Personnel Assessment and Decisions

Volume 5
Issue 3 *Advanced Technologies in Assessment: A Science–Practice Concern*

Article 4

2019

Developing Device-Equivalent and Effective Measures of Complex Thinking with an Information Processing Framework and Mobile First Design Principles

Darrin M. Grelle
SHL, darrin.grelle@shl.com

Sara L. Gutierrez
SHL, sara.gutierrez@shl.com

Developing Device-Equivalent and Effective Measures of Complex Thinking with an Information Processing Framework and Mobile First Design Principles

Darrin M. Grelle[1] and Sara L. Gutierrez[1]

1. SHL

KEYWORDS

online assessment, mobile testing, mobile first design, equivalence, mobile/non-mobile comparison, unproctored testing, cognitive assessment

ABSTRACT

Organizations are increasingly offering pre-employment assessments on mobile devices to evaluate candidates. The aim of this study is to investigate whether employing a mobile first responsive web design based on an information processing framework will result in device-equivalent measures of cognitive ability. Tests of numerical and deductive reasoning composed of interactive item types were tested for measurement equivalence across device types. Hypotheses were tested using data collected from paid participants over 3 weeks in 2018. Participants completed the test on both a PC and a mobile device. Paired samples t-tests indicated no mean differences in scores or number of items completed across device type. Additional analyses indicated that these item types demonstrated configural invariance and at least partial metric and scalar invariance across device types. The results of this study provide strong support that mobile first design can yield a valid and reliable test that can be used on any device and for any job level.

Organizations are increasingly offering pre-employment assessments on mobile devices (portable computing devices such as smartphones or tablet computers) to screen candidates. The most recent Global Assessment Trends Report indicates that 15% of organizations currently use mobile assessments compared to only 4% in 2014 (Kantrowitz, Tuzinski, & Raines, 2018). The increased interest in engaging mobile delivered assessment is likely driven by key challenges and issues facing organizations across the globe. In a candidate-centric market (Sullivan, 2014), organizations desire to use the recruitment and hiring process as a means to attract top talent. One way to accomplish this is to offer an efficient and accessible candidate experience through the use of new technologies, such as interactive elements and mobile-delivered assessment.

Technology continues to play a significant role in the way industrial-organizational (I-O) psychologists design valid and reliable assessments (Stone, Deadrick, Lukaszewski, & Johnson, 2015). In addition to basing decisions for which assessments to use on validity, reliability, and traditional psychometric requirements, many organizations now consider nonpsychometric attributes, such as whether assessments are available and/or are optimized for mobile administration, and how engaging, innovative, and "good for the brand" the test experience might be (Sullivan, 2014). To this end, a shift has occurred in assessment science whereby the role of the technology is as important as the psychometrics in order to meet the evolving needs of assessment consumers (Arthur, Doverspike, Kinney, & O'Connell, 2017; Morelli, Potosky, Arthur, & Tippins, 2017).

The intersection between testing and technology brings opportunities and challenges for assessment design, particularly for mobile assessments. New scoring models and alternative item types are being considered, which has tremendous potential to improve the measurement efficiency and accuracy associated with assessment. When developing a mobile assessment, one must consider the candidate expe-

Corresponding author:
Darrin M. Grelle
Email: darrin.grelle@shl.com

rience and the associated psychometric challenges, most notably the potential for score differences that may arise from candidates completing assessments on a range of small and large screen devices, including smart phone, tablet, and computer. The ideal mobile assessment will therefore balance creating a positive and engaging candidate experience with demonstrating measurement equivalence across device types.

Given the speed at which technology is changing, limited guidance exists within the I-O psychology literature to direct and guide practitioners who are in the trenches of mobile assessment design. The purpose of the current article is to supplement the little existing guidance with an applied example of ground-up mobile assessment development that employs psychometric rigor while also satisfying consumers' increasing demands for engaging mobile assessment. We describe the design criteria that utilizes mobile first responsive web design (Marcotte, 2010; Ward, 2017) and follows design principles anchored by theory in industrial-organizational and cognitive psychology to create two innovative and engaging computer adaptive measures of cognitive ability.

Equivalence of Mobile to PC-Based Testing

Past research has consistently demonstrated that non-cognitive assessments such as personality, biodata, and situational judgment show little evidence of score degradation for tests completed on mobile devices (Arthur, Doverspike, Munoz, Taylor, & Carr, 2014; Illingworth, Morelli, Scott, & Boyd, 2015; Lawrence, Wasko, Delgado, Kinney, & Wolf, 2013; Morelli, Mahan, & Illingworth, 2014). In contrast, the majority of initial research investigating the measurement equivalence for cognitive testing does show a decrement in scores for those completing on mobile device as compared to a PC (Arthur et al., 2014; Impelman, 2013; King, Ryan, Kantrowitz, & Grelle, 2014; LaPort, Huynh, Stemer, Ryer, & Moretti, 2016). That said, cognitive ability tests continue to be one of the most valid predictors of job performance (Schmidt & Hunter, 2004) and remain one of the most commonly used assessment types (Kantrowitz et al., 2018).

Thus, attention has turned to methods of designing cognitive assessments that mitigate device related differences through mobile-optimized and mobile first responsive web design to drive measurement equivalence (Boyce & Gutierrez, 2018). Rather than a shrunken down version of what is displayed on a larger screen, mobile first responsive web design starts with the smallest supported device and works up to larger devices to provide the user an experience that is optimized for and consistent across all device types (Ward, 2017). The concern with simply displaying traditional cognitive ability measures on smaller screen devices is that construct-irrelevant variance is introduced into the test (Arthur, Keiser, & Doverspike, 2017). Mobile first responsive web design features can include single column layouts, simple navigation, large graphics, reduced text, no need to type, and uncluttered design (Lyerly, n.d.). Indeed, mobile-optimized cognitive ability assessments have been shown to be equivalent across device types, specifically for measures of working memory (Frost, Carpenter, & Ferrell, 2018; Morgan, LaPort, Lowery, Cottrell, Rangel, Martin, & Boyce, 2018) or general entry-level cognitive tests (Gutierrez & Grelle, 2018).

The mobile-optimized design requirements of reduced text and streamlined presentation of stimuli have posed a dilemma for the development of a robust measure of cognitive ability. Traditionally, items with substantial amounts of text and elaborate infographics have often been required to assess more complex, higher order thinking processes required of individuals in mid to high level jobs. For instance, assessments of deductive reasoning or reading comprehension commonly present a paragraph of text or information in tables or graphs, and candidates are required to read or review this information in order to answer a question. Previous research has clearly demonstrated that traditional items presented on mobile devices show score decrements. With limited screen sizes on mobile devices, traditional cognitive items are not tenable. For this reason, many of the mobile-equivalent cognitive measures developed to date tend to measure only lower level abilities such as memory, working memory, and compare/contrast tasks where only simple item stimuli are needed. Although these types of assessments may be predictive for entry-level roles, they will be less relevant for professional-level roles and above where the job demands higher levels of problem solving ability. For these roles, alternative item types that present test stimuli in new and unique ways are needed that allow for the assessment of complex abilities.

Technology Enhanced Test Design Principles

When developing cognitive tests that may be taken on a mobile device, there are several factors that need to be considered in order to ensure measurement equivalence and construct validity. The structural characteristic/information processing model (SCIP; Arthur et al., 2017) provided a useful and theoretically grounded framework on which to make assessment design decisions that attempted to mitigate or eliminate construct-irrelevant variance. This model discusses four structural characteristics of computers and mobile devices that can yield score differences and introduce construct irrelevant variance.

Screen Size

Screen size can have significant impact on the test taking experience. According to the SCIP framework, if any piece of a question does not fit on the screen of smaller devices that does fit on larger devices, then working mem-

ory is introduced into the test (Arthur, Keiser, Hagen, & Traylor, 2018). This is because the candidate must retain the pieces of the question they cannot see in their working memory while working out the rest of the question. Responsive web design is ideally suited to help with this problem; by adjusting how information appears on screen based on screen size, one can ensure that the amount of information that appears on screen is consistent across devices. It is also extremely important to write content that can fit on smaller screens.

Screen Clutter

Screen clutter is defined as the amount of text, images, and other objects on screen. Arthur et al. (2018) propose that as screen clutter increases, there is an increase in the visual acuity and perceptual speed demands on the candidate. This aspect of the framework was not empirically evaluated, but having too much information and/or fine detail on screen is clearly problematic. In order to create question types that optimized the limited screen space on small mobile devices while keeping screen clutter to a minimum, the cognitive ability test content experts worked extensively with a creative agency that specializes in building consumer-grade mobile applications using mobile first responsive web design. Using both the SCIP framework and responsive web design heuristics adapted from Gomez, Caballero, and Sevillano (2014), the following design principles were followed to make the best use of limited screen space:

- Use minimal text
- Utilize graphics to convey information wherever possible
- Eliminate the need for traditional multiple-choice response options given screen size limitations
- Eliminate any need for horizontal scrolling– all content must fit within the width of the screen
- Minimize vertical scrolling to every extent possible
- Where vertical scrolling must exist on smaller screens, ensure it is also required on larger screens

Response Interface

Every question within an assessment requires some way for the candidate to provide a response. The input can be as simple as the traditional multiple-choice response format or as complex as free-text entry via a keyboard. The complexity of the method of response entry will have an impact on the degree to which candidates must draw upon their psychomotor abilities to input their responses (Pais, 2018). Although simple multiple-choice selection is likely the method least demanding in psychomotor ability, providing candidates with a list of options from which to choose takes up valuable screen space. The questions in the current study were designed so that candidates enter their responses directly into the question using drag/drop/tap functionality. This includes behaviors like changing the size of a wedge on a pie chart, adding tasks to a daily planner, or selecting dates on a calendar. The following design principles were followed to ensure that psychomotor demands were kept to a minimum:

- Ensure input mechanisms (tap, drag, slide, rotate) could as easily be conducted on smaller screens as they could be on larger screens to minimize user error with working on smaller screens
- Ensure input mechanisms (tap, drag, slide, rotate) could as easily be conducted touch screen devices as they could be with a mouse or touchpad
- Ensure that every question had multiple ways to input responses to avoid the "fat finger problem" (Pais, 2018)
- Provide detailed instructions and guided practice questions to ensure candidates are familiar with each question type

Permissibility

Permissibility refers to the freedom candidates have to take assessments in a setting of their choosing. A candidate with a mobile device has more freedom to complete an assessment virtually anywhere they choose as compared to someone completing the assessment on a desktop computer. Candidates are freer to choose settings where distractions are prevalent when completing a test on a mobile device. Distractions increase the selective attention demands on the candidate (Lavie, 2005). From a design perspective, this structural characteristic is the most challenging of the four to remedy via test design. There are no design principles that can directly influence a candidate's decision to take a test in an appropriate venue. For the tests designed in the current study, the test instructions strongly urge candidates to find a place free from distractions and to turn off phone notifications if using a mobile device. Candidates are permitted to exit the test at any time and return where they left off for a limited number of times (for content security reasons), and they can switch to a different device if they choose. Also, it is our hope that the fun, engaging nature of the test will encourage candidates to make choices that set themselves up for success when taking the time to complete the test, such as finding a quiet location to test.

An ancillary goal of utilizing mobile-optimized design, beyond driving equivalence across devices, was to ensure a positive candidate experience for those completing the tests on smaller screen devices. A summary of the literature's findings for test-takers' reactions and preferences for completing tests on mobile devices as compared to non-mobile devices provided by Arthur et al. (2017) indicated that in most cases, test takers' reactions were less positive when completing a test on a mobile device. It should be noted that very few studies included in the review utilized tests designed to be mobile optimized. Applying design princi-

ples during test construction with the express purpose of creating an equivalent experience regardless of device type is expected to enhance test takers' satisfaction when completing these tests on a mobile device.

Summary

The principal goal of the test design process was to create an engaging, job-relevant test of cognitive ability that has measurement equivalence across devices. Due to the move away from traditional multiple-choice entry to alternate response entry formats, an alternative scoring method was required. Whereas traditional cognitive assessments typically utilize dichotomously scored items, partial credit scoring is more suited to item types with multiple response entry points. In these tests, a candidate is asked to solve multiple problems within a single question. As such, the assumption of local independence within the item is violated, which renders the use of a three parameter logistic item response theory (IRT) model inappropriate. The response capture design for these cognitive tests allows us to apply a partial credit model of scoring, which generates more information about a candidate while utilizing fewer items. Therefore, an added benefit of the mobile first multiple-data-point item type is added precision with less candidate time required.

The Current Study

Recent research has shown promise regarding the potential for equivalent design of cognitive ability tests through the use of mobile-optimized design principles (Brown & Grossenbacher, 2017, Castillo & Doe, 2017; Frost et al., 2018; Gutierrez & Grelle, 2018; Morgan et al., 2018). Unfortunately, few cognitive ability measures discussed in the mobile equivalence literature to date measure complex critical thinking skills that would be appropriate for reliably assessing the cognitive ability of individuals applying for mid to high level job roles. Additionally, many existing measures do not present item content that is job relevant or face valid.

The current study aims to examine the measurement equivalence and efficacy of two newly designed interactive and mobile optimized tests of cognitive ability. These tests are not serious games but do not consist of simple dichotomous right/wrong questions, either. Instead, these tests utilize modern input mechanisms (e.g., tap, rotate, drag) to build interactive and engaging, work-relevant scenarios that can be utilized to measure cognitive ability for all job levels. As the line between what is considered a "smartphone," a "tablet," and a PC becomes blurrier with every new device added to the market, we chose to conduct this study by comparing the largest and smallest devices in the range and did not consider tablets. We conducted a review of the most commonly used devices on the market and classified them as smartphones, tablets, and PCs by screen width in pixels (as a unit of measure, not actual screen resolution – 1px = 1/96th of one inch). Devices that fall in the "tablet" category by pixel width were not included in this study. Given the careful consideration to the optimized design of each item type within these two tests, we hypothesize:

Hypothesis 1: Reliability: For each test, reliability will be sufficient and consistent across the spectrum of ability.

Hypothesis 2: Test Performance: For each test, no mean score differences will be found between those completing the test on mobile devices (screens smaller than 768 pixels in width) and those completing on personal computers (PCs; screens larger than 992 pixels in width).

Hypothesis 3: Measurement Invariance: For each test, invariance across device types will be supported, indicating the tests are measuring the same construct regardless of device type.

Hypothesis 4a: Test Time: For each test, the ability to complete all items within the allotted time will not be impacted based on device type utilized.

Hypothesis 4b: Test Time: For each test, the average time spent completing the test will not be impacted based on device type utilized.

Although the current measures of cognitive ability utilize new design attributes, item types, and scoring system, they were designed to measure the same constructs traditionally found to be predictive of performance in the workplace: deductive reasoning and numerical reasoning.

Hypothesis 5: The new measures of deductive and numerical reasoning will strongly correlate to established measures of their respective construct.

METHOD

Materials

Interactive assessment. The two interactive assessments included in this study contain different question types that simulate work relevant activities. The deductive reasoning test includes items where candidates are provided with a set of rules or restrictions and then asked to place employee avatars into different offices on a map, schedule tasks on their daily planner, select dates on a calendar, or

rank employees. The numerical reasoning test provides the candidate with some numerical information that must be computed and then entered into pie charts, spreadsheets, bar graphs, or line graphs. There are nine distinct interactions/question types across the two tests, and candidates provide their responses by interacting with the map, calendar, and so forth using drag, drop, and tap/click features. The deductive reasoning test has 12 questions and the numerical reasoning test has 10. Both tests have an 18 minute timer. Both tests also provide detailed instructions on how to complete the test that includes a video demonstrating how the different question types should be completed. Though the question types were designed to be intuitive, the video helps control for any differences in familiarity with the different response entry formats.

The operational version of the interactive assessments is computer adaptive and has large item banks with questions spanning the full spectrum of difficulty, with more targeted at medium difficulty where most people fall on the ability distribution. In order to conduct measurement equivalence and invariance analyses, however, all participants in the analysis have to have seen the same questions. To facilitate this analysis and to ensure there were no practice effects due to participants seeing the same set of questions twice, two equivalent forms of each test were constructed. Form 1 of the test was always administered in Part 1 and Form 2 was always administered in Part 2. Therefore, measurement equivalence analyses were conducted using independent samples at both time points, and mean differences were compared using paired samples that completed both parts. Questions were selected from the middle range of difficulty and equated using the IRT parameters such that the range and mean of the discrimination parameters and difficulty parameters of each test form were equivalent. Each test form also contained equal numbers of the different item types each test contained.

Traditional cognitive measure. The two "traditional" measures of cognitive ability are operational tests used for selection. They are administered online and usually completed on a PC. Both tests are computer adaptive with five-option multiple choice questions scored dichotomously using the 3PL model. The deductive reasoning assessment is 18 questions with a 20 minute time limit, and the numerical reasoning assessment is 16 questions with a 20 minute time limit. Both tests include detailed written instructions on how to complete the assessment.

Procedure

The deductive and numerical reasoning studies were conducted separately with unique participant samples, but the method for each study was the same. Participants were randomly assigned to one of two conditions. In the first condition, participants were asked to complete the new interactive assessment on a desktop or laptop computer. The interactive assessment was followed by the traditional measure of the same construct, which was also completed on a desktop or laptop computer. Twenty-four hours after completing the first part of the study, participants received a link to the second part of the study and had two days to complete it. In the second part of the study, participants were asked to complete the interactive assessment again but this time on a smartphone. They were specifically instructed not to use a tablet device to maximize the difference in device type under study. Participants assigned to the second condition completed the two parts of the study in the reverse order to control for order and practice effects. The test design and sample sizes are outlined in Figure 1. Participants were asked in both parts of the study to complete a three-question survey about the device they were using to complete the assessments. Our system also captured the screen size of the device used as a validation of the participants' answers. If participants indicated that they used a device different from what they were instructed to use, they were excluded from the analyses.

Participants

Separate samples were recruited for the numerical reasoning and deductive reasoning studies. Participants were recruited from a data crowdsourcing pool and were paid for their participation. Each test was administered in two parts separated by 24 to 48 hours. One part is the mobile administration of the test and second is the PC administration. We sought 300 participants who completed both parts, so we recruited roughly twice that amount for the first part of the study with the assumption that about half of the candidates would not return. For deductive reasoning, 593 participants completed Part 1 of the study, and for numerical reasoning 551 completed Part 1. Table 1 contains the demographic details for the 228 deductive reasoning participants and 215 numerical reasoning participants that completed both parts of the study and provided usable data (participants were excluded from all analyses if they did not follow the in-

FIGURE 1.

Deductive reasoning			
Condition 1	N	*Condition 2*	N
PC first	301	Mobile first	292
Mobile second	116	PC second	111
Numerical reasoning			
Condition 1	N	*Condition 2*	N
PC first	254	Mobile first	297
Mobile second	115	PC second	100

TABLE 1.
Participant Characteristics

	Numerical reasoning		Deductive reasoning	
Demographic category	N	%	N	%
Male	49	22.8	39	17.2
Female	164	76.3	186	81.9
Prefer not to answer	2	0.9	2	0.9
American Indian or Alaska Native	2	0.9	2	0.9
Asian	8	3.7	7	3.1
Black or African American	15	7.0	26	11.5
Hispanic or Latino	17	7.9	14	6.2
Two or more races	3	1.4	2	0.9
White	168	78.1	176	77.5
Prefer not to answer	2	0.9	0	0.0

structions described in the Procedure section or spent fewer than 30 seconds per question). The entire sample indicated that they currently reside in the United States, with 98% reporting English as their first language. Our sample was predominantly female; however, we have no reason to believe that this will make our results less generalizable. As women have been found to be more likely than men to complete selection assessments using a mobile device (Golubovich & Boyce, 2013), this sample may actually be more representative of the population that would take assessments on a mobile device.

RESULTS

Descriptive statistics for all of the tests included in the study were calculated and are included in Table 2. The test scores are raw theta scores, which generally range from -3.0 to 3.0 with a mean of about 0.0 in the general population. Participants who demonstrated low effort based on time spent per question were removed from all analyses. The low scores on the assessments most likely reflect a lower ability sample. The standard error estimates are low (mean standard error estimates ranging from 0.34 to 0.40), indicating that participant ability was still accurately measured. Table 2 also includes the descriptive statistics for the standard error of the interactive assessments and the 10th and 90th percentile standard error estimates. The standard error can be converted to a standard reliability estimate with the formula (Embretson & Reise, 2000):

$$r_{xx} = 1 - SE^2 \quad (1)$$

Using this formula we see that for most candidates, the reliability estimate ranges from 0.74 to 0.95, with an average reliability between 0.85 and 0.88, which is well within the accepted range of reliability for a test (Nunnally, 1978), thus supporting Hypothesis 1. It should be noted that the forms used in this study are static. The operational versions of these tests are computer adaptive, which yield lower standard errors because the items are tailored to a candidate's ability level (Weiss, 2011).

In both studies, Form 2 was always presented after Form 1, and though the forms were equated for difficulty based on the IRT parameters, scores improved between the first and second sitting of the assessment. When looking only at candidates that completed both parts of the study, we see significant improvement in scores. A paired sample *t*-test was conducted for deductive reasoning, $t(275) = -6.50$, $p < .01$ and numerical reasoning, $t(261) = -5.57$, $p < .01$. (Note: the samples are slightly larger for these comparisons because they include participants who completed the tests on the same device for both parts of the study. These participants were removed from all subsequent analyses.)

Hypothesis 2 was tested via a paired sample *t*-test comparing the participant scores on the mobile sitting of the test to the PC version of the test. Results indicated that for both deductive reasoning, $t(226) = -1.02$, $p = 0.31$, and numerical reasoning, $t(214) = 0.05$, $p = 0.96$, scores did not significantly differ between the mobile and PC administration of the interactive tests, supporting Hypothesis 2. The correlations between scores on both administrations were significant and of a large enough magnitude to indicate that the same construct is being assessed across both device

TABLE 2.
Descriptive Statistics

	N	*M*	*SD*	Mean SE	*SD* of SE	10th	90th
Deductive Reasoning Form 1	593	-0.53	0.74	0.37	0.11	0.28	0.51
Deductive Reasoning Form 2	227	-0.12	0.73	0.40	0.08	0.27	0.46
Standard Deductive Test	391	-0.77	0.76				
Numerical Reasoning Form 1	551	-0.83	0.62	0.36	0.11	0.23	0.49
Numerical Reasoning Form 2	215	-0.41	0.71	0.32	0.10	0.23	0.49
Standard Numerical Test	385	-1.12	0.91				

types. For deductive reasoning, the correlation was $r(227) = 0.67$, $p < .001$. The correlation for numerical reasoning was $r(215) = 0.77$, $p < .001$. When we controlled for the order in which the assessments were taken to account for the practice effect using regression, we found partial correlation of $r(227) = 0.71$, $p < .001$ for deductive reasoning and $r(215) = 0.80$, $p < .001$ for numerical reasoning.

Though we found support for Hypothesis 2, it is still important to ensure that the questions are functioning equivalently across device types. Mplus 8.1 (Muthen & Muthen, 2018) was used to test the four test forms used in the study for configural, metric, and scalar invariance. For a detailed review of the methods for testing measurement equivalence and invariance (MEI), please see Vandenberg and Lance, 2000. Participants completed two different forms of the interactive tests, and MEI analyses require that the entire sample completes the same set of questions, so the MEI analyses conducted compared independent samples. The item scores were analyzed as ordered categorical data using the weighted least square mean and variance adjusted (WLSMV) estimator. Because the chi-square generated for a model using this estimate is mean and variance adjusted, one must use the DIFFTEST function in Mplus to perform chi-square difference tests for nested models. This is why the values in the chi-square difference columns may seem out of alignment with the chi-square values reported for the overall model fit. All of the models tested had all questions loading on a single factor; however, in the interest of evaluating equally feasible alternative models (Vandenberg & Grelle, 2009), a multifactor model was tested where questions were loaded on different factors based on the different item types included in each test form. The single factor model had better fit in all four cases. However, modification indices suggested that the residual variance for questions of the same type were often correlated, leading to less than perfect fit. In order to test the most parsimonious model, though, the residuals were left uncorrelated. The analyses are summarized in Tables 3–6.

The results demonstrate that for both deductive reasoning forms, configural, metric, and scalar invariance held supporting Hypothesis 3. The test for residual invariance failed. For Form 1 of numerical reasoning, configural and metric invariance held. Form 1 failed the test of scalar invariance, however, unless the item thresholds for the 10th item were freely estimated across the mobile and PC samples. As full scalar invariance did not hold, the test of residual invariance was not conducted. Form 2 of numerical reasoning met configural invariance; however, it failed the test of metric invariance unless the factor loadings for the 3rd and 4th items were free to vary across samples. Though only partial metric invariance held, the test of scalar invariance was conducted. The test failed unless the thresholds of the 3rd and 4th items were also allowed to vary across samples as well as the thresholds for the 1st item. The test of residual invariance was not conducted for this form. These results indicate partial support for Hypothesis 3.

Next, we looked at differences in the number of items completed by device type to determine if participants found it easier to complete more items on one device type versus another. Descriptive statistics for completion times and number if items completed are included in Table 7. A *t*-test of the number of items completed show that there is no difference for deductive reasoning, $t(227) = -0.31$, $p = 0.80$, or numerical reasoning $t(214) = -0.90$, $p = 0.37$, supporting Hypothesis 4a. We also looked at the average time spent completing the test and found only partial support for Hypothesis 4b. A *t*-test of the time to complete the test in minutes showed no difference for numerical reasoning $t(214) = -0.13$, $p = 0.89$, but did show a difference for deductive reasoning $t(227) = 5.09$, $p < 0.001$, with participants spending about 1 minute longer on the mobile version of the assessment.

Finally, in order to determine whether our interactive tests are measuring the constructs of interest, we correlated scores between the interactive version of the test and the standard five-option multiple choice version. The correlations were calculated between the PC version of the interactive test and the standard version because those two

TABLE 3.
Deductive Form 1

Model	Chi square	*df*	RMSEA	Delta chi	Delta *df*	*p*-value
PC	239.1	54	0.107			
MOBILE	177.54	54	0.089			
Omnibus model	408.65	54	0.105			
Baseline group model	419.71	108	0.099			
Invariant factor loadings	415.95	118	0.092	15.11	10	0.128
Invariant factor loadings and thresholds	429.85	139	0.084	31.11	21	0.072
Invariant factor loadings, thresholds, and residual variance	427.51	151	0.079	28.45	12	0.005

TABLE 4.
Deductive Form 2

Model	Chi square	*df*	RMSEA	Delta chi	Delta *df*	*p*-value
PC	72.8	54	0.046			
MOBILE	89.49	54	0.070			
Omnibus model	88.45	54	0.046			
Baseline group model	163.81	108	0.059			
Invariant factor loadings	169.76	118	0.054	10.42	10	0.404
Invariant factor loadings and thresholds	190.45	136	0.052	23.54	18	0.171
Invariant factor loadings, thresholds, and residual variance	221.54	148	0.058	30.12	12	0.003

TABLE 5.
Numerical Form 1

Model	Chi square	*df*	RMSEA	Delta chi	Delta *df*	*p*-value
PC	65.78	35	0.062			
MOBILE	70.84	35	0.063			
Omnibus model	103.06	35	0.063			
Baseline group model	136.73	70	0.062			
Invariant factor loadings	137.57	79	0.055	6.09	9	0.731
Invariant factor loadings and thresholds	175.8	108	0.05	46.08	29	0.023
Invariant factor loadings, thresholds (except item 10)*	158.79	105	0.046	30.12	26	0.263

*This model is compared to the invariant factor loadings model allowing the thresholds for item 10 to vary across devices.

TABLE 6.
Numerical Form 2

Model	Chi square	*df*	RMSEA	Delta chi	Delta *df*	*p*-value
PC	64.47	35	0.075			
MOBILE	63.93	35	0.077			
Omnibus model	90.01	35	0.074			
Baseline group model	128.12	70	0.076			
Invariant factor loadings	147.89	79	0.078	20.41	9	0.016
Invariant factor loadings (except items 3 & 4)*	135.08	76	0.05	9.15	6	0.165
Invariant factor loadings, thresholds (except items 1,3, & 4)**	157.25	95	0.067	28.88	19	0.068

*This model is compared to the baseline group model allowing the factor loadings for Items 3 and 4 to vary across devices.
**This model is compared to the adjusted invariant factor loadings model allowing the thresholds for Items 1, 3, and 4 to vary across devices.

TABLE 7.
Timer Statistics*

	Avg. # items completed	Time taken in minutes
Deductive reasoning - Mobile	11.7	13.6
Deductive reasoning - PC	11.8	12.5
Numerical reasoning - Mobile	9.0	15.0
Numerical reasoning - PC	9.1	15.0

*Includes only cases that completed both parts of the study.

tests were administered in the same sitting and on the same device. The correlations for deductive reasoning, $r(361) = 0.66$, $p < .001$, and numerical reasoning, $r(331) = 0.70$, $p < .001$, demonstrate a strong relationship between the two versions of each test, supporting Hypothesis 5.

DISCUSSION

The results of this study provide strong support for our hypotheses that using construct-oriented mobile first design can yield a valid and reliable test that can be used on any device and for any job level. When developing these tests, our goal was not only to create tests that showed measurement equivalence across devices but to also measure specific cognitive constructs that research has shown to be related to job performance using item types that candidates will find more engaging.

Hypothesis 1 was well supported by the data. There were no mean differences found between scores on each device. The use of a paired sample design further supports this hypothesis because participants had to take the assessment on both device types. In independent samples, there is always the chance that sampling error accounts for differences (or lack thereof), especially when participants are not randomly assigned to different conditions. We also found a strong correlation between scores on both devices after controlling for the order effect, indicating that the test is measuring the same construct across devices. Deviation from a perfect correlation between the two may be due to individual differences in comfort levels with different device types. In a real world setting, candidates would choose which device they felt most comfortable using.

Though we found no mean differences between scores on the two device types, it is important to ensure that the questions are functioning in the same way across devices. If the questions do not operate in the same way across devices, then mean comparisons can be rendered invalid. Generally, configural, metric, and scalar invariance is all that is required to make meaningful comparisons (Vandenberg & Lance, 2000). Deductive reasoning met all three but lacked residual invariance. Lack of residual invariance indicates differences in the amount of error in measurement for each item. As there are many possible sources of error in measurement, it is difficult to know exactly why error variance varies across devices, but it could be due to different levels of distraction across device types and the ability to ignore those distractions. What is important is that the first three tests of measurement invariance demonstrate that items have equivalent difficulties and ability to discriminate between high and low performers across devices. For numerical reasoning, the tests were not as clear. Both test forms showed configural invariance supporting a single factor structure. Only partial metric and scalar invariance was supported however. That said, full invariance is unlikely in practice due to the sensitivity of the statistical tests and sampling error (Byrne, Shavelson, & Muthén, 1989). Most of the items in each test met full invariance, supporting Hypothesis 2. We examined the content of the items that did not demonstrate full invariance looking for characteristics that might cause issues across devices like scrolling, specific response functionality, and presence of images. We did not find any features unique to these items that would lead to a lack of invariance.

The reliability analysis demonstrated that, even using a fixed form of the test, participants could be reliably measured across the ability distribution. Using interactive questions with multiple data points per question, we are able to get more information about the candidate with fewer questions: The average standard error of our standard adaptive tests is similar to what we see with the interactive tests; however, the standard adaptive tests typically require 16-22 items to achieve this compared to 10-12 items in the interactive tests. More information leads to a more reliable assessment. We anticipate that because the questions in the adaptive versions of the interactive tests cover a range of difficulty, and the measurement model used takes advantage of the multiple data points per question, we should find high reliability in measurement across a wide range of scores. This would mean that the test can be used to accurately distinguish between high and low ability candidates at all job levels.

A concern when developing mobile-enabled tests is that certain aspects of the test might take more time on one device versus another. The interface used to evaluate the information presented in the question might take more time because one device might require scrolling in order to see all of the information. It also might take more time to operate using a touch screen versus using a mouse when entering responses. We did find that candidates were spending about a minute longer on a mobile when completing the deductive reasoning test; however, in support of Hypothesis 4a, we found that candidates are able to complete the same number of items in the time allowed regardless of which device was used. We did not see any differences in test time for numerical reasoning, so there is no concern with candidates using one device type having less time to complete the full test because the device is causing item responses to take longer. We will be collecting additional data and conducting item level analyses of the deductive reasoning content to determine if there are any adjustments that can be made to the content to reduce differences in time to complete.

Many mobile-enabled cognitive tests cited in the existing literature measure very narrow facets of cognitive ability like numerical calculation or working memory. More complex concepts like numerical and deductive reasoning have traditionally taken too much screen space to be suitable for a smaller screen. The convergent validity analysis between the interactive assessments and the traditional multiple choice tests demonstrates that it is possible to accurately measure those constructs on a mobile device.

Limitations and Future Directions

This work documents the successful implementation of innovative, interactive elements within the design and construction of mobile-enabled cognitive assessments. Given the sustained interest in mobile-enabled cognitive testing within the field of personnel selection, there are many opportunities for additional research. To reduce the testing burden on our participants, this study was limited to the two interactive and two traditional tests of cognitive ability. In future research, we hope to include personality and biodata measures to round out the construct validity analysis by providing evidence of discriminant validity. Additionally, because measurement equivalence/invariance using the fixed forms of the tests has been demonstrated, future research will seek to replicate these results using the computer adaptive versions of these assessments. Finally, given the increased focus on ensuring tests created for personnel selection are engaging, innovative, and "good for the brand," future research should investigate the extent to which the addition of mobile first, interactive elements are seen as more interesting and provide an enhanced candidate experience.

REFERENCES

Arthur, W., Jr., Doverspike, D., Kinney, T. B., & O'Connell, M. (2017). The impact of emerging technologies on selection models and research: Mobile devices and gamification as exemplars. In J. L. Farr & N. T. Tippins (Eds.), Handbook of employee selection (2nd ed.). New York, NY: Taylor & Francis/ Psychology Press.

Arthur, W., Jr., Doverspike, D., Muñoz, G. J., Taylor, J. E., & Carr, A. E. (2014). The use of mobile devices in highstakes remotely delivered assessments and testing. International Journal of Selection and Assessment, 22, 113–123.

Arthur, W., Jr., Keiser, N.L., & Doverspike, D. (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, 31, 1-32.

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

Boyce, A. & Gutierrez, S. (2018, April). Mobile first design: The key to effective mobile cognitive testing? Symposium presented at 33rd Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL

Brown, M. I., & Grossenbacher, M.A. (2017). Can you test me now? Equivalence of GMA tests on mobile and non-mobile devices. International Journal of Selection and Assessment, 25(1), 61–71.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. Psychological Bulletin, 105, 456- 466.

Castillo, M. S., & Doe, R. (2017). Mobile and nonmobile assessment in organizations: does proctoring make a difference? Psychology, 8 (6), 878–891.

Embretson, S. E., & Reise, S. P. (2000). Multivariate Applications Books Series. Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum Associates.

Frost, C., Carpenter, J., & Ferrell, J. (2018, April). Demonstrating equivalence of high-fidelity cognitive measures on mobile devices. In A. S. Boyce and S. Gutierrez (Chairs), Mobile first design: The key to effective mobile cognitive testing? Symposium conducted at the 33th annual Society for Industrial and Organizational Psychology Conference, Chicago, IL.

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

Gomez, R. Y., Caballero, D. C., & Sevillano, J. L. (2014). Heuristic evaluation on mobile interfaces: A new checklist. Scientific World Journal, 2014, 1-19.

Gutierrez, S. L. & Grelle, D. (2018, April). Impact of mobile-first design on equivalence for cognitive tests. In S. L. Gutierrez & A. S. Boyce (Chairs), Mobile first design: The key to effective mobile cognitive testing? Symposium presented at the 33rd Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Illingworth, A. J., Morelli, N. A., Scott, J. C., & Boyd, S. L. (2015). Internet-based, unproctored assessments on mobile and non-mobile devices: Usage, measurement equivalence, and outcomes. Journal of Business and Psychology, 30, 325–343.

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

Kantrowitz, T. M., Tuzinski, K. A., & Raines, J. M. (2018). 2018 Global Assessment Trends Report. SHL.

King, D. D., Ryan A. M., Kantrowitz, T., & Grelle, D. (2014, April). MIT versus PCIT: Assessing equivalence, individual differences, and reactions. In T. Kantrowitz & C. M. Reddock (Chairs), Shaping the future of mobile assessment: Research and practice update. Symposium presented at the 29th Annual Conference of the Society for Industrial and Organizational Psychology, Honolulu, HI.

LaPort, K., Huynh, C. T., Stemer, A., Ryer, J. A., & Moretti, D. M. (2016). Mobile assessment: Comparing traditional cognitive, cognitive-reasoning, and non-cognitive performance. In T. D. McGlochlin (Chair), Mobile equivalence: Expanding research across assessment methods, levels, and devices. Symposium conducted at the 31st Annual Conference of the Society for Industrial and Organizational Psychology, Anaheim, CA.

Lavie, N. (2005). Distracted and confused?: Selective attention under load. Trends in Cognitive Sciences, 9(2), 75-82.

Lawrence, A. D., Wasko, L., Delgado, K., Kinney, T. B., & Wolf, D. (2013, April). Understanding the mobile experience: Data across device and industry. Paper presented at the 28th Annual Conference of the Society for Industrial and Organizational Psychology, April, Houston, TX.

Lyerly, T. (n.d.). Mobile-friendly vs. mobile-optimized vs. responsive design for websites. Retrieved from http://torspark.com/mobile-friendly-vs-mobile-optimized-vs-responsive-design/

Marcotte, E. (2010). Responsive web design. A List Apart (306). Retrieved from http://alistapart.com/article/responsive-web-design/

Morelli, N. A., Mahan, R. P., & Illingworth, A. J. (2014). Establishing the measurement equivalence of online selection assessments delivered on mobile versus nonmobile devices. International Journal of Selection and Assessment, 22, 124–138.

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

Morgan, K. E., LaPort, K. A., Lowery, B. S., Cottrell, J. M., Rangel, B., Martin, N. R., & Boyce, A. S. (2018, April). The quest for equivalence: Mobile-first working memory assessment. S.L. Gutierrez and A. S. Boyce (Chairs), Mobile first design: The key to effective mobile cognitive testing? Symposium presented at the 33rd Annual conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Muthén, L. K., & Muthén, B. O. (1998-2017). Mplus User's Guide.

8th Edition. Los Angeles, CA: Muthén & Muthén.
Nunnally, J. C. (1978). Psychometric theory (2nd ed.). New York: McGraw-Hill.
Pais, J. (2018) UX for mobile: The rise of fat-finger design. Outsystems. Retrieved from https://www.outsystems.com/blog/posts/ux-for-mobile-fat-finger-design/
Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. Journal of Personality and Social Psychology, 86(1), 162-173.
Stone, D. L, Deadrick, D. L, Lukaszewski, K. M., & Johnson, R. D. (2015). The influence of technology on the future of human resource management. Human Resource Management Review, 25, 216-231.
Sullivan, J. (2014). The power has shifted to the candidate, so current recruiting practices will stop working. Retrieved from http://www.eremedia.com/ere/the-power-has-shifted-to-the-candidate-so-current-recruiting-practices-will-stop-working/
Vandenberg, R. J., & Grelle, D. (2009). Alternative model specifications in structural equation modeling: Facts, fiction, and truth. In C. E. Lance & R. J. Vandenberg (Eds), Statistical and methodological myths and urban legends: Received doctrine, verity, and fable in the organizational and social sciences (pp. 167-193). New York, NY: Routledge.
Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. Organizational Research Methods, 3(1), 4-69.
Ward, C. (2017). Jump start responsive web design, 2nd edition. Melbourne, Victoria, Australia: SitePoint Pty. Ltd.
Weiss, D. J. (2011). Better data from better measurements using computer adaptive testing. Journal of Methods and Measurement in the Social Sciences, 2, 1-27.

RECEIVED 11/30/18 ACCEPTED 05/20/19