary task will be adversely affected by competing cognitive demands or additional cognitive load. Figure 2 presents an illustration of a general information-processing model that has been modified to reflect the conceptual underpinnings of the SCIP framework.

**UIT-device-type construct-irrelevant information-processing demands**

It is our proposition that the four structural characteristics of UIT assessment devices identified here—(a) screen size, (b) screen clutter, (c) response interface, and (d) permissibility—engender the role of four corresponding information-processing variables—(a) working memory, (b) perceptual speed and visual acuity, (c) psychomotor ability, and (d) selective attention. The framework focuses on these four information-processing variables because they are the most aligned with the focal structural characteristics in that they broadly pertain to the ability to process information while holding relevant information in the short term, the ability to detect differences between standard and comparative (other) figures/targets, the ability to search for and detect discrepancy–cause relations, and the ability to respond quickly and accurately. So, as illustrated in Figure 2, to the extent that these information-processing variables play a role in using the UIT device, they then result in additional construct-irrelevant cognitive load that is likely to influence performance on the test when said cognitive variables or associated demands are not the focal construct of interest (e.g., see Arthur et al., 2014).

The general effects of screen size have been examined in the computer–human interaction literature (e.g., Chae & Kim, 2004; De Bruijn, De Mul, & Van Oostendorp, 1992). In terms of the SCIP framework, the role of working memory (Baddeley, 2012; Harrison et al., 2013; Mrazek, Franklin, Phillips, Baird, & Schooler, 2013; Shelton, Elliott, Matthews, Hill, & Gouvier, 2010)
based on the treatise that differences in screen size (small to large) engender differences in working memory demands (high to low) because more screens are typically needed to present the required information, and therefore more information has to be held in working memory in order to complete the assessment (Sanchez & Goolsbee, 2010).

Working memory has been defined as “a brain system that provides temporary storage and manipulation of the information necessary for … complex cognitive tasks” (Baddeley, 1992, p. 556). That is, working memory capacity reflects limits of an individual’s ability to retrieve information from memory that has been lost from the focus of attention due to competing cognitive demands (Baddeley, 2012; Harrison et al., 2013; Mrazek et al., 2013). In the context of testing, working memory capacity facilitates solving cognitive problems by enabling the test-taker to draw comparisons between alternatives and responses and then select a response by simultaneously matching the demands of the item, prior knowledge, and relevant information retrieved from memory (Baddeley, 2012; Harrison et al., 2013; Mrazek et al., 2013). Consequently, for small-screen devices, to the extent that there is limited or only partial information in the test-taker’s visual field (i.e., the screen), working memory will play a greater role in processing the information required to complete the item, as larger amounts of it (e.g., portions of current items—either stem or response options; responses to previous similar items) have to be retrieved from working memory (e.g., see Sanchez & Branaghan, 2011). Furthermore, on assessments optimized for small-screen devices such as smartphones, it is not uncommon to present only one item per screen and so test-takers do not have ready access to their responses to previous items, which may inform their responses to subsequent items (e.g., Couper, Conrad, & Tourangeau, 2007; Rivers, Meade, & Fuller, 2009; Schwarz, 1999). Consequently, consonant with the ideas and propositions of information-processing theory, the SCIP framework posits that when UIT devices at the high end of the device-engendered cognitive load continuum (e.g., smartphones) are used to measure cognitive constructs, there will be a stronger relationship between working memory and the focal cognitive construct scores compared to devices at the low end of the continuum (e.g., desktop computers).

In addition to working memory, differences in screen clutter—the density of information (text and symbols) within a limited amount of display space—which is associated with screen size, will be expected to result in performance differences as well (Horrey & Wickens, 2004; Kroft & Wickens, 2002). It is our proposition that higher levels of screen clutter will increase the role that perceptual speed and visual acuity play in the test-taker’s interaction with the device and, subsequently, their test performance. Visual acuity is the ability to see and distinguish fine details and thus speaks to the clarity or sharpness of vision. It is recognized as being important to the reading of text, and recognizing symbols, among others. As per O*NET (n.d.—a), perceptual speed is the “ability to quickly and accurately compare similarities and differences among sets of letters, numbers, objects, pictures, or patterns.” It entails the ability to compare targets presented at the same time or one after the other.

Differences in device interface engender differences in construct-irrelevant psychomotor ability demands. Psychomotor abilities broadly refer to abilities that “influence the capacity to manipulate and control objects” (O*NET, n.d.—b) and entail a range of UIT-device-relevant abilities such as arm–hand steadiness, control precision, finger dexterity, manual dexterity, rate control, reaction time, and wrist–finger speed. Thus, the use of a full size keyboard and/or mouse versus touch screen and finger swipes places differential demands on psychomotor ability to the extent that test-takers will likely have more difficulty interfacing with a handheld small-screen device requiring touch screen and finger swipes (due to the so-called fat finger problem; e.g., Henze, Rukzio, & Boll, 2011; Siek, Rogers, & Connelly, 2005) versus a desktop computer for which one can use a full-size keyboard and/or mouse (or even a desktop with touch-screen interface because, given the larger screen size, they would be less susceptible to the fat finger problem). Hence, it is interesting to note that using a within-subjects design in a comparison of desktop computers and smartphones to complete a FFM measure, Huff (2015) reported more errors for smartphone assessments. Errors were operationalized as the respondent hitting the Submit button too many times resulting in
multiple submissions, or the respondent failing to properly submit the completed measure resulting in no submission record.

As previously noted, UIT devices also differ in permissibility, that is, the degrees of freedom that the test-taker has in terms of the locations and situations in which the assessment can be taken. Thus, devices vary in terms of the extent to which they readily permit the test-taker to take the assessment in a private versus public space, while moving versus being stationary, and indoors versus outdoors (Gray et al., 2015). In psychological terms, these differences translate into variation in the extent to which the test-taker is subjected to distractions while taking the assessment (Chang et al., 2016b). The additional cognitive load engendered here pertains to selective attention, which is the ability to focus attentional resources on a task in the presence of irrelevant distractions (Arthur et al., 2004). That is, individual differences in selective attention are germane when there are a number of sources competing for user’s attention, and it requires effortful cognitive control (Lavie, Hirst, de Fockert, & Viding, 2004). Furthermore, increases in cognitive demands are associated with increases in distractibility (Lavie & de Fockert, 2005; Lavie et al., 2004) and consequently, the enhanced role of selective attention. So, consonant with this, selective attention has been demonstrated to be related to outcomes such as driving crashes, and performance on a host of complex tasks (Arthur & Doverspike, 1992; Arthur et al., 1995; Arthur, Strong, & Williamson, 1994). Thus, to the extent that the UIT device permits the test-taker to take assessments in environments where the test-taker will be subjected to higher levels of distraction, higher demands will be placed on the test-takers selective attention ability.

What is a mobile versus nonmobile device?

Framed in the context of construct-irrelevant information-processing demands that are engendered by the specific structural characteristics of UIT devices identified here, a question that arises from a psychological perspective is, What is a mobile device, and when should device-type affect test scores? It is our treatise that in terms of psychological research and understanding, conceptualizing UIT device types as being untethered from the wall (mobile) versus tethered to the wall (nonmobile) is neither scientifically informative nor meaningful. Hence, the SCIP framework is presented as a framework for classifying UIT devices on factors that may account for or explain observed device-type effects (or lack thereof) as a result of the specific testing situation or circumstances. Consequently, any UIT device can be appraised in terms of the level of posited construct-irrelevant information demands to predict whether or not there should be device-type effects. Thus, Figure 3 is presented as an illustrative example of how current prototypical UIT devices might be classified in terms of the level of device-engendered construct-irrelevant cognitive load. However, once again, it is emphasized that the SCIP framework is not device specific; that is, it is not tied or yoked to any specific UIT device (e.g., smartphone) or set of devices (e.g., mobile devices). Instead, as long as an

---

**Figure 3.** Illustrative example of how current prototypical unproctored Internet-based testing devices might be classified on construct-irrelevant information-processing cognitive load with hypothesized similar and dissimilar device-type effect outcomes.
assessment device engenders the construct-irrelevant information-processing cognitive load posited by the SCIP framework, it then subsequently allows one to appraise and make inferences as to whether one is likely to have device-type effects. Therefore, the devices presented in Figure 3 are merely used as examples of how the framework permits the psychological conceptualization of this particular set of UIT devices. It is also recognized that other configurations of these exemplar devices are possible (e.g., desktops with a touch screen) and could be incorporated into this figure as well. So, with that as a backdrop, desktops are presented at the lower end of the construct-irrelevant cognitive load continuum because they are characterized by (a) lower working memory demands because of their larger screens, (b) lower perceptual speed and visual acuity demands because of lower screen clutter, (c) lower psychomotor ability demands because of a less challenging (non-challenging) response interface, and (d) lower selective attention demands because of lower levels of distraction resulting from lower levels of permissibility. In contrast, smartphones are at the higher end of the construct-irrelevant cognitive load continuum because they are characterized by (a) higher working memory demands because of their smaller screens, (b) higher perceptual speed and visual acuity demand because of higher screen clutter, (c) higher psychomotor ability demands because of a relatively more challenging response interface, and (d) higher selective attention demands because of higher levels of distraction resulting from higher levels of permissibility.

However, again, although the proceeding example used exemplar devices that were simultaneously “high” or “low” on all the information-processing demands, this is of course not necessarily always the case. As with the structural characteristics, any configuration of information-processing demands is possible; it is a function of the specific configuration of structural characteristics that underlie the specific device. Thus, the SCIP framework also has the added advantage of permitting the integration of a wide range of UIT-device-type configurations (e.g., using an Xbox on a 55-in. TV; using a desktop in a noisy public space such as a university computer lab or a public library), as well as even newer technologies as they emerge, as long as they engender and subsequently can be conceptualized in terms of the degree of construct-irrelevant information-processing demands.

In summary, instead of framing mobile and nonmobile devices in terms of whether they are tethered to the wall, the SCIP framework allows one to place UIT devices on a continuum in terms of how they vary on the additional cognitive load engendered by the presence and role of construct-irrelevant information-processing variables (i.e., working memory, perceptual speed and visual acuity, psychomotor ability, and selective attention) that are required for their effective use as a result of how they differ on the specified structural characteristics, that is, screen size, screen clutter, response interface, and permissibility. Hence, to the extent that UIT device types are similar or are not different in terms of the level of construct-irrelevant cognitive load engendered by said devices, one would not expect device-type effects in test scores. In contrast, to the extent that they differ in the level of construct-irrelevant cognitive load engendered by the devices, then one would expect device-type effects in assessment and test scores. Thus, as suggested in Figure 3 and posited in the next section, whereas we expect UIT-device-type effects in desktop (nonmobile) versus smartphone (mobile) comparisons (when measuring cognitive constructs), we do not expect device-type effects for desktop (nonmobile) versus notebook (mobile) comparisons. Furthermore, we would also expect differences in tablet (mobile) versus smartphone (mobile) comparisons. These examples highlight the conceptual advantages to using the SCIP framework to conceptualize UIT device-types instead of the mobile versus nonmobile distinction.

Some research propositions

Test performance

There are a number of testable propositions that arise from the SCIP framework. For instance, to the extent that a UIT device places additional construct-irrelevant cognitive load on the test-taker, if said device is used to take a cognitive assessment (which already engenders higher levels of cognitive processing and demands), then the test-taker should obtain lower scores than if the same assessment
were completed on a device that places lower construct-irrelevant cognitive load on the test-taker. This issue can also be framed in terms of test difficulty in that introducing additional construct-irrelevant cognitive load on a cognitive assessment makes responding to items more difficult, resulting in lower scores. In contrast, because test difficulty, as traditionally conceptualized, is more of an issue for cognitive than noncognitive assessments, then noncognitive assessments should be less influenced by the demands engendered by the differences in construct-irrelevant cognitive load (because noncognitive assessments by definition require lower levels of cognitive processing and demands). Thus, UIT devices that differ in terms of cognitive demands on the latter should not display differences in test scores when used to complete noncognitive assessments; and even if there are differences, they should be relatively substantially smaller than those observed for cognitive constructs. As reflected in the preceding review of the literature, this proposition is in line with the pattern of results that have been observed for score differences (or lack thereof) on mobile and nonmobile devices for cognitive and noncognitive assessments, respectively.

**Proposition 1:** When UIT devices at the higher end of the device-engendered cognitive load continuum (e.g., smartphones) are used to measure cognitive constructs, mean scores on the focal constructs will be lower than when said constructs are assessed using devices at the lower end of continuum (e.g., desktop computers).

**Proposition 2:** When UIT devices at the higher end of the device-engendered cognitive load continuum (e.g., smartphones) are used to measure noncognitive constructs, mean scores on the focal constructs will not be different from those assessed on devices at the lower end of continuum (e.g., desktop computers).

**Information-processing variable relationships**

Another set of testable propositions pertain to the information-processing variable relationships. Specifically, to the extent that the tenets of the SCIP framework are sound, then when used to assess cognitive constructs, because of the higher cognitive load associated with taking cognitive assessments, the relationships between the information-processing variables (i.e., working memory, perceptual speed and visual acuity, psychomotor ability, and selective attention) and scores on UIT devices on the higher end of the continuum should be stronger than those for devices on the lower end of the continuum. On the other hand, because of the lower cognitive load associated with taking noncognitive assessments, when used to measure noncognitive constructs, there should not be any differences in the observed relationships between the information-processing variables and device-type scores at the higher and lower ends of the continuum.

**Proposition 3:** When UIT devices at the higher end of the device-engendered cognitive load continuum (e.g., smartphones) are used to measure cognitive constructs, there will be stronger relationships between the construct-irrelevant information-processing variables and the focal cognitive construct scores compared to devices at the lower end of the continuum (e.g., desktop computers).

**Proposition 4:** When used to measure noncognitive constructs, the magnitude of the relationships between the construct-irrelevant information-processing variables and the focal noncognitive construct scores will be (a) weak and (b) similar for UIT devices at the higher (e.g., smartphones) and lower end (e.g., desktop computers) of the device-engendered cognitive load continuum.
Criterion-related validity
As previously noted, the absence of any examinations of the comparative criterion-related validity of UIT device types is a major gap in the literature. However, the viability of the SCIP framework as a conceptually sound framework can be examined in reference to propositions pertaining to criterion-related validity. The SCIP framework poses the specified information-processing variables as sources of construct-irrelevant cognitive load because these variables are not the focal construct of interest in the assessment (see Figure 2). However, many of these variables have established demonstrable relationships with organizationally relevant outcomes and performance. For instance, working memory has been shown to be predictive of performance on a wide range of simple and complex tasks and activities that are relevant across jobs, including problem solving, multitasking, and decision making (e.g., Colom, Martínez-Molina, Shih, & Santacreu, 2010; Cowan et al., 2005; Edwards, Franco Watkins, McAbee, & Faura, 2017). Perceptual speed and visual acuity have similarly been linked to job performance across multiple domains (Bertua, Anderson, & Salgado, 2005) and demonstrate incremental validity above cognitive ability for particular jobs (air traffic control: Ackerman & Ciancio, 2000; clerical: Hunter & Schmidt, 1996; warehousing: Mount, Oh, & Burns, 2008). There is also a well-established body of literature that supports psychomotor ability as a predictor of job performance, including the performance of pilots (e.g., Carretta & Ree, 2000; Wheeler & Ree, 1997) and utility workers (technicians, electricians, mechanics; Levine, Spector, Menon, & Narayanan, 1996), to name a few. Selective attention has also been linked to job performance and particularly in the context of driving behavior and accidents (Arthur & Doverspike, 1992; Clay et al., 2005; Myers, Ball, Kalina, Roth, & Goode, 2000).

Consequently, whereas they may not be the focal constructs of interest, to the extent that the specified SCIP information-processing variables may be relevant in the context of the job in question (e.g., perceptual speed for air traffic controllers: Ackerman & Ciancio, 2000; psychomotor ability for utility workers: Levine et al., 1996), then use of UIT devices that engender the additional cognitive load may unintendly contribute to the criterion-related validity of the measure. That is, although, for instance, the use of a smartphone to complete a GMA test may increase the role of working memory in the assessment score (more so than using a desktop), to the extent that working memory is relevant to successful performance on the criterion, using a smartphone in this context should result in a higher criterion-related validity than the use of a desktop computer. As another example, the use of UIT devices that have challenging response interfaces, on the basis of the tenets of the SCIP framework, would be expected to exhibit higher levels of criterion-related validity in the context of jobs for which perceptual speed and visual acuity are important for success (relative to devices with a nonchallenging response interface). This reasoning is consistent with tenets of Proposition 3 (and 4), and the results of Arthur, Keiser, Hagen, and Traylor (2017), who found the working memory/GMA relationship to be stronger for assessments completed on smartphones ($r = .29$) compared to desktop computers ($r = .14$).

Proposition 5. When UIT devices at the higher end of the device-engendered cognitive load continuum (e.g., smartphones) are used to measure constructs (both cognitive and noncognitive) in the context of jobs for which a specified SCIP information-processing variable is also important for successful performance on the criterion, the observed criterion-related validity for the test scores should be higher compared to that obtained for completion of the assessment on devices at the lower end of the continuum (e.g., desktop computers).

Subgroup differences
A conclusion drawn for the review of the literature is that, whereas there are no mean differences on noncognitive (e.g., personality) assessments taken on mobile and nonmobile devices (and when present, they are very small), under high-stakes conditions where test-takers select their assessment device, there are pronounced differences for cognitive constructs with scores on mobile devices being
substantially lower. This has implications for subgroup differences and subsequently potential adverse impact. That is, if taking cognitive tests on mobile devices results in lower scores, and the tendency to take assessments on mobile devices covaries with specified protected group status, then this raises the specter of observed subgroup differences and higher adverse impact potential resulting from the use of certain UIT devices (e.g., smartphone) in employment-related assessments but not others. However, there is no research of which we are aware that has examined this issue for cognitive constructs (cf. Golubovich & Boyce, 2013; Kinney et al., 2014; McClure Johnson & Boyce, 2015, for noncognitive constructs; Rossini, 2016, which failed to distinguish constructs from methods). In an initial attempt to rectify this, we further examined Arthur, Doverspike, et al.’s (2014, Table 9) results, which reported an overall $d$ of 0.90, reflecting lower scores on mobile devices. Disaggregating these results by race/ethnicity indicated that the White–African American $d$ was 0.68 for mobile devices and 0.84 for nonmobile devices; that is, in Arthur, Doverspike, et al., UIT device type did not interact with demography to result in larger subgroup differences. Indeed, paradoxically, in these data, it appears to make the subgroup differences smaller, a pattern of results that is similar to those reported by Arthur, Edwards, and Barrett (2002) and Edwards and Arthur (2007) in their comparisons of constructed-response and multiple-choice tests.

In spite of the preceding, from the perspective of the SCIP framework, because UIT device types are differentiated in terms of the additional information-processing cognitive load, the issue is one of whether there are subgroup differences on the specified information-processing variables that would subsequently further contribute to observed differences. Of the four information-processing variables, psychomotor ability stands out as one on which there are well-established sex differences. These differences, favoring male individuals (e.g., Barnett, van Beurden, Morgan, Brooks, & Beard, 2010; Hyde, 2005; Sanchez-Ku & Arthur, 2000; Thomas & French, 1985; Jarrett, Glaze, Schurig, & Arthur, 2017), are commonly advanced as explanations for observed differences on tasks with high psychomotor demands, and even video-game-based tasks. Consequently, consonant with this literature, to the extent that UIT devices differ in the amount of cognitive load engendered by differing levels of psychomotor demands, one could reasonably posit that when UIT devices on the higher end of the device-engendered cognitive load continuum, especially in terms of psychomotor ability, are used to complete cognitive assessments, observed sex-based subgroup differences will be larger compared to assessments completed on the lower end of the continuum. Furthermore, the observed differences will be larger for cognitive constructs compared to noncognitive constructs, and maybe even more so with high-fidelity synthetic tasks environments, simulations, and gamification (Arthur, Doverspike, et al., 2017).

**Proposition 6:** When UIT devices at the higher end of the device-engendered cognitive load continuum, especially in terms of psychomotor ability, are used to measure cognitive constructs, there will be sex-based mean score differences, which will be larger than those observed for devices at the lower end of the continuum (with less psychomotor ability demands).

**Proposition 7:** The observed sex-based UIT-device-type differences will be larger for cognitive constructs compared to noncognitive constructs.

Concerning Propositions 6 and 7, given the well documented age-related differences in information processing ability (Hasher & Zacks, 1988; Salthouse & Babcock, 1991; Verhaeghen & Salthouse, 1997) similar corresponding age-based propositions could be advanced for these propositions as well.

**Test completion time**

Test difficulty has been construed in terms of not only content but also the method or mode of assessment (e.g., Bergstrom, Lunz, & Gershon, 1992; Ponsoda, Olea, Rodriguez, & Revuelta, 1999; Thiede, 1996; Tonidandel, Quiñones, & Adams, 2002). Hence, to the extent that the increased
cognitive load associated with the construct-irrelevant information-processing variables introduces cognitive demands that make the assessment more difficult, then one would expect test-takers to take longer to complete the assessments. This reasoning explains the finding that for noncognitive constructs that are untimed, the use of handheld small screen devices is associated with longer completion times (Arthur et al., 2014; Dages & Jones, 2015; Illingworth et al., 2015). So, consonant with this, the SCIP framework predicts stronger relationships between the information-processing variables and completion time for devices at the higher end of the device-engendered cognitive load continuum compared to those at the lower end.

Proposition 8: For assessments that are untimed (not speeded), UIT devices at the higher end of the device-engendered cognitive load continuum (e.g., smartphones) will display longer completion times.

Proposition 9: For assessments that are untimed (not speeded), there will be a stronger relationship between the construct-irrelevant information-processing variables and completion times for devices at the higher end of the device-engendered cognitive load continuum (e.g., smartphones) compared to those at the lower end.

Test-taker reactions and preferences
Test difficulty, again broadly construed in terms of both content and method of assessment, has also been demonstrated to influence reactions to assessments (e.g., Hong, 1999; King et al., 2015; Tonidandel et al., 2002) with higher levels of difficulty being associated with more negative reactions. Hence, as per the SCIP framework, to the extent that the increased cognitive load associated with construct-irrelevant information-processing variables introduces additional challenges to completing assessments on UIT devices, then one would expect reactions to assessments completed on devices at the higher end of the device-engendered cognitive load continuum to be more negative. This is consonant with the general negative reactions observed for assessments completed on mobile devices in both operational and lab settings (e.g., Chang et al., 2016a; Fursman & Tuzinski, 2015; Gutierrez & Meyer, 2013; Huff, 2015; King et al., 2015; Smeltzer, 2013).

Proposition 10: Test-takers will have comparatively less favorable reactions to and lower preferences for UIT devices at the higher end of the device-engendered cognitive load continuum (e.g., smartphones) compared to those at the lower end of the continuum.

Proposition 11: The difference in reactions and preferences observed for UIT device types will be stronger when UIT devices are used to assess cognitive compared to noncognitive constructs.

In summary, the SCIP framework presents a conceptual explanation for the observed relationships between UIT device types and score differences or lack thereof on cognitive and noncognitive assessments. More important, not only does it adequately and parsimoniously explain and account for the measurement outcome-related effects observed in the extant literature, it also permits the formulation of empirically testable propositions over and beyond what is currently present in the extant literature.

Discussion and conclusion
As previously noted, the theoretical or conceptual work that seeks to explain UIT-device-type effects on assessment and test scores is limited. Indeed Potosky (2008) is the only other formal conceptual framework of which we are aware. Potosky’s framework is embedded in communication theory; that is, Potosky views personnel tests and assessments as a communicative act between the test-taker and
organization (or the individual or anyone who wants to measure attributes of the test-taker). Consequently, the focus is on how the communication channel (e.g., face-to-face vs. telephone interview) or the medium’s structural attributes affect the message quality and, hence, test scores. Various attributes of the test administration medium (i.e., transparency, social bandwidth, interactivity, and surveillance) are used in this communication process that can in turn affect assessment outcomes. In contrast, the SCIP framework represents a framework that seeks to explain and predict the conditions under which UIT devices have different effects on a range of assessment-related outcomes (test performance, test completion time, criterion-related validity, subgroup differences, test-taker reactions and perceptions) as a function of the construct assessed (cognitive vs. noncognitive), and the interaction between construct-irrelevant information-processing demands and associated structural characteristics of the UIT device. In addition, its breadth is reflected in the proposition that the tenets of the framework apply to not only a wide range of UIT-device-type configurations but indeed any assessment media or context in which construct-irrelevant information-processing demands associated with the testing method or medium are germane.

Thus, as frameworks for conceptually explaining and accounting for UIT-device-type assessment-related outcomes, Potosky’s (2008) framework and the SCIP framework are fundamentally dissimilar. Potosky conceptualized attributes of test medium based on how they influence communication or message quality and test scores, whereas the SCIP framework explains test score differences as a function of a finite pool of cognitive information-processing resources differentially drawn from test-takers as they complete assessments. As such, consonant with the pivotal role that individual differences play in personnel assessment and testing (Sackett, Lievens, Van Iddekinge, & Kuncel, 2017), the SCIP framework conceptualizes assessment device-type effects in terms of how individual differences of the test-taker on specified information-processing attributes interact with the structural characteristics of the assessment device to generate construct-irrelevant cognitive load on the test-taker that subsequently influences the individual’s performance on the assessment (see Figure 2).

Although up to this point the role and influence of the construct-irrelevant information-processing variables have been discussed somewhat independently, this is not intended to discount the fact that conceptually they likely have related and interactive effects. So, for instance, the posited psychomotor effects are not unrelated to screen size, as smaller screen sizes usually translate into more screens being required to present the same amount of information and consequently, a higher frequency or amount of interaction with the device, and hence higher levels of finger dexterity demands, thus further magnifying the role of psychomotor ability. The presentation of more screens would also amplify the working memory demands as well. As an additional example of these interactive effects, smaller screen sizes also mean higher levels of clutter and thus higher perceptual speed and visual acuity demands. A question that arises from this interrelatedness is the relative importance of the information-processing variables in explaining or accounting for the observed device-type effects. For instance, it is interesting to note that all the lab studies that we are aware of, where UIT devices were compared but in a proctored environment, which by so doing (inadvertently) eliminated the distraction structural characteristic and thus the role of selective attention (e.g., Smeltzer, 2013), failed to obtain device type differences on cognitive constructs. This suggests that permissionality and thus distractions might be a particularly important or pivotal factor (e.g., see Chan et al., 2016b; Lawrence, Kinney, O’Connell, & Delgado, 2017). In summary, it would be informative for future research to examine the relative importance of the four construct-irrelevant information-processing variables in accounting for the observed device-type effects. On a related note, future research could also examine the relationships between the information-processing variables for a specified device, and whether these relationships systematically vary across device types as well.

As illustrated in Figure 3, we posit that as exemplars, devices like desktops, laptops, notebooks, and tablets are likely similar enough on the additional cognitive load engendered to not result in assessment-related outcome differences between these devices, in contrast to smartphones. This then
suggests that the mobile versus nonmobile classification that characterizes the literature is conflated with other factors that likely obfuscate the research findings, as on the basis of the SCIP framework, tablets and smartphones, for instance, will result in different patterns of results and thus should not be collapsed into a single “mobile” category. Hence, the SCIP framework highlights and seeks to address methodological flaws with the extant literature, which has generally failed to recognize fine distinctions between UIT device types (cf. Brown et al., 2016; Dages & Jones, 2015; Parker & Meade, 2015) as reflected in the different information-processing demands that they engender. In short, researchers need to report the specific UIT device used to operationalize their mobile and nonmobile conditions, as the SCIP framework would suggest differences between some mobile devices (e.g., smartphones vs. tablets) but not others (e.g., tablets vs. notebooks)—indeed, no differences between some mobile (e.g., laptops, notebooks) and nonmobile devices (e.g., desktops).

In summary, the SCIP framework offers a number of advantages and distinguishing features. Specifically, first, it explains and accounts for UIT-device-type score differences by relying on theories of individual differences and associated information-processing demands and their interaction with device-type structural characteristics. Second, it advances several empirically testable formal propositions, and support or lack thereof for them should ultimately determine the viability of the framework as an explanatory framework. Third, it provides guidance to future research on how to classify and describe UIT device types in comparative studies, that is, psychologically and conceptually instead of technologically (i.e., mobile [wireless] vs. nonmobile [wired]). Fourth, it permits the examination of a wide range of UIT-device-type configurations (e.g., an Xbox on a 53-in. TV; a desktop with a touch-screen interface; a tablet with an external keyboard, etc.), as well as newer technologies as they emerge—as long as they engender and subsequently can be conceptualized in terms of the level of construct-irrelevant information-processing demands. Finally, it conceptually informs discussions of a wider range of assessment methods and modes, particularly in any domain in which construct-irrelevant information-processing demands associated with the testing method or medium are pertinent.

In conclusion, technological advancements have led to the widespread use of UIT assessments and, concomitantly, the increasing use of different UIT device types to complete these assessments. However, a comprehensive literature search and subsequent review indicated that the empirical literature is quite limited, with the vast majority being conference presentations instead of peer-reviewed articles; nevertheless, the detailed review presented here (see appendices) should provide both researchers and practitioners access to a detailed summary of the current state of the pertinent literature that can inform their scholarly and applied practices. Another conclusion is that the conceptualization of UIT devices as mobile versus nonmobile has probably played a retardant role in advancing psychologically sound conceptual developments in this domain; being mobile versus nonmobile cannot be the psychological process via which UIT device types affect test scores. So, in an effort to advance the psychological basis of the UIT-device-type literature, this article presents the SCIP framework as a framework for conceptualizing UIT device types in terms of the construct-irrelevant information-processing demands (i.e., working memory, perceptual speed and visual acuity, psychomotor ability, and selective attention) that are engendered by the structural characteristics of the device (i.e., screen size, screen clutter, response interface, and permissibility) that differentially affect test scores as a function of the constructs assessed. So, in addition to explaining general findings in the literature, the SCIP framework also permitted the formulation of several empirically testable propositions that follow from the framework.

Finally, we would also note that although our focus in this article has been specifically and exclusively on UIT assessment, as previously noted, the SCIP framework could conceptually inform discussions of a wider range of assessments, particularly in any domain in which construct-irrelevant information-processing demands associated with the testing method or medium are germane. So, for instance, it would be germane in the context of other technologically mediated assessment methods such as virtual role-plays, immersive simulations, and gamified assessments (Arthur et al., 2017). As another example, one could envisage characteristics associated with SJT items, for instance (e.g., item
length, complexity of instructions, response format) that could very well engender differential construct-irrelevant cognitive load (Arthur et al., 2014). The basic approach represented by the SCIP framework could also be used to explain the observed score differences between SJT modes (e.g., paper-and-pencil vs. video; Chan & Schmitt, 1997), and for paper-and-pencil tests, the observed differences between multiple-choice and constructed-response formats (Arthur et al., 2002; Edwards & Arthur, 2007). Nevertheless, it would seem that the first step is for future research to undertake empirical tests of the propositions presented here in an effort to evaluate the viability of the framework as an explanatory framework and, on the basis of these studies, refine and/or expand on it (and maybe even reject it) as warranted.

Notes

1. We recognize that even desktop computers can have wireless Internet cards that technologically untether them from the wall. However, they nevertheless would still be fixed-location devices.
2. Of the 11 studies that did not report differences, the effect sizes were not zero; however, they were very small and, in terms of their direction, did not consistently favor one device type (nonmobile devices) compared to the other (mobile devices).
3. For instance, when contacted for copies of their symposium presentations, authors almost always sent us copies of presentation slides and not manuscripts or papers. That being said, because these are generally not as accessible as peer-reviewed articles, the summaries provided in the appendixes should be of some informational value to scholars and practitioners.
4. Voice is obviously another possible response interface; however, it is unclear how many high-stakes test-takers will or actually use this as a mode of responding.
5. Distractions can take the form of not only noise and other factors external to the device but may include factors such as low or erratic connectivity, and other distractors such as pop-up messages, buzzes, notifications, and the like.
6. See Morelli, Potosky, Arthur, and Tippins (2017) for a fuller discussion of the differences between these two frameworks.

References

*Reviewed and summarized in Appendix A.
**Summarized in Appendix B.


Appendix A

Summary of Empirical Comparative Unproctored Internet-Based Testing Device-Type Studies—Unproctored Assessments Only


<table>
<thead>
<tr>
<th>Authors</th>
<th>Source</th>
<th>Operational vs. Lab</th>
<th>Sample Size</th>
<th>Constructs Assessed</th>
<th>Test Type/Method</th>
<th>Timed vs. Untimed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Arthur et al. (2014)</td>
<td><em>International Journal of Selection and Assessment</em></td>
<td>Operational</td>
<td>N = 3.5M Mobile = 1.93% (69,144) - mobile = larger numbers of AA, H, women, and younger</td>
<td>Noncognitive (FFM) and cognitive (GMA)</td>
<td>Likert scale and MC</td>
<td>Personality = untimed GMA = speeded</td>
</tr>
<tr>
<td>2. Brown et al. (2016)</td>
<td>SIOP conference</td>
<td>Lab (MTurk sample)</td>
<td>N = 1,769; desktop or laptop = 724; tablet = 46; smartphone = 569 - tablet and smartphone assignment by device ownership</td>
<td>Cognitive (GMA)</td>
<td>MC</td>
<td>Yes (speeded)</td>
</tr>
<tr>
<td>3. Chang et al. (2016a)</td>
<td>SIOP conference</td>
<td>Lab (MTurk sample)</td>
<td>N = 143; within-subjects design with ≥ 2-week interval; Time 1 = PC or laptop; Time 2 = mobile device (tablet or smartphone) of participant’s choice.</td>
<td>Not specified</td>
<td>High-fidelity interactive simulation</td>
<td>NR</td>
</tr>
<tr>
<td>4. Dages and Jones (2015)</td>
<td>SIOP conference</td>
<td>Operational</td>
<td>Sample 1 = 11,371; 156-item management-level assessment, mobile = 14% Sample 2 = 7,545; 50-item, retail and manufacturing applicants, mobile = 8%</td>
<td>Noncognitive</td>
<td>Likert scale</td>
<td>No</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measurement Equivalence</th>
<th>Mean Differences</th>
<th>Criterion-Related and Differential Validity</th>
<th>Test-Taker Reactions</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Yes</td>
<td>GMA = yes, d = 0.90</td>
<td>NR</td>
<td>NR</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Personality = no</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- personality: completion time differences</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. NR</td>
<td>No significant differences in scores</td>
<td>NR</td>
<td>NR</td>
<td>No formal tests of measurement equivalence, but convergent validities suggestive of equivalence</td>
</tr>
<tr>
<td></td>
<td>- computer/tablet d = −0.05; computer/smartphone d = −0.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>No significant differences in # items completed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- computer/tablet d = −0.08; computer/smartphone d = −0.10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. NR</td>
<td>Yes; lower mobile device scores, d = 0.43</td>
<td>NR</td>
<td>For mobile condition, lower levels of satisfaction for screen sizes less than 4 in.</td>
<td></td>
</tr>
<tr>
<td>4. NR</td>
<td>Similar pass rates</td>
<td>NR</td>
<td>NR</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Longer completion times</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Of multiple device users, large percentage (89% for 156-item measure; 74% for 50-item) started on smartphone and switched to computer</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note. AA = African American; H = Hispanic; FFM = five-factor model of personality dimensions; GMA = general mental ability; MC = multiple-choice text/exam; SIOP = Society for Industrial and Organizational Psychology; PC = personal computer; NR = not reported.*
Table A2. Summaries of Studies A5–A8.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Source</th>
<th>Operational vs. Lab</th>
<th>Sample Size</th>
<th>Constructs Assessed</th>
<th>Test Type/Method</th>
<th>Timed vs. Untimed</th>
</tr>
</thead>
<tbody>
<tr>
<td>5. Fursman (2016)</td>
<td>SIOP conference</td>
<td>Lab → (&quot;potential job candidates via online website offering practice tests&quot;; entered into drawing for &quot;small monetary prize&quot;)</td>
<td>Operational</td>
<td>Cognitive ability (numerical ability)</td>
<td>MC</td>
<td>NR</td>
</tr>
<tr>
<td>6. Fursman and Tuzinski (2015)</td>
<td>SIOP conference</td>
<td>Operational</td>
<td>N = 8,176; % mobile unspecified</td>
<td>Noncognitive and cognitive</td>
<td>Likert scale, NR</td>
<td>NR</td>
</tr>
<tr>
<td>7. Golubovich and Boyce (2013)</td>
<td>SIOP conference</td>
<td>Operational</td>
<td>N = 12.9M Mobile = 7.88% = 912,045 - mobile increasing trend over 5 yrs; 2009 = 3.1%, 2013 = 14.31% - mobile, greater #s of AA, H, &amp; women</td>
<td>Noncognitive (e.g., conscientiousness, emotional stability)</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>8. Grossenbacher et al. (2016)</td>
<td>SIOP conference</td>
<td>Lab (MTurk sample)</td>
<td>N = 2,787; computer = 1,700; tablet = 494; smartphone = 593</td>
<td>Cognitive (GMA)</td>
<td>MC</td>
<td>Yes (speeded)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measurement Equivalence</th>
<th>Mean Differences</th>
<th>Criterion-Related and Differential Validity</th>
<th>Test-Taker Reactions</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>5. Yes; no DIF</td>
<td>Yes</td>
<td>NR</td>
<td>Reported for mobile no warning and warning conditions</td>
<td>- no systemic difference between them; similar levels of preference</td>
</tr>
<tr>
<td>6. NR</td>
<td>NR</td>
<td>NR</td>
<td>Survey of reactions to smartphone assessments; low preference for completing employment assessments on smartphones; considered unfair; availability of mobile assessment does not influence decision to apply or accept offer</td>
<td>—</td>
</tr>
<tr>
<td>7. NR</td>
<td>No adverse impact</td>
<td>NR</td>
<td>NR</td>
<td>—</td>
</tr>
<tr>
<td>8. Yes; no DIF</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>—</td>
</tr>
</tbody>
</table>

Note. SIOP = Society for Industrial and Organizational Psychology; MC = multiple-choice text/exam; NR = not reported; SJT = situational judgment test; AA = African American; H = Hispanic; GMA = general mental ability; DIF = differential item functioning.
**Table A3.** Summaries of Studies A9–A12.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Source</th>
<th>Operational vs. Lab</th>
<th>Sample Size</th>
<th>Constructs Assessed</th>
<th>Test Type/Method</th>
<th>Timed vs. Untimed</th>
</tr>
</thead>
<tbody>
<tr>
<td>9. Gutierrez et al. (2015)</td>
<td>SIOP conference</td>
<td>Lab (MTurk sample)</td>
<td>Tablet = 130; mobile phone = 180 (58%); computer (desktop/laptop) n = 6,557; mobile n = 3,130 (32%)</td>
<td>Noncognitive and cognitive</td>
<td>MC</td>
<td>NR</td>
</tr>
<tr>
<td>11. Illingworth et al. (2015)</td>
<td>Journal of Business and Psychology</td>
<td>Operational</td>
<td>N = 929,341; Mobile = 0.83% = 7,743</td>
<td>Noncognitive (i.e., conscientiousness, customer service, integrity, interpersonal, stress tolerance, &amp; teamwork)</td>
<td>Personality &amp; biodata</td>
<td>37 self-report items; 75 min</td>
</tr>
<tr>
<td>12. Impelman (2013)</td>
<td>SIOP conference</td>
<td>Operational</td>
<td>N = 615,350; Mobile = 4.5% = 27,691</td>
<td>Noncognitive (i.e., “facets of the Big 5”) &amp; cognitive ability</td>
<td>NR</td>
<td>NR</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Authors</th>
<th>Source</th>
<th>Operational vs. Lab</th>
<th>Sample Size</th>
<th>Constructs Assessed</th>
<th>Test Type/Method</th>
<th>Timed vs. Untimed</th>
</tr>
</thead>
<tbody>
<tr>
<td>9. NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>More satisfied as screen size increases</td>
<td>Low-stakes MTurk sample</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Smartphones less satisfied with item types with more text than those with graphics</td>
<td>40% → regardless of completing assessment on phone or tablet indicated preference to completing assessment on mobile devices to computers</td>
<td></td>
</tr>
<tr>
<td>10. NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>Between-subjects design</td>
<td>Assessments were all noncognitive</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Only 29% indicated more likely to apply if mobile option available. Across computer, tablet, and phone users, no differences in reported ease of test completion, and opportunity to performance. Only 55% of tablet and phone users reported preference for mobile device over computer; compared to only 7% of computer users. 89% and 90% of tablet and phone users considered mobile device use to be fair; only 49% of computer users.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Yes</td>
<td>No</td>
<td>NR</td>
<td>NR</td>
<td>Personality = no; cognitive ability d = 0.31</td>
<td>Comparison of hourly and management. “Less educated and more internal candidates utilizing mobile devices at greater rate.”</td>
<td></td>
</tr>
<tr>
<td>12. NR</td>
<td>Personality = no; cognitive ability d = 0.31</td>
<td>NR</td>
<td>NR</td>
<td>Comparison of hourly and management. “Less educated and more internal candidates utilizing mobile devices at greater rate.”</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. SIOP = Society for Industrial and Organizational Psychology; MC = multiple-choice text/exam; NR = not reported.
<table>
<thead>
<tr>
<th>Authors</th>
<th>Source</th>
<th>Operational vs. Lab</th>
<th>Sample Size</th>
<th>Constructs Assessed</th>
<th>Test Type/Method</th>
<th>Timed vs. Untimed</th>
</tr>
</thead>
<tbody>
<tr>
<td>13. Kinney et al. (2014)</td>
<td>SIOP conference</td>
<td>Operational</td>
<td>N = 129,000</td>
<td>Noncognitive (i.e., safety awareness, quality focus, impulsivity, dependability, positive attitude)</td>
<td>Personality, SJT, &amp; biodata</td>
<td>NR</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- mobile ≤ smartphone = 10% = 12,908; tablet = 3.12% = 4,019</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- mobile = larger # of minorities &amp; males (except healthcare)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. LaPort (2016)</td>
<td>SIOP conference</td>
<td>Operational</td>
<td>N = 14,211 applicants (desktop, laptop, tablet, smartphone); device-type ns not reported</td>
<td>Noncognitive¹ (managing people and interactions, managing and adapting to change, work and achievement orientation) Cognitive reasoning² (minimal scrolling)</td>
<td>Likert scales¹ MC²</td>
<td>NR</td>
</tr>
<tr>
<td>15. Lawrence et al. (2016)</td>
<td>SIOP conference</td>
<td>Lab (MTurk sample)</td>
<td>N = 300; within-subjects design; PC then mobile device with 10-30-day interval</td>
<td>Noncognitive</td>
<td>High-fidelity simulation</td>
<td>NR</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- mobile → 30% tablet, 70% smartphone</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16. Lawrence et al. (2013)</td>
<td>SIOP conference</td>
<td>Operational</td>
<td>N = 200,000+ Mobile = 8% = 15,738</td>
<td>Noncognitive (i.e., attention to detail, stress tolerance, likelihood of turnover &amp; absenteeism, services &amp; sales potential)</td>
<td>NR</td>
<td>NR</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measurement Equivalence</th>
<th>Mean Differences</th>
<th>Criterion-Related and Differential Validity</th>
<th>Test-Taker Reactions</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>13. Yes</td>
<td>No; adverse impact analyses at 15% &amp; 50% “cut rates”</td>
<td>NR</td>
<td>Between-subjects design</td>
<td>Comparisons across industrial sectors (i.e., manufacturing, retail, &amp; healthcare). However, because there are no differences, the preceding is collapsed across sections.</td>
</tr>
<tr>
<td>14. NR</td>
<td>Yes</td>
<td>NR</td>
<td>NR</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>- in all cases smartphone users scored lower than other device types; desktop &amp; laptop vs. smartphone differences larger than tablet vs. smartphone - negligible to small differences for noncognitive tests (M d = 0.13) - medium differences for cognitive test with minimal scrolling (M d = 0.46) - large differences for cognitive test with higher scrolling requirements (M d = 0.60)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15. NR</td>
<td>Yes</td>
<td>NR</td>
<td>NR</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>- higher scores for PC; accuracy d = 0.16, speed d = 0.64, time-to-completion d = 0.61.</td>
<td>Reported for mobile only; across various criteria, ranged from r = .11-.22 (avg = .17)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16. Yes</td>
<td>No</td>
<td>NR</td>
<td>NR</td>
<td>Comparisons based on balanced samples (i.e., randomly selected subset of nonmobile users to match mobile users sample size)</td>
</tr>
</tbody>
</table>

*Note. Matching superscripts specify which constructs were assessed using what test type/method. SIOP = Society for Industrial and Organizational Psychology; SJT = situational judgment test; NR = not reported; GMA = general mental ability; MC = multiple-choice text/exam.*
Table A5. Summaries of Studies A17–A20.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Source</th>
<th>Operational vs. Lab</th>
<th>Sample Size</th>
<th>Constructs Assessed</th>
<th>Test Type/Method</th>
<th>Timed vs. Untimed</th>
</tr>
</thead>
<tbody>
<tr>
<td>17. McClure Johnson and Boyce (2015)</td>
<td>SIOP conference</td>
<td>Operational</td>
<td>N = 18,557,571; PC = 84%; mobile = 12%, in-store kiosk = 4%</td>
<td>Noncognitive</td>
<td>&quot;Personality, biodata, and situational judgment items&quot;</td>
<td>Untimed</td>
</tr>
<tr>
<td>18. Morelli, Illingworth, Moon, Scott, and Boyd (2013)</td>
<td>SIOP conference</td>
<td>Operational</td>
<td>N = 664,469 Mobile = 0.78% = 5,182</td>
<td>Noncognitive (i.e., conscientiousness, interpersonal skills, teamwork, curiosity, customer service, adaptability)</td>
<td>Personality &amp; biodata</td>
<td>32 self-report items; 75 min</td>
</tr>
<tr>
<td>19. Morelli et al. (2014)</td>
<td>International Journal of Selection and Assessment</td>
<td>Operational</td>
<td>N = 608,518 Mobile = 3.98% = 24,192</td>
<td>Cognitive = “learning”; noncognitive (i.e., conscientiousness, customer service orientation, neatness, customer service)</td>
<td>MC, biodata, multimedia/simulation, &amp; SJT</td>
<td>NR</td>
</tr>
<tr>
<td>20. Parker and Meade (2015)</td>
<td>SIOP conference</td>
<td>Lab (MTurk sample)</td>
<td>N = 692; random assignment to computer (nonmobile), smartphone-optimized, and smartphone-accessible conditions</td>
<td>Cognitive and noncognitive</td>
<td>Raven’s APM (MC); IPIP Likert scales</td>
<td>Yes?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measurement Equivalence</th>
<th>Mean Differences</th>
<th>Criterion-Related and Differential Validity</th>
<th>Test-Taker Reactions</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>17. NR</td>
<td>Small but lower mobile device scores for entry-level jobs, but less so for manager assessment</td>
<td>NR</td>
<td>NR</td>
<td>—</td>
</tr>
<tr>
<td>18. Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19. Yes</td>
<td>No with exception of customer service SJT, d = 0.25</td>
<td>NR</td>
<td>NR</td>
<td>&quot;learning&quot; = &quot;ability to learn and follow a set of written instructions for how a guest room or facility area should be arranged or maintained&quot;</td>
</tr>
<tr>
<td>20. IRT; scores invariant across device types</td>
<td>No mean differences across device types</td>
<td>NR</td>
<td>NR</td>
<td>—</td>
</tr>
</tbody>
</table>

Note. SIOP = Society for Industrial and Organizational Psychology; PC = personal computer; MC = multiple-choice text/exam; APM = advanced progressive matrices; IPIP = international personality item pool; UIT = unproctored Internet-based testing; SJT = situational judgment test; NR = not reported; IRT = item response theory.
### Table A6. Summaries of Studies A21–A23.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Source</th>
<th>Operational vs. Lab</th>
<th>Sample Size</th>
<th>Constructs Assessed</th>
<th>Test Type/Method</th>
<th>Timed vs. Untimed</th>
</tr>
</thead>
</table>
| 21. Rossini (2016) | SIOP conference         | Operational         | 5 years of data (2012-2016) across 3 job levels  
- entry/support n = 83,196; individual contributor n = 79,798; leader n = 10,347  
- did not report device-type | Did not distinguish constructs from methods  
→ "mix of SJT, dispositional, and biodata" | NR               | NR                |
| 22. Rupayana and Hedricks (2013) | SIOP conference         | Operational         | 3M references for 1M candidates  
- mobile = 2.6% in 2010; 9.2% in 2012 | Use of mobile devices to submit references | NR               | NR                |
Mechanical aptitude n = 396 – 941; mobile n = 23% - 30%  
GMA n = 1433 - 6151; mobile n = 5% - 6%  
Work attitudes n = 1018; mobile n = 39% | Cognitive and noncognitive  
Personality & work attitudes | NR               | NR                |

**Measurement Equivalence**

<table>
<thead>
<tr>
<th>Authors</th>
<th>Measurement Equivalence</th>
<th>Mean Differences</th>
<th>Criterion-Related and Differential Validity</th>
<th>Test-Taker Reactions</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>21. NR</td>
<td>Yes</td>
<td>Personal</td>
<td>NR</td>
<td>Between-subjects</td>
<td></td>
</tr>
</tbody>
</table>
| 22. NR  | GMA = yes; d = 0.46 & 0.35  
Mechanical apt = yes; d = 0.93 & 0.26  
Personality = no; work attitudes = no | NR               | NR                                          | "mobile-accessible rather than mobile-optimized" |
| 23. NR  | GMA = yes; d = 0.46 & 0.35  
Mechanical apt = yes; d = 0.93 & 0.26  
Personality = no; work attitudes = no | NR               | NR                                          | Referees who used a mobile device to do so submitted references over a shorter period of time |

**Note.** SIOP = Society for Industrial and Organizational Psychology; SJT = situational judgment test; NR = not reported; GMA = general mental ability; MC = multiple-choice text/exam; AI = adverse impact; AIRs = adverse impact ratios.
## Appendix B

### Summary of Empirical Comparative Unproctored Internet-Based Testing Device-Type Studies—Proctored Assessments Only

<table>
<thead>
<tr>
<th>Authors</th>
<th>Source</th>
<th>Operational vs. Lab</th>
<th>Sample Size</th>
<th>Constructs Assessed</th>
<th>Test Type/Method</th>
<th>Timed vs. Untimed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gutierrez and Meyer (2013)</td>
<td>SIOP conference</td>
<td>Lab</td>
<td>N = 282</td>
<td>Noncognitive and cognitive * Personality, SJT, &amp; cognitive test* - repeated measures 3 weeks apart</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Huff (2015)</td>
<td>Computers in Human Behavior</td>
<td>Lab</td>
<td>N = 47; repeated measures (no time separation between conditions), desktops and participants' personal smartphones, counterbalanced</td>
<td>Noncognitive (i.e., FFM)</td>
<td>Likert scale</td>
<td>NR</td>
</tr>
<tr>
<td>King et al. (2015)</td>
<td>International Journal of Selection and Assessment</td>
<td>Lab</td>
<td>N = 253; within-subjects design with 3-week interval; randomly assigned to device type at Time 1</td>
<td>Cognitive ability(^1) Customer service orientation(^2) Supervisory SJT(^3)</td>
<td>MC(^1) True/False(^2) “Time pressure was removed”</td>
<td></td>
</tr>
<tr>
<td>Smeltzer (2013)</td>
<td>Thesis</td>
<td>Lab</td>
<td>N = 80</td>
<td>Noncognitive Personality &amp; biodata * repeated measures 3 weeks apart</td>
<td>NR</td>
<td>NR</td>
</tr>
</tbody>
</table>

### Measurement Equivalence

<table>
<thead>
<tr>
<th>Measurement Equivalence</th>
<th>Mean Differences</th>
<th>Criterion-Related and Differential Validity</th>
<th>Test-Taker Reactions</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. NR</td>
<td>NR</td>
<td>NR</td>
<td>Within-subjects design - mobile = less fair, more difficult, less comfortable, no more likely to apply, not better place to work; preference for PC</td>
<td>Summary = negative reactions to mobile &amp; no positives/advantages over nonmobile</td>
</tr>
<tr>
<td>2. NR; but higher coefficient alphas for smartphones for all FFM scores.</td>
<td>NR</td>
<td>NR</td>
<td>Within-subjects design - smartphone = more submission errors; longer completion time; lower usability ratings.</td>
<td>“website to complete the assessment was not optimized for completion on a mobile device”</td>
</tr>
<tr>
<td>3. Yes (SJT); no (GMA); no MEI analyses for customer service orientation</td>
<td>Yes for GMA, (d = 0.16) No for customer service orientation, (d = 0.03), and supervisory SJT, (d = -0.02)</td>
<td>NR</td>
<td>Test ease, and chance to perform = more positive results when tested on PC vs. mobile device</td>
<td>Summary = negative reactions to mobile &amp; no positives/advantages over nonmobile</td>
</tr>
<tr>
<td>4. NR</td>
<td>No</td>
<td>NR</td>
<td>Within-subjects design Mobile = interfered with opportunity to perform; did not improve perceptions of organization; did not think organization would view taking on mobile negatively or affect hiring decision; preference for PC</td>
<td>Summary = negative reactions to mobile &amp; no positives/advantages over nonmobile</td>
</tr>
</tbody>
</table>

Note. Matching superscripts specify which constructs were assessed using what test type/method. SIOP = Society for Industrial and Organizational Psychology; SJT = situational judgment test; NR = not reported; FFM = five-factor model of personality dimensions; MC = multiple-choice text/exam; PC = personal computer; GMA = general mental ability; MEI = measurement equivalence/invariance.