ABSTRACT

PARKER, BRANDY NICOLE. Smartphones in Selection: Exploring Measurement Invariance using Item Response Theory. (Under the direction of Dr. Adam W. Meade.)

The use of mobile devices (e.g., smartphones) by applicants when completing assessments is a growing phenomenon in the area of selection. Like the transition from paper-and-pencil to online testing, research is needed in order to understand whether measurement invariance holds across device types and website formats. The aim of this study was to examine the equivalence of the psychometric properties for two measures used in selection across smartphones and non-mobile devices. Data were collected from 693 Mechanical Turk participants who were randomly assigned to complete a survey using one of the three formats: non-mobile, mobile-friendly, and mobile-optimized. Item response theory was used to explore whether measurement invariance held for both a cognitive ability measure (i.e., Raven's Advanced Progressive Matrices; APM) and a personality measure (i.e., conscientiousness). Analysis of differential functioning revealed that both the APM and the conscientiousness measure were invariant across formats. Post-hoc analysis found no group means differences for either measure; however attrition rates were statistically significantly higher for the mobile-friendly group.

Smartphones in Selection: Exploring Measurement Invariance using Item Response Theory

by
Brandy Nicole Parker

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Psychology

Raleigh, North Carolina

2014

APPROVED BY:

Adam W. Meade, Ph.D. Committee Co-Chair	Mark A. Wilson, Ph.D. Committee Co-Chair
S. Bartholomew Craig, Ph.D.	Lori Foster Thompson, Ph.D.

BIOGRAPHY

Brandy Nicole Parker was born and raised in Louisville, KY. She graduated from Louisville Male Traditional High School in 2003 and then attended Eastern Kentucky University where she was an Honors Scholar, receiving a Bachelor of Science Degree in Psychology in 2007. After spending a year working in a variety of interesting jobs, Brandy decided what she wanted to be when she grew up; she moved to Raleigh, North Carolina to study Industrial and Organizational (I/O) Psychology at North Carolina State University. While at NC State she discovered that she had a broad range of interests within the field, including selection, training, and psychometrics. Brandy was fortunate to have several great work opportunities throughout her time in graduate school, all of which helped to shape her interests and pay her bills. She completed her Master of Science Degree in I/O Psychology in 2011. While she continued her graduate school experience, Brandy also discovered that there were other fun things outside of I/O, including singing, swing dancing, team triva, and building wonderful friendships.

ACKNOWLEDGEMENTS

I want to start by acknowledging my dissertation committee: Dr. Bart Craig, Dr. Adam Meade, Dr. Lori Foster Thompson, and Dr. Mark Wilson. They have all served as wonderful mentors and role models throughout my graduate school experience. Their advice, support, and occasional teasing helped me not only complete my degree, but helped me grow as a person.

Completing this degree has been a challenge and I would not have had the energy to finish without having my family and friends in my cheering section. I am fortunate to have met so many great people that it would be impossible to mention them all by name, but there are a few I want to single out. My parents, Jim and Cheryl, and my brother, Nick, have made me feel loved and supported, all the way from Kentucky and Minnesota. My cohort: I couldn’t have asked for better people to share my graduate school experience with! Amy, Jen, and Ruchi: I have loved talking and laughing with you ladies. You made the past six years fun! Dave and Lisa Sharek: You have been my family since I first moved to Raleigh and I love you both for it. Lodge McCammon: You have challenged me, provided me with amazing opportunities (who knew I could sing!), and given me great advice and support over the past three years. You are an invaluable friend and I appreciate having you in my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Smartphones in Selection: Exploring Measurement Invariance using Item Response Theory

Over the past several decades, technology has greatly impacted the field of I/O psychology. The area of selection in particular has seen dramatic changes as a result of technological developments. The computer brought forth a more efficient way of processing applicants (McBride, 1998); next followed unproctored internet testing, resulting in cost savings and expansion of the applicant pool (Tippins, 2009). With the continued advances in technology people can now carry computers in their pockets, as many cell phones come equipped with internet connectivity (i.e., smartphones), allowing potential applicants to browse and apply for jobs whenever and wherever they would like.

According to research conducted by the Pew Research Center's Internet and American Life project, 91% of American adults own a cell phone of some kind and 56% own a smartphone (Smith, 2013). Though currently there is no specific information regarding the percentage of cell phone owners who use their device to search or apply for jobs, there is evidence to suggest that this occurs. Some organizations have started tracking the operating system and browser types used by online applicants and they are finding that anywhere from less than 1% to 14% of applicants are using a mobile device (i.e., any portable device with a limited operating system and internet connectivity, such as a smartphone or tablet) to apply for jobs (for examples see Doverspike, Arthur, Taylor, & Carr, 2012; Impelman, 2013; Lawrence, Wasko, Delgado, Kinney, Wolf, 2013; Morelli, Illingworth, Moon, Scott, & Boyd, 2013).

Whether or not organizations are anticipating or prepared for applicants to use smartphones and other mobile devices, the number of mobile device applicants is only expected to grow. In 2011, Fallaw and Kantrowitz (2013) found that 9% of human resource professionals surveyed reported they had candidates request to complete application forms and/or assessments on their mobile device; this number increased to 19% in 2012 and 23% in 2013. Golubovich and Boyce (2013) saw an increase in mobile device use from 3.1% of applicants in 2009 to 14.3% in 2013. Additionally, human resource professionals are growing more interested in testing applicants via mobile device (Fallaw & Kantrowitz, 2013; Fallaw, Kantrowitz, & Dawson, 2012).

Like the transition from paper-and-pencil to internet testing, research is needed to determine if applicant test scores are comparable between mobile and non-mobile devices. This paper aims to add to the small but growing body of research around mobile device use in selection. More specifically, item response theory was used to explore whether measurement invariance held for both a cognitive ability measure (i.e., Raven's Advanced Progressive Matrices; APM) and a personality measure (i.e., conscientiousness).

Mobile Device Usage among Applicants

Organizations are seeing diversity across both race and gender with mobile device applicants. In their sample from a large organization in the restaurant/retail category, Golubovich and Boyce (2013) reported that higher proportions of African-American and Hispanic applicants were using mobile devices (including both smartphones and tablets) to apply for jobs than White applicants. This finding was consistent from 2009 through 2013.

Additionally, there were more female applicants using a mobile device to apply than male applicants. This trend was also observed in applicant data from four organizations in the hospitality industry (Impelman, 2013). Impelman (2013) found that with hourly positions, most mobile device applicants were African-America (50%) and 66% of mobile device applicants were female. Though Doverspike and colleagues (2013) found that the majority of applicants were White (regardless of device used), they did note slightly higher proportions of African-American and Hispanic applicants in the mobile device user category. They also found that the majority of mobile device applicants were female (59%).

Smartphone Use

The racial diversity of mobile applicants is not surprising given the research on smartphone ownership and internet use. According to findings from the Internet and American Life Project, minorities are less likely than Whites to have a home broadband internet connection. Only 49% of African-Americans and 51% of Hispanics have high-speed broadband connection at home, compared to 66% of Whites (Zickuhr & Smith, 2012). Part of the reason for the low percentages can be attributed to cost, as the biggest demographic differences center around household income and education (Zickuhr & Smith, 2012). It seems the smartphone has helped to address disparity in internet access. Of those who were surveyed, 64% of African-Americans and 60% of Hispanics reported that they own a smartphone (Smith, 2013). These two minority groups are also more active with respect to accessing the internet using a phone when compared to Whites. Sixty percent of African-American cell phone owners and 66% of Hispanic cell phone owners reported that they use their phone to access the internet, compared to only 52% of Whites (Duggan & Rainie, 2012). Furthermore, 38% of African-Americans surveyed reported they go online mostly using their smartphone (Smith, 2013).

While direct statistics are lacking, it is plausible that those of lower socioeconomic status are more likely to use a smartphone as their primary means of accessing the internet, including searching and applying for jobs. It can be more cost-effective to own a smartphone (which allows for making calls, accessing email, the internet, playing games, etc.) than to own a regular cell phone (or landline), a computer, and pay for broadband internet. In the U.S., ethnicity and race are linked to socioeconomic status (APA Task Force on Socioeconomic Status, 2007). House and Williams (2000) found that race/ethnicity correlates with almost every indicator of a person's socioeconomic status. It would logically follow that higher proportions of minorities are using their smartphones to apply for jobs, as can be seen in recent research on mobile device use in selection. Thus, offering testing via smartphone may help organizations increase the diversity of their applicant pools.

Mobile Devices in Selection

Research on the use of smartphones and other mobile devices in selection is just beginning. A literature search revealed that there have been no journal publications focused specifically on the use of mobile devices in job applicant testing. The limited available research in this area has come from I/O practitioners, some collaborating with academics, presenting their findings at the Society for Industrial and Organizational Psychology conference. Though much of the research has focused on understanding mobile device use (e.g., Gutierrez & Meyer, 2013) and capturing the demographics of mobile device applicants (e.g., Golubovich & Boyce, 2013), a few studies have examined test invariance (e.g., Morelli et al., 2013) and performance differences (e.g., Doverspike et al., 2012) across mobile and non-mobile applicants.

Data collected by various organizations imply that applicants have been using mobile devices for at least the past few years. Golubovich and Boyce (2013) reported on mobile applicant data from as early as 2009. Despite the limited research, many organizations have their applications accessible via smartphone. In their survey of HR professionals, Fallaw and Kantrowitz (2013) found that approximately 40% of respondents indicated they would allow applicant testing via mobile device, if the option existed. Organizations like Aon Hewitt ("Mobile Enhanced Assessments," n.d.) and PeopleAnswers ("PeopleAnswers Launches Mobile App," 2012) are already offering mobile-compatible assessments. But organizations should proceed with caution. The psychometric properties of a scale used across different mediums of administration, such as a computer and a smartphone, should be examined in order to determine whether the scale is functioning the same way across formats.

Measurement invariance (MI) is the degree to which, under different conditions or formats, scales yield identical measures of the same construct (Horn & McArdle, 1992). If a test/scale is invariant, persons having equal standing on a latent trait should have equal probability of obtaining the same observed score, regardless of being from different samples or groups (Meade & Wright, 2012). If there is a lack of MI (i.e., differential functioning, DF), findings of differences across groups or individuals cannot be reliably interpreted. In employee selection, it is critical to know whether an applicant test is invariant across formats; hiring decisions are made, in part, based on individual scores. If DF is suspected for a measure used across different methods of administration, then the psychological constructs cannot be assumed to be identical (Horn & McArdle, 1992). To date, there have been few studies that compared the psychometric properties of assessments completed on mobile and non-mobile devices (Illingworth et al., 2013; Lawrence et al., 2013; Mitchell & Blair, 2013; Morelli et al., 2012; Morelli et al., 2013).

Non-cognitive measures with mobile devices. Much of the available research on applicant testing via mobile devices (i.e., smartphones and tablets) has focused on personality and other types of non-cognitive measures. Unlike the research on cognitive ability measures, I/O psychologists have already begun to test for MI of non-cognitive measures across device types (Illingworth et al., 2013; Lawrence et al., 2013; Mitchell & Blair, 2013; Morelli et al., 2012; Morelli et al., 2013). Morelli and colleagues (2012) collected data from over 900,000 customer support job applicants on five personality constructs: conscientiousness, customer service, integrity, interpersonal skill, stress tolerance, and teamwork, using both a Likert-type scale and biodata. Using multiple-group confirmatory factor analysis (MGCFA), the authors found the measures of conscientiousness, customer service, integrity, and teamwork to be invariant across devices, except for construct means. These findings were later replicated and extended by Morelli et al (2013). Using data from 664,469 online applicants for a retail sales position, the authors conducted MGCFAs to examine MI for three measures: conscientiousness, curiosity, and customer service. Based on several iterations of model fit, there did not appear to be any differences between mobile and non-mobile users. Additionally, the authors also found no practically significant performance differences (i.e., mean differences) across devices.

Lawrence and colleagues (2013) examined data collected from nearly 200,000 applicants for retail positions on several personality and situational judgment measures: attention to detail, stress tolerance, productivity, likelihood for absenteeism, likelihood for turnover, service potential, and sales potential. Eight percent of the applicants in their sample used a mobile device. Using MGCFA, they found no meaningful differences in model fit when comparing mobile and non-mobile devices, nor did they find meaningful performance differences.

Cognitive ability measures with mobile devices. There have been a few recent studies that utilized mobile applicant data from cognitive ability measures (Doverspike et al., 2012; Hawke, 2013; Impelman, 2013), however, findings focused only on performance differences between mobile and non-mobile applicants. Doverspike and colleagues (2012) examined performance differences on a general mental ability measure comprised of both a verbal and numerical component. Over one million job applicants were included in their study, with applicants free to use the device of their choosing; approximately 1.7% of applicants used a mobile device. The authors found that mobile device users had significantly lower performance scores than non-mobile users on the numeric component, verbal component, and the overall general mental ability measure. Using data from management position applicants across four organizations in the hospitality industry, Impelman (2013) found that the 2.8% of applicants who completed a cognitive ability measure via mobile device performed worse than those using a non-mobile device. Interestingly, differences between mobile and non-mobile applicants in cognitive ability scores were less pronounced among racial minorities; results were mixed for gender.

While these findings imply that mobile device applicants perform worse on cognitive ability tests, there are no studies that have examined whether MI exists across mobile and non-mobile devices for cognitive ability measures. It is possible that the performance differences observed by both Doverspike et al. (2012) and Impelman (2013) could be attributed to a lack of MI. As stated by Vandenberg and Lance (2000), "violations of measurement equivalence assumptions are as threatening to substantive interpretations as is an inability to demonstrate reliability and validity" (p. 6).

Smartphone Formatting and Usability

Though the current research implies that personality measures are invariant across mobile and non-mobile devices, not many studies have examined whether MI will hold under different mobile device conditions. Illingworth and colleagues (2013) explored whether non-cognitive measures were invariant across different device browsers and operating systems, using data from 660,269 retail sales applicants. The authors used MGCFA to explore whether conscientiousness, openness, and customer service measures were invariant across five browser types and five operating systems (to include both non-mobile and mobile devices). Results suggested that all three measures were invariant across all operating systems and browser types.

Display size. Research on the effects of small displays (like that of a smartphone) on information processing supports the idea that I/O psychologists need to explore MI of mobile devices if organizations are to move forward with their use. With smaller displays, textual information often flows across multiple screens, requiring the user to scroll in order to read the entire text (Albers & Kim, 2002). Even when websites are intentionally designed to better display information for a mobile platform, there is still difficulty in displaying all information, in a readable format, on a single screen (Sanchez & Branagahn, 2011). Sanchez and Branagahn (2011) hypothesized that the restrictions of a smaller screen would affect an individual's ability to reason using the information displayed on the screen. They had participants read several emails containing information that was necessary in order to correctly answer a short multiple choice test and found that compared to a full-size display, reasoning deficits occurred when using a small display.

One explanation for these deficits is the effect of scrolling. A small display usually necessitates scrolling in order to view all the information available. If an individual is trying to retain information, scrolling can be taxing because there is a level of stress in maintaining large amounts of information in short-term memory (Albers & Kim, 2002). Albers and Kim (2002) explained that the limitations of short-term memory dictate how much information a person can mentally process while moving (scrolling) from screen to screen. Smaller screens require a person to hold more information in short-term memory for a longer period of time, in order to comparing and evaluate the information. In "overloading" short-term memory, a person may struggle with reasoning and information processing.

In addition to scrolling, Sanchez and Goolsbee (2010) examined whether character size is responsible for reasoning deficits. They found that when reading from a small device, the interaction between screen size and character size used to portray textual information can result in reduced reading performance and comprehension. The authors explained that if characters are too small, it becomes difficult to distinguish the letters and numbers, leading to perceptual jumbling of the characters. The jumbling might increase processing load for the reader, which would take resources away from things like retaining information. However, having too large of characters on a small display resulted in more scrolling, which was found to reduce information recall. The authors found that when the scrolling was kept to a minimum (but the characters weren't too small), factual recall was equivalent to that of a full-size display. Sanchez and Branaghan (2011) found similar results when they had participants recall a series of emails. When participants used a small display with vertical (portrait) orientation, which necessitated scrolling, they performed worse on a multiple choice recall test than those using a large display; however, when the orientation was horizontal (landscape), performance decrements were eliminated. This was because scrolling was reduced.

Mobile website format. Organizations realize the importance of having a website that is easy to view and use on a smartphone. According to Latitude (a research company), 61% of surveyed mobile device users (*N* = 909) said they had a better opinion of a brand when that company offered a good mobile experience (Latitude, 2012). Many IT and marketing companies are shifting focus to mobile users as more people browse websites and shop using their smartphones or tablets. Currently, most companies consider a website to be "mobile-friendly" if it is accessible from a mobile device (e.g., Gallizzi, 2013; "Mobile Friendly vs Mobile Optimized," 2012). A mobile-friendly website looks identical across devices, but the smaller screen of the smartphone means users must scroll from left to right or zoom to better view the webpage. While a mobile-friendly website is functional, it is not ideal. To create a better user experience, many companies and organizations will create a mobile-optimized website. Websites that are mobile-optimized do not require zooming or scrolling left and right and will often have larger navigation buttons and text. Mobile-optimized websites can either be existing websites that auto-detect mobile devices and reformat accordingly ("Mobile Friendly vs Mobile Optimized," 2012) or websites that are specifically designed for a smartphone or tablet (Gallizzi, 2013). Mobile-optimized websites are considered easier to use and navigate.

Organizations that currently allow or are considering smartphone-based assessments should be aware of the differences in mobile website design. If selection assessments are not optimally formatted, it is possible that the aforementioned effects could result in DF. Though Illingworth and colleagues (2013) found that MI held for non-cognitive measures across browser type and operating system, there is more to consider. Of the available research on mobile devices in selection, it is unclear whether the assessments used in various studies were hosted on websites that reformatted specifically for a mobile device (i.e., mobile-optimized) or if they were simply accessible via mobile device (i.e., mobile-friendly). It is plausible that poorly formatted assessments, like those on mobile-friendly websites, would results in DF when compared with either mobile-optimized assessments or non-mobile assessments.

There is still much research to be done in order to understand the effects of using smartphones in applicant testing. Current mean-difference studies, in particular those looking at cognitive ability measures, have assumed MI across devices; thus there is a definite need to identify whether DF exists across devices for cognitive ability measures. Furthermore, studies that have tested for DF in non-cognitive measures have not explicitly examined the effect of smartphone website format, with only two studies (Morelli et al., 2012 and Morelli et al., 2013) differentiating among mobile device types (i.e., smartphone and tablet). The current study seeks to address these gaps by examining whether MI holds for a cognitive ability measure and a personality measure across three device categories: non-mobile, smartphone with mobile-optimized website (smartphone-MO) and smartphone with mobile-friendly website (smartphone-MF). The following research questions were investigated:

Research Question 1: For a cognitive ability measure, does MI hold across non-mobile, smartphone-MO, and smartphone-MF conditions?

Research Question 2: For a personality measure, does MI hold across non-mobile, smartphone-MO, and smartphone-MF conditions?

This study makes several unique contributions to the literature on mobile devices in selection. First, it is the first to examine whether a cognitive ability measure exhibits DF across non-mobile and mobile devices. Second, by employing random assignment, this study eliminates respondent characteristics associated with device choice as a potential effect on measurement properties. Third, this study explores the extent to which mobile website compatibility may influence DF. As stated by Meade et al. (2007), one goal of MI research "should be to specify guidelines regarding when MI likely would be present or absent" (p. 326).

Method

Sample

Data were collected using Mechanical Turk, a crowdsourcing website hosted by the company Amazon. Research by Behrend, Sharek, Meade, and Wiebe (2011) found that participants sourced from Mechanical Turk were more diverse and had more work experience than the traditional participant pool of university students. Additionally, the authors found that the reliability of the data from Mechanical Turk participants was as good as or better than university participants. All participants received a payment of $1.00 upon completion of the study.

The total number of logged survey responses was 943; however, several participants dropped out of the study just after giving consent or after answering only demographic questions, reducing the sample size to 772. Upon inspecting the data, individuals were removed from the sample if they did not follow directions (i.e., they did not use the device assigned to them, *n* = 54), if they did not complete any of the items within the constructs of interest (*n* = 15), if they spent more than 30 minutes completing the cognitive ability measure (*n* = 3), or if they were deemed poor respondents on the conscientiousness measure (*n* = 7). The final survey sample size was 693 participants.

The sample was predominately White (74%) and the majority of the participants were female (58%). The average age for the sample was 31.7 years (*SD* = 9.9; range = 18-69). These demographics were consistent across the three conditions (see Tables 1 and 2). Nearly all participants (*n* = 628) indicated that they were currently employed, though all participants reported having some work experience. Based on those who provided additional information (*n* = 608), the length of current employment ranged from less than 1 week to 30 years. Participants held a wide variety of jobs; reported jobs included nurse, teacher, sales manager, writer, construction worker, waiter, transcriptionist, administrative assistant, and customer service representative, to name a few. The average of the total number of jobs held by participants was five, though the most frequently reported total number of jobs held was three.

Measures

Cognitive ability measure. Cognitive ability was assessed using the 12-item short form of the Raven's Advanced Progressive Matrices Test (APM; Arthur & Day, 1994). The original test, developed by John Raven, is a series of 36 matrix problems that increase in difficulty. For each item, participants are required to select the piece (out of eight options) that completes the pattern (see Appendix A for an example item). The original APM typically takes 40-60 minutes to administer. Arthur and Day (1994) created a shorter version of the APM that still provides a sound assessment of general intelligence. The 12-item APM short form, which takes approximately 15 minutes to complete, includes items from the original test, each increasing in difficulty, and demonstrates psychometric properties similar to the full 36-item test (Arthur & Day, 1994; Arthur, Tubre, Paul, & Sanchez-Ku, 1999).

Personality measure. Conscientiousness was assessed using a 20-item scale (Appendix B) taken from the International Personality Item Pool (IPIP; Goldberg, 1999). This scale was administered as part of the IPIP 100-item measure of the Big Five personality constructs (i.e., extroversion, openness to experience, neuroticism, agreeableness, and conscientiousness). Conscientiousness was selected as a measure of interest because it is one of the most frequently used scales in selection (Schmidt & Hunter, 1998) and it has been shown to be a valid predictor of job performance criteria (Barrick & Mount, 1991). Additionally, conscientiousness has previously been examined for MI across formats. Meade et al. (2007) found that it exhibited strong MI across paper-and-pencil and computer formats. Respondents indicated how accurately each item described themselves using a 5-point scale ranging from 1 (*very inaccurate*) to 5 (*very accurate*). Half of the items were negatively worded and were reverse-coded prior to analysis.

Design

This study employed an experimental design. Participants were randomly assigned to one of three conditions: non-mobile, smartphone-MO, or smartphone-MF. In addition to asking participants to report their device, each participant's operating system and browser information were collected in order to confirm the type of device used. Over 20 different kinds of non-mobile devices were used by participants. Some of the most commonly reported devices used were HP laptop, Macbook Pro, and Dell laptop. Over 50 different smartphone models were reported by participants assigned to the smartphone-MF and smartphone-MO conditions. The most popular models included iPhone 4, iPhone 5, and Samsung Galaxy S3. Tablets (e.g., mobile devices that do not also function as a phone, such as iPads) were not permitted in this study.

Procedure

Participants first completed a qualifying questionnaire (Appendix C). Participants who qualified were randomly presented one of three prompts that included a link to the survey (one link for non-mobile and smartphone-MO, one for smartphone-MF) and a message indicating that the participant should use either a smartphone or non-mobile device to complete the survey. Participants were required to use the type of device assigned to them. Those that did not were dropped from the study and did not receive payment.

At the start of the survey (prior to consent), participants were told that the intent of the survey was to collect data on several types of questions, using different formats, and that the questions in the survey were similar to what might be used by organizations in candidate screening measures. They were asked to treat the survey as though it were a job application, taking time to answer each question to the best of their ability. In order to simulate the effects of completing an actual job application and assessment, participants were told that the top 5% of scorers would receive an additional $0.50.

The average survey completion time was approximately 22 minutes. To verify that the smartphone surveys were perceived as being mobile-optimized or mobile-friendly, a few questions were included that asked participants how they interacted with the survey (Appendix D). Figure 1 shows an example of both the mobile-optimized and mobile-friendly survey formats, as viewed on an iPhone. Additionally, the survey included text entry items (e.g., “What is your job title?”) to approximate an applicant completing biodata-type questions.

Analysis

Model fit. Prior to testing for DF, both the cognitive ability test and the personality assessment were tested for unidimensionality. Mplus Version 5 (Muthén & Muthén, 1998-2011) was used to conduct an exploratory factor analysis (EFA) on each measure. The eigenvalues were examined, looking for a clear delineation of one factor. Next, a two-parameter logistic model (2PLM) was fit to the APM data and a graded-response model (GRM) was fit to the conscientiousness data using IRTPRO (Cai, Thissen, du Toit, 2011). IRT model fit was assessed by examining the M_2 value, Root Mean Square Error of Approximation (RMSEA), and $S\text{-}X^2$ Item Level Diagnostics.

The M_2 statistic is a test-level goodness-of-fit statistic proposed by Maydeu-Olivares and Joe (2005, 2006). This statistic, distributed as χ^2, is recommended for large and/or sparse contingency tables, which is often the case when using a measure with several items and only a few hundred observations. RMSEA is examined along with M_2 to help determine model fit because, like χ^2, M_2 is affected by sample size. The item-level $S\text{-}X^2$, proposed by Orlando and Thissen (2000, 2003), is an indicator of model fit in which observed percentages of item correct scores are compared with those implied by the item characteristic curve (ICC). This statistic is distributed as χ^2, with significant item values indicating model misfit. It can be used with both dichotomous and polytomous data (LaHuis, Clark, & O'Brien, 2011).

DF analysis. Individual items were examined for DF using IRTPRO. An IRT approach is considered a more desirable method than MGCFA when examining the equivalence of a single scale because more information is available for MI testing (Meade & Lautenschlager, 2004). An IRT approach uses a log-linear model (rather than linear, like in MGCFA) to describe the relationship between observed item responses and the underlying trait. This non-linear model is more appropriate than a linear model for dichotomous test data (i.e., the APM test) but is also useful for polytomous data (i.e., the conscientiousness measure).

According to Meade and Wright (2012), the most common method for testing DF in IRT analysis is the all others as anchors (AOAA) approach. In the AOAA approach, the baseline model has all estimated item parameters for like items constrained to be equal across the two groups. One limitation of this approach is that some of the items may exhibit DF, yet are treated as anchors. Meade and Wright (2012) tested several different approaches for selecting anchor items and testing for DF using IRT. Based on their findings, the authors proposed a series of steps to optimally identify anchor items. Though their recommendations were focused around the use of the likelihood ratio test (which has been more commonly used in IRT DF analysis), these steps are also useful with the Wald test. Table 3 outlines four steps (adapted from Meade & Wright, 2012) which were used to test for DF.

Step 1. First, a Wald test was conducted using IRTPRO. The Wald test is an updated version of the Lord's (1980) Wald χ^2 test and provides results similar to those of the more commonly used likelihood ratio test (Woods, Cai, & Wang, 2013). An advantage of the Wald test is that it requires only a single model to be estimated (Woods et al., 2013), using a "test all items, anchor all items" approach. The model compares item parameter estimates between the reference and focal groups, divided by the standard error of their difference. The resulting χ^2 values generated for each item are examined, with significant values signaling potential DF. Like the AOAA approach, the Wald test also suffers from possible DF contamination of the anchor set because all items are both tested and used as anchors, hence the need to follow the steps outlined in Table 3.

Step 2. Looking at only the non-significant items from Step 1, the *a* parameters for each item in the reference group were examined and rank-ordered. Meade and Wright (2012) recommended selecting the five items with the largest *a* parameters to serve as anchors, based on the 20 item measure they used, which was 25% of the total items. Therefore five of the conscientiousness items and three of the APM items were selected from Step 1 to serve as anchors.

Step 3. A second Wald test was conducted, specifying the anchor items identified in Step 2. In this second Wald test, each of the candidate items (i.e., those tested for DF) were evaluated and those with significant χ^2 values were flagged as exhibiting DF.

Step 4. DF effect size indices were computed using the parameter estimates from Step 1 (for anchor items, using the reference group) and Step 3 (for all other items). These effect size indices are important because they provide information on the extent to which items and scales function differently (Meade, 2010). Six effect size indices were examined: signed item difference in the sample (SIDS), unsigned item difference in the sample (UIDS), expected score standardized difference (ESSD), signed test difference in the sample (STDS), unsigned expected test score difference in the sample (UETSDS), and expected test score standardized difference (ETSSD). A description of each of these indices is provided in Table 4 (adapted from Meade, 2010). SIDS, UIDS, and ESSD were included because they are item-level effect size indices while STDS, UETSDS, and ETSSD are test-level effect size indices. Meade's (2010) VisualDF program was used to compute the six effect size estimates.

Though the Wald tests allows for comparison among three groups, pair-wise comparisons between two groups were examined. Limiting the comparison to two groups for each analysis allowed for examination of specifically which group's items were exhibiting DF.

Results

Descriptive Statistics

APM. A total of 623 participants answered at least one of the questions for the APM. There were 258 participants in the non-mobile group, 209 in the smartphone-MO group, and 156 in the smartphone-MF group. Responses were coded as 0 (*incorrect*) or 1 (*correct*); non-responses to an item were coded as missing. Based on this sample, reliability was 0.79.

Scores on the APM ranged from 0-12. Item difficulty statistics appear in Table 5. The distribution of participant scale scores is displayed in Figure 2.

Conscientiousness. A total of 692 participants answered at least one of the questions for the conscientiousness measure. There were 261 participants in the non-mobile group, 217 in the smartphone-MO group, and 214 in the smartphone-MF group. Based on this sample, reliability was 0.94. Participant scores on the conscientiousness measure ranged from 27-100. For each item, the number of respondents per response category was examined. The lowest end of the response scale (1, *very inaccurate*) had small numbers of respondents; in particular, the number of participants who selected a response of 1 ranged from $n = 6$ to $n = 46$. Therefore the two lowest response categories (1, *very inaccurate*; 2, *inaccurate*) were collapsed, changing the conscientiousness measure from a 5-point scale to a 1 to 4 scale for the purposes of the DF analysis. The distribution of participant scale scores is displayed in Figure 3.

Manipulation Check

At the end of the survey, participants in both smartphone groups were asked a few questions about their experience with the mobile website (Appendix D). The mobile-optimized website (smartphone-MO) was expected to be easier to use than the mobile-friendly (smartphone-MF) website. Specifically, the mobile-friendly version was expected to have more zooming, scrolling, and more difficult-to-use navigation buttons (e.g., “next” to advance the survey, response option buttons) compared to the mobile-optimized version, due to its format. As presented in Table 6, higher proportions of smartphone-MF participants indicated having to zoom and scroll during the survey, as well as having more difficultly with the navigation buttons. A χ^2 test revealed that zooming ($\chi^2 = 142.44$, $df = 1$, $p < .01$) and scrolling ($\chi^2 = 29.29$, $df = 1$, $p < .01$) occurred significantly more frequently for the smartphone-MF participants compared to the smartphone-MO participants and a significantly higher proportion of smartphone-MF participants indicated difficulty with navigation buttons ($\chi^2 = 16.66$, $df = 1$, $p < .01$).

Model Fit

APM. The EFA of the APM items indicated unidimensionality, as the first factor had an eigenvalue of 5.46, which was 5 times larger than the second largest eigenvalue (1.03). An attempt was made to fit a three-parameter logistic model (3PLM) to the items in order to model the *c* parameter (pseudo-guessing) because the APM was multiple-choice. The 3PLM resulted in an estimation error during multi-group comparisons, as the maximum sample size in any one of the three groups was $n = 256$ (non-mobile condition). Though a 3PLM would have been preferred, with eight response options per item, the *c* parameter would likely have been small and thus would have had only a small (if any) influence on the *a* or *b* parameters. Therefore a 2PLM was fit to the APM items for entire sample ($n = 623$). The model appeared to fit well to the test as a whole ($M_2 = 60.70$, $p = 0.25$; RMSEA = 0.01). The $S\text{-}X^2$ item-level tests also indicated good model fit for the entire sample (Table 7), though there were four items with statistically significant $S\text{-}X^2$ values, two with $p \leq .05$ and two with $p \leq .01$ (meaning the observed ICC did not fit the expected ICC well). Given that model fit could possibly be influenced by DF (i.e., the estimated parameters may not be the same for all groups, thus skewing fit between expected ICCs and observed ICCs), model fit was examined separately for the three groups (non-mobile, smartphone-MO, smartphone-MF). The 2PLM appeared to fit slightly better when the three groups' parameters were estimated separately ($M_2 = 106.00$, $p = 0.99$; RMSEA = 0.00). Only two items showed statistically significant $S\text{-}X^2$ values, one in the non-mobile group, one in the smartphone-MO group (Table 8).

Conscientiousness. The EFA results showed a first factor eigenvalue of 9.74 and a second factor eigenvalue of 1.27, which indicated unidimensionality for the conscientiousness measure. The GRM was then fit to the conscientiousness items for the entire sample ($n = 693$). The model fit moderately well, given the fit statistics ($M_2 = 4093.95$, $p < 0.01$; RMSEA = 0.05) and that there were only five items with significant $S\text{-}X^2$ values (Table 9). Four of these items were significant at the $p \leq .01$ level, meaning there was a statistically significant difference between the expected ICC and the observed ICC. Like the APM items, model fit was checked for the three groups and model fit improved when the three groups' parameters were estimated separately ($M_2 = 440.61$, $p = 0.029$; RMSEA = 0.01). However, the $S\text{-}X^2$ values (Table 10) varied widely across the three groups. For the non-mobile group, all but five of the items had significant $S\text{-}X^2$ values, but the smartphone-MO group had only one statistically significant item and the smartphone-MF had only four statistically significant items.

APM DF Analysis

Non-mobile versus smartphone-MO. First all items were tested for DF using all items as anchors via the Wald test (Step 1). As seen in Table 11, there was no indication of potential DF in any of the items. Although none of the items had statistically significant χ^2 values, it is still possible that one or more items could exhibit DF later in the process. Thus, the analysis continued on to Steps 2 and 3, with items 2, 4, and 5 selected as anchors. As seen in Table 12, none of the items exhibited DF in Step 3. Because there were no DF items, Step 4 (i.e., computing DF effect size indices) was not completed.

Non-mobile versus smartphone-MF. In Step 1 all items were assessed for DF. As seen in Table 13, none of the items indicated potential DF. Like the non-mobile versus smartphone-MO analysis, items 2, 4, and 5 had the largest reference group *a* parameters and were selected to serve as anchor items (Step 2). Running the analysis again with the three anchor items (Step 3), item 10 was found to have a statistically significant χ^2 value (Table 14). Continuing with Step 4, effect size indices were computed using VisualDF. The SIDS (0.09) is the average difference in expected scores across respondents in the focal group, meaning the smartphone-MF group could be expected to score 0.09 points higher on item 10 than the non-mobile group. Comparing SIDS and UIDS (0.14) indicated that the DF for item 10 was not uniform, which was evident from the ICCs presented in Figure 4. The ESSD for item 10 was 0.49, indicating a medium effect, based on Cohen’s (1988) effect size guidelines. The STDS, which is in the metric of the observed score, was 0.32, meaning the scale scores for the non-mobile and smartphone-MF groups could be expected to differ by

0.32. The UETSDS (the test-level STDS) was 0.52 and the ETSSD, similar to the ESSD, was 0.12 (a small effect size). Given the magnitude of these effect sizes and considering the scale score range of 0-12, DF did not appear to be practically meaningful.

Smartphone-MO versus smartphone-MF. The initial DF analysis (Step 1) yielded only one potential DF item (Table 15). Items 1, 3, and 5 were selected as anchor items in Step 2. Using these three anchor items in the DF analysis, the Wald test (Step 3) revealed that none of the items exhibited DF (Table 16), therefore the analysis did not continue to Step 4.

Summary. Research Question 1 asked "For a cognitive ability measure, does MI hold across non-mobile, smartphone-MO, and smartphone-MF conditions?" Both the non-mobile versus smartphone-MO and smartphone-MO versus smartphone-MF IRT analyses demonstrated no evidence of DF. Although the non-mobile versus smartphone-MF IRT analysis found evidence of one DF item, the associated effect size indices indicated that the DF was not practically meaningful. In answering Research Question 1, MI held for a cognitive ability measure across the three conditions.

Conscientiousness DF Analysis

Non-mobile versus smartphone-MO. Data were first tested for DF using all items as anchors via the Wald test. The results from Step 1 (Table 17) indicated that item 4 might be functioning significantly differently. For Step 2, items 6, 9, 10, 14, and 17 were selected as anchors. The Wald test was conducted again (Step 3) and results showed no statistically significant χ^2 values (Table 18) for any of the items

Non-mobile versus smartphone-MF. In the analysis of the non-mobile and smartphone-MO groups, none of the items were found to have a statistically significant χ^2 (Step 1; Table 19). Continuing with the process, items 4, 6, 9, 10, and 17 were selected as anchors (Step 2). The output from Step 3 was examined, revealing no items that exhibited DF (Table 20).

Smartphone-MO versus smartphone-MF. The analysis (Step 1) of the smartphone-MO and smartphone-MF groups yielded no potential DF (Table 21). Items 4, 5, 9, 10, and 17 were selected to serve as anchors (Step 2) in the second Wald test (Step 3), which found no items with statistically significant χ^2 values (Table 22).

Summary. Research Question 2 asked "For a personality measure, does MI hold across non-mobile, smartphone-MO, and smartphone-MF conditions?" None of the IRT analyses (i.e., non-mobile versus smartphone-MO, non-mobile versus smartphone-MF, smartphone-MO versus smartphone-MF) found evidence of DF. In answering Research Question 2, MI held for a personality measure across the three conditions.

Post-Hoc Analyses

Because both the APM and the conscientiousness measure demonstrated invariance, DF analyses were followed with one-way ANOVAs in order to examine whether there were group mean difference on the two measures. Group means and *SD*s for both measures appear in Table 23. A Levene's test, conducted for both measures, was not statistically significant, meaning the variances of the groups were the same. The results from the ANOVA indicated that the group means for both the APM ($F(2, 620) = 1.82$, $p = .16$) and conscientiousness ($F(2, 689) = 0.07$, $p = .93$) were not statistically significantly different.

One point of interest in this study is the group differences in attrition. Data collection time was extended due to the high rate of attrition from those in the smartphone groups, particularly the smartphone-MF group. As previously explained, the starting sample size of 943 was reduced to 772 due to several participants dropping out just after consent or after answering only a few demographic questions. In this first reduction of sample size, the number of participants dropping from the smartphone-MF condition appeared to be considerably larger than the other two groups (Table 24). A χ^2 test indicated that the difference in attrition rate between the starting and the reduced sample was statistically significant ($\chi^2 = 69.90$, $df = 2$, $p < .01$), with the smartphone-MF group having the greatest reduction ($n = 110$). After cleaning the data (i.e., removing participants who did not follow instructions, who did not answer any of the items for either of the constructs of interest, who spent more than 30 minutes on the APM, or who responded poorly on the conscientiousness measure), the sample was further reduced to the final sample size of 693 (Table 24). Comparing the reduced sample and the final sample, a χ^2 test indicated that the attrition rate was statistically significantly different among the groups ($\chi^2 = 55.34$, $df = 2$, $p < .01$), with the non-mobile group having less attrition than the two smartphone groups.

Discussion

With the increase in smartphone ownership, more and more applicants have been using their smartphones and other mobile devices to apply online for jobs in various industries (Doverspike et al., 2013; Impelman, 2013). Organizations are seeing more diversity in the mobile device applicant pool compared to the non-mobile applicant pool (e.g., Golubovich & Boyce, 2013) and, of the available research using data from organizations, many studies found higher proportions of African-American and Hispanic applicants in the mobile device applicant pool compared to the non-mobile applicant pool (Doverspike et al., 2013; Golubovich & Boyce, 2013). Given the current trends on smartphone ownership and use (Zickuhr & Smith, 2012; Smith, 2013) it appears that the use of mobile devices in selection will continue to grow. Currently there is very little research investigating whether mobile applicants are at a disadvantage compared to non-mobile applicants. The available research indicates that mobile device applicants perform worse on cognitive ability measures compared with non-mobile applicants (Doverspike et al., 2013), while other studies show no mean differences on non-cognitive measures (Morelli et al., 2013). A limitation of those studies, however, is that applicants self-selected the device they used to complete the application.

The purpose of the present study was to examine whether a cognitive ability measure (the APM short form) and a personality measure (conscientiousness) were invariant across non-mobile and mobile device users. Additionally, this study examined whether smartphone website format affected DF under random assignment conditions. Participants were randomly assigned to complete a survey in one of three conditions: non-mobile, smartphone-MO, and smartphone-MF. Both the APM and the conscientiousness measure were tested for DF using IRT.

Despite finding one item on the APM that exhibited statistically significant DF (between the non-mobile and smartphone-MF groups), this DF was not practically meaningful; thus the measure was invariant across the three groups. This finding adds to the literature because to date there have been no studies of MI across non-mobile and mobile devices for a cognitive ability measure. The findings from this study suggest that if an organization uses a cognitive ability measure in their online application process, applicants are not at risk of being inappropriately evaluated should they choose to use a mobile device. The post-hoc analysis of the APM showed no group mean difference, meaning that mobile and non-mobile applicants performed similarly on the cognitive ability measure. This finding was not consistent with the two other studies that examined cognitive ability (Doverspike et al., 2013; Impelman, 2013); those authors found statistically significant mean differences across non-mobile and mobile applicants. One possible explanation for this difference in findings is that the cognitive ability measures used in those studies were not invariant across device types, as neither study reported tests of invariance. In fact, Doverspike and colleagues (2013) noted that the mean differences might be reduced by the introduction of a device-specific website, implying the mobile website format could have resulted in DF for the cognitive ability measure. Another explanation is that the participants in those studies who chose to use a mobile device to complete the online assessments did in fact have lower cognitive ability than those who used a computer.

The conscientiousness measure showed no signs of DF, therefore it too demonstrated MI. This is consistent with other researchers' findings of MI for personality constructs

(Morelli et al., 2012; Illingworth et al., 2013; Lawrence et al., 2013; Mitchell & Blair, 2013; Morelli et al., 2013). The post-hoc analysis showed no group mean difference for conscientiousness. Again, this is consistent with previous research (e.g., Lawrence et al., 2013; Morelli et al., 2013).

A particularly interesting finding was that there was a statistically significant difference in attrition. Far fewer non-mobile participants quit the survey than did smartphone-MF participants (Table 22). All survey respondents were paid for their participation so it was somewhat unexpected that the smartphone-MF group would drop out at a statistically significantly higher rate. A possible explanation is that those who started the survey on their smartphone found the mobile-friendly survey too difficult to use. The mobile-friendly survey was less user-friendly than the mobile-optimized survey, as evidenced by the responses to the manipulation check (Table 6). Perhaps the smartphone-MF participants did not feel that a payment of $1.00 was sufficient for the challenge of completing the survey on a mobile-friendly website.

Furthermore, both groups of smartphone participants were removed from the sample during data cleaning at a statistically significantly higher rate than the non-mobile participants. The biggest reason participants were dropped was not following directions. Both smartphone groups had several participants who either used, and reported using, a computer or who used a computer but reported using a smartphone, which resulted in them being removed from the sample. It is possible that the participants who falsely reported using a smartphone anticipated that completing the survey on a smartphone would be difficult and thus, they chose not to follow instructions.

These findings related to attrition imply two things. First, some applicants may not want to use their smartphones. The participants who did not follow directions are an indication that some individuals prefer to use a computer to do things like complete an online survey. Second, organizations may be at risk of losing applicants simply because of poorly formatted application websites. A statistically significantly higher proportion of smartphone-MF applicants quit the survey soon after starting it, likely because of the difficultly of using a mobile-friendly website. Though many applicants may attempt to apply via smartphone, having a poorly formatted mobile website may cause organizations to lose out on many potentially well-qualified applicants due to attrition. This is particularly important because organizations are seeing higher proportions of minority applicants within the mobile-device category, so a mobile-friendly website could also cost the organization to lose out on potentially well-qualified minority applicants.

Limitations and Future Research

One limitation of this study was the somewhat small sample size. Having a larger sample would likely have allowed the use of a 3PLM for the APM. Though a 2PLM is appropriate to use, a 3PLM would have better modeled potential guessing and thus a better analysis of DIF. A larger sample would have also given more power to all of the statistical analyses.

Another limitation of this study is the generalizability of the findings. Though efforts were made to approximate the attentiveness of a real job applicant (i.e., paying participants, informing them of a bonus for top scorers), the participants were likely not invested in the same way an actual job applicant would be invested. Real job applicants might feel pressure/stress when completing an online job application because they are striving to do their best in order to obtain a job. The participants in this study likely felt minimal stress while completing the survey because they would be rewarded (paid $1.00) so long as they finished. It is possible that, given the minimal stress, these participants were less prone to making errors while on a mobile device than a real applicant might make. Future research should attempt to create a high-fidelity situation, perhaps having participants complete the study in a style more similar to a selection context (e.g., have rewards for the best applicants, use a mock job application).

Mechanical Turk was used to collect data because these participants typically have more work experience and vary more in age than do undergraduate samples (Behrend et al., 2011), making the study sample more similar to the current work population. Furthermore, the survey respondents reported having wide variety of jobs, which lends to generalizability of these results across job types. However, Mechanical Turk users are unique simply because of their use of Mechanical Turk. The individuals who choose to spend their time (as a full time job, as supplemental income, or just for fun) completing surveys and other tasks online may not be similar to the average worker, therefore it is possible that these participants are more “tech savvy” than most and are less likely to experience stress or difficulty in completing a survey using a mobile device.

Conclusion

With the rise in mobile device applicants, organizations need to ensure whether the measures they use in their online job applications are invariant across device types. If comparisons are to be made between non-mobile and mobile applicants, it is critical that the tests or surveys used yield identical measures of the same construct. A lack of invariance in a selection measure could result in an organization making mistakes in their selection process. This is particularly concerning considering current research suggests that higher proportions of minority applicants are using mobile devices than Whites. Though the results of this study found the APM and a conscientiousness measure to be invariant across three conditions, organizations should be aware that participants could be deterred from applying for positions if the job application website is poorly formatted.

REFERENCES

References preceeded by an asterisk were included only in the proposal (see Appendix E).

Albers, M., & Kim, L. (2002). Information design for the small-screen interface: An overview of web design issues for personal digital assistants. *Technical Communication, 49*, 45-60.

APA Task Force on Socioeconomic Status. (2007). *Report of the APA Task Force on Socioeconomic Status*. Retrieved from: http://www.apa.org/pi/ses/resources/publications/task-force-2006.pdf

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

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

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44,* 1-26.

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods, 43*, 800-813. doi: 10.3758/s13428-011-0081-0

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Lincolnwood, IL: Scientific Software International.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Earlbaum.

Duggan, M., & Rainie, L. (2012). Cell phone activities 2012. Retrieved from Pew Research Center website: http://www.pewinternet.org/Reports/2012/Cell-Activities.aspx

Doverspike, D., Arthur, Jr., W., Taylor, J., & Carr, A. (2012, April). Mobile mania: The impact of device type on remotely delivered assessments. In J. Scott (chair), *Chasing the tortoise: Zeno's paradox in technology-based assessment*. Symposium conducted at the 27th annual conference of The Society for Industrial and Organizational Psychology, San Diego, CA.

Fallaw, S. S., & Kantrowitz, T. M., & Dawson, C. R. (2012). Global assessment trends report. Retrieved from SHL website: http://www.shl.com/assets/GATR_2012_US.pdf

Fallaw, S. S., & Kantrowitz, T. M. (2013). Global assessment trends report. Retrieved from SHL website: http://www.shl.com/assets/GATR_2013_US.pdf

Gallizzi, M. (2013, March 10). Mobile-friendly or mobile-optimized websites. Retrieved from http://www.moiremarketing.com/blog/mobile-friendly-or-mobile-optimized-websites

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De

Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe*, *Vol. 7* (pp. 7-28). Tilburg, The Netherlands: Tilburg University Press.

Golubovich, J., & Boyce, A. S. (2013, April). Hiring tests: Trends in mobile device usage. In N. 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.

Gutierrez, S., & Meyer, J. (2013, April). Assessments on the go: Applicant reactions to mobile testing. In N. 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.

Hawkes, B. (2013, April). Developing evidence-based guidelines for testing on mobile devices. In C. A. Hedricks (Chair), *Goin' mobile: Employers, applicants, and their references*. Symposium conducted at the 28th annual conference of The Society for Industrial and Organizational Psychology, Houston, TX.

House, J. S., & Williams, D. R. (2000). Understanding and reducing socioeconomic and racial/ethnic disparities in health. In B. D. Smedley & S. L. Syme (Eds.), *Promoting health: Intervention strategies from social and behavioral research* (pp. 81-125). Washington, DC: National Academy Press.

Horn, J. L., & Mcardle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*, 117-144. doi: 10.1080/03610739208253916

Illingworth, A. J., Morelli, N. A., Scott, J. C., Moon, S., & Boyd, S. (2013, April). Equivalence of assessments on mobile devices: Impact of device software. In N. 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.

Impelman, K. (2013, April). Mobile assessment: Who's doing it and how it impacts selection. In N. 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.

LaHuis, D. M., Clark, P., & O'Brien, E. (2011). An examination of item response theory item fit indices for the graded response model. *Organizational Research Methods, 14*(1), 10-23. doi: 10.1177/1094428109350930

Latitude. (2012, December). Next-gen retail: Mobile and beyond. Retrieved from: http://files.latd.com/Latitude-Next-Gen-Retail-Study.pdf

Lawrence, A., Wasko, L., Delgado, K., Kinney, T., & Wolf, D. (2013, April). Does mobile assessment administration impact psychological measurement? In N. 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.

Lopez Rivas, G. E., Stark, S., & Chernyshenko, O. S. (2009). The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test. *Applied Psychological Measurement, 33*, 251–265. doi: 10.1177/0146621608321760

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum

Maydeu-Olivares, A., & Joe, H. (2005). Limited and full information estimation and testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association, 100*, 1009–1020. doi: 10.1198/016214504000002069

Maydeu-Olivares, A. & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71*, 713-732. doi: 10.1007/s11336-005-1295-9

McBride, J. R. (1998). Innovations in computer-based ability testing: Promise, problems, and perils. In M. D. Hakel (Ed.), *Beyond multiple choice: Evaluating alternatives to traditional testing for selection* (pp. 23-39). Mahwah, NJ: Lawrence Erlbaum.

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology, 95*, 728–743. doi: 10.1037/a0018966

Meade, A. W., & Lautenschlager, G. J. (2004, April). Same question, different answers: CFA and two IRT approaches to measurement invariance in C. E. Lance (chair), *Issues and advances in measurement equivalence/invariance research*. Symposium conducted at the 19th annual conference of The Society for Industrial and Organizational Psychology, Chicago, IL.

Meade, A. W., Michels, L. C., & Lautenschlager, G. J. (2007). Are internet and paper-and-pencil personality tests truly comparable: An experimental design measurement invariance study. *Organizational Research Methods, 10*, 322-345. doi: 10.1177/1094428106289393

Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology, 97*, 1016-1031. doi: DOI: 10.1037/a0027934

Mitchell, D., & Blair, M. (2013, April). Goin' mobile: A mobile provider's foray into mobile assessments. In C. A. Hedricks (Chair), *Goin' mobile: Employers, applicants, and their references*. Symposium conducted at the 28th annual conference of The Society for Industrial and Organizational Psychology, Houston, TX.

Mobile-enhanced assessments. (n.d.). Retrieved from http://www.aon.com/human-capital-consulting/consulting/Mobile-Enabled_Assessments.jsp

Mobile friendly vs mobile optimized vs responsive design: What you need to know about the mobile version of your website. (2012, June 14). Retrieved from http://www.signalfire.us/mobile-friendly-vs-mobile-optimized-vs-responsive-design/

Morelli, N., Illingworth, A. J., Moon, S., Scott, J., and Boyd, S. (2013, April). *Equivalence of Assessments on Mobile Devices: A Replication and Extension*. Poster presented at the 28th annual conference of the Society for Industrial & Organizational Psychology in Houston, TX.

Morelli, N. A., Illingworth, A. J., Scott, J. C., & Lance, C. E. (2012, April). Are internet-based, unproctored assessments on mobile and non-mobile devices equivalent? In J. Scott (chair), *Chasing the tortoise: Zeno's paradox in technology-based assessment*. Symposium presented at the 27th annual conference of The Society for Industrial and Organizational Psychology, San Diego, CA.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (Seventh Edition). Los Angeles, CA: Muthén & Muthén.

Orlando, M., & Thissen, D. (2000). Likelihood-based item fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 50-64. doi: 10.1177/01466216000241003

Orlando, M. & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*, 289-298. doi: 10.1177/0146621603027004004

PeopleAnswers launches mobile app for employee assessment software. (2012, October 17). Retrieved from https://www.peopleanswers.com/aboutus_news_pressRelease_detail?reqItemId=52

Qualtrics [Survey Software]. Provo, UT: http://www.qualtrics.com.

Sanchez, C. A., & Branaghan, R. J. (2011). Turning to learn: Screen orientation and reasoning with small devices. *Computers in Human Behavior, 27*, 793-797. doi: 10.1016/j.chb.2010.11.004

Sanchez, C. A., & Goolsbee, J. Z. (2010). Character size and reading to remember from small displays. *Computers & Education, 55,* 1056-1062. doi: 10.1016/j.compedu.2010.05.001

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262-274.

Smith, A. (2013). *Smartphone ownership – 2013 update*. Retrieved from Pew Research Center webiste: http://www.pewinternet.org/Reports/2013/Smartphone-Ownership-2013.aspx

*Stark, S. (2001). MODFIT: A computer program for model-data fit. Urbana-Champaign: University of Illinois at Urbana- Champaign. Retrieved from http://work.psych.uiuc.edu/irt

Tippins, N.T. (2009). Internet alternatives to traditional proctored testing: Where are we now? *Industrial and Organizational Psychology, 2*, 2-10.

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*, 4-70. doi: 10.1177/109442810031002

Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement, 73*, 532-547. doi: 10.1177/0013164412464875

Zickuhr, K., & Smith, A. (2012). Digital differences. Retrieved from Pew Internet & American Life Project website: http://pewinternet.org/Reports/2012/Digital-differences.aspx

Table 1

Participant Gender and Race by Condition

	Condition			
Variable	Non-Mobile	Smartphone-MO	Smartphone-MF	Total
Gender				
Male	110	92	91	293
Female	151	125	123	399
Race				
African and/or African American	24	14	14	52
Asian and/or Asian American	16	15	21	52
Caucasian and/or European American	192	170	154	516
Hispanic	23	15	11	49
Native American and/or Alaskan Native	2	1	5	8
Native Hawaiian and/or Pacific Islander	0	0	1	1
Other	4	2	9	15

Table 2

Descriptive Statistics for Participant Age by Condition

Condition	*n*	*M*	*SD*	Range
Non-Mobile	260	33.63	11.38	19-69
Smartphone-MO	216	30.90	8.88	18-65
Smartphone-MF	214	30.16	8.64	18-55
Total	690	31.70	9.94	18-69

Table 3

Recommended Best-Practice Steps in Conducting IRT Invariance Analyses

Step	Description
1	Conduct DF analyses via the "test all items, anchor all items" approach.
2	Of the non-significant items, rank-order the items by the largest reference group *a* parameters and select approximately 25% of total items to serve as anchors.
3	Conduct DF analysis using the anchor items identified in Step 2, evaluating DF significance levels.
4	Compute DF effect size indices using parameter estimates from Step 1 (for anchor items, using reference group) and Step 3 (for all other items).

Note. IRT = item response theory; DF = differential functioning. Adapted from "Solving the Measurement Invariance Anchor Item Problem in Item Response Theory," by A. W. Meade and N. A. Wright, 2012, *Journal of Applied Psychology, 97,* p. 1030.

Table 4

IRT Differential Functioning Effect Size Index Descriptions

Index		Description
SIDS	Signed item difference in the sample	Average difference in ESs across focal group respondents as compared to reference group respondents. DF across respondents is allowed to cancel in cases of nonuniform differences in ESs.
UIDS	Unsigned item difference in the sample	The average difference in ESs across focal group sample respondents had differences been uniform in nature. Comparing UIDS and SIDS gives an indication of the extent to which differences in ESs cancel across different respondents.
ESSD	Expected score standardized difference	An ES version of Cohen's *d*. Mean ESs are computed for the focal group respondents using both focal and reference item parameters. The difference between these means are divided by the pooled SD of the two sets of ESs. The metric can be interpreted using the guidelines given by Cohen (1988).
STDS	Signed test difference in the sample	The difference in observed summed scale scores expected, on average, across all focal group respondents, due to DF alone. Allows cancellation of DF across both items and persons.
UETSDS	Unsigned expected test score difference in the sample	The hypothetical difference in expected scale scores that would have been present if scale-level DF had been uniform across respondents. Allows cancellation of DF across items but not persons.
ETSSD	Expected test score standardized difference	An ETS version of Cohen's *d*. The metric can be interpreted using the guidelines on effect size given by Cohen (1988).

Note. In all cases, focal group expected scores are computed using item parameters estimated in both the focal group sample and the reference group sample. Uniform differences in expected scores mean that one group is favored. Nonuniform differences in expected scores indicate that for some respondents, expected scores will be higher for the focal group and for other respondents expected scores will be higher for the reference group. ES = expected score; DF = differential functioning; ETS = expected test score. Adapted from "A

Taxonomy of Effect Size Measure for the Differential Functioning of Items and Scales," by A. W. Meade, 2010, *Journal of Applied Psychology, 95,* pp. 732-733.

Table 5

APM Item Difficulty by Group

	Non-Mobile		Smartphone-MO		Smartphone-MF		Total	
Item	*n*	Proportion Correct	*n*	Proportion Correct	*n*	Proportion Correct	*n*	Proportion Correct
1	257	0.70	209	0.70	156	0.72	622	0.71
2	258	0.67	207	0.68	156	0.71	621	0.68
3	257	0.62	208	0.66	155	0.68	620	0.65
4	256	0.66	209	0.65	154	0.70	619	0.67
5	256	0.63	209	0.60	154	0.65	619	0.62
6	256	0.48	207	0.48	154	0.52	617	0.49
7	256	0.37	206	0.35	153	0.42	615	0.37
8	256	0.30	206	0.28	152	0.38	614	0.31
9	256	0.40	205	0.44	151	0.48	612	0.44
10	255	0.24	204	0.26	149	0.35	608	0.27
11	256	0.28	203	0.24	148	0.28	607	0.27
12	256	0.16	203	0.17	147	0.22	606	0.18

Table 6

Smartphone Manipulation Check

	Smartphone-MO			Smartphone-MF		
Smartphone Use Questions	*n*	No	Yes	*n*	No	Yes
At any point during the survey, did you enlarge (i.e., “zoom in”) the text at any point to better read or answer questions?	203	85%	15%	145	21%	79%
At any point during the survey, did you have to scroll left and right at any point to better read or answer questions?	203	40%	60%	144	13%	87%
Did you think the navigation buttons (i.e., the “next page” arrow for the survey, the multiple choice response options) were difficult to use?	203	80%	20%	144	60%	40%

Table 7

S-X2 Item Level Diagnostic Statistics for 2PLM of APM Items

Item	χ^2	*df*	*p*
1	4.94	8	0.76
2	15.53	8	0.05*
3	10.94	9	0.28
4	4.60	8	0.80
5	4.40	9	0.88
6	11.55	10	0.32
7	14.30	10	0.16
8	7.07	10	0.72
9	7.23	10	0.70
10	22.36	10	0.01**
11	22.50	9	0.01**
12	20.08	9	0.02*

Note. $n = 623$.
*$p \leq .05$. **$p \leq .01$.

Table 8

S-X2 Item Level Diagnostic Statistics for 3PLM of APM Items by Group

	Non-Mobile (*n* = 258)			Smartphone-MO (*n* = 209)			Smartphone-MF (*n* = 156)		
Item	χ^2	*df*	*p*	χ^2	*df*	*p*	χ^2	*df*	*p*
1	4.76	7	0.69	4.73	7	0.69	6.65	5	0.25
2	3.35	7	0.85	5.12	7	0.65	12.51	8	0.13
3	10.26	8	0.25	5.74	7	0.57	5.03	6	0.54
4	2.75	7	0.91	3.96	8	0.86	7.98	6	0.24
5	4.91	6	0.56	5.94	7	0.55	3.73	7	0.81
6	8.60	9	0.48	6.76	8	0.56	5.24	8	0.73
7	5.68	8	0.68	11.86	9	0.22	7.02	7	0.43
8	9.11	9	0.43	7.28	9	0.61	4.86	9	0.85
9	4.18	8	0.84	2.29	7	0.94	6.60	10	0.76
10	12.45	9	0.19	22.15	10	0.01**	3.14	7	0.87
11	16.73	9	0.05*	4.20	9	0.90	7.55	7	0.38
12	10.99	8	0.20	11.57	8	0.17	4.33	6	0.63

*$p \leq .05$. **$p \leq .01$.

Table 9

S-X2 Item Level Diagnostic Statistics for GRM of Conscientiousness Items

Item	χ^2	*df*	*p*
1	109.91	93	0.11
2	100.78	72	0.01**
3	107.92	100	0.28
4	88.64	70	0.07
5	87.46	76	0.17
6	78.77	68	0.17
7	101.33	97	0.36
8	104.99	99	0.32
9	112.67	70	0.00**
10	58.72	61	0.56
11	98.12	94	0.36
12	80.41	91	0.78
13	147.40	110	0.01**
14	92.03	78	0.13
15	140.06	97	0.00**
16	101.35	105	0.58
17	100.79	81	0.07
18	124.45	94	0.02*
19	102.79	96	0.30
20	109.24	101	0.27

Note. $n = 692$.
*$p \leq .05$. **$p \leq .01$.

Table 10

S-X2 Item Level Diagnostic Statistics for GRM of Conscientiousness Items by Group

	Non-Mobile (*n* = 261)			Smartphone-MO (*n* = 217)			Smartphone-MF (*n* = 214)		
Item	χ^2	*df*	*p*	χ^2	*df*	*p*	χ^2	*df*	*p*
1	70.69	46	0.01**	45.30	47	0.54	48.94	49	0.48
2	55.75	42	0.08	47.70	38	0.13	35.08	36	0.51
3	71.52	56	0.08	75.54	56	0.04*	81.60	55	0.01**
4	65.69	37	0.00**	32.53	31	0.39	43.34	31	0.07
5	52.39	38	0.06	42.45	33	0.13	45.38	41	0.29
6	52.57	31	0.01**	27.40	29	0.55	43.21	38	0.26
7	70.88	49	0.02*	50.02	48	0.39	54.67	51	0.34
8	88.19	55	0.00**	40.97	49	0.79	42.95	50	0.75
9	74.94	33	0.00**	42.58	37	0.24	61.68	34	0.00**
10	44.17	25	0.01**	28.03	28	0.46	42.51	31	0.08
11	70.85	57	0.10	51.46	51	0.46	53.14	47	0.25
12	65.06	50	0.07	51.15	52	0.51	53.49	46	0.21
13	98.65	59	0.00**	47.68	51	0.61	57.51	53	0.31
14	67.81	40	0.00**	56.01	41	0.06	42.26	38	0.29
15	79.10	52	0.01**	59.51	47	0.10	64.90	48	0.05*
16	69.97	47	0.02*	59.25	51	0.20	58.01	47	0.13
17	57.83	40	0.03*	42.94	40	0.35	52.57	37	0.05*
18	61.01	43	0.04*	43.80	41	0.35	47.58	42	0.26
19	83.21	48	0.00**	59.89	47	0.10	48.51	46	0.37
20	85.27	55	0.01**	67.43	61	0.27	54.63	57	0.57

*$p \leq .05$. **$p \leq .01$.

Table 11

Step 1 for APM Non-Mobile vs Smartphone-MO

				Non-Mobile		Smartphone-MO	
Item	χ2	*df*	*p*	*a*	*b*	*a*	*b*
1	0.10	2	0.95	1.83	-0.72	2.02	-0.69
2	0.20	2	0.91	2.04	-0.57	1.83	-0.62
3	1.50	2	0.46	1.41	-0.48	1.91	-0.54
4	1.30	2	0.52	1.87	-0.55	1.35	-0.61
5	0.60	2	0.76	2.17	-0.41	1.89	-0.33
6	3.00	2	0.23	0.99	0.09	1.65	0.06
7	2.10	2	0.35	1.52	0.49	1.02	0.76
8	0.90	2	0.65	0.78	1.25	1.04	1.10
9	1.30	2	0.53	1.42	0.38	1.81	0.20
10	0.30	2	0.86	0.65	1.95	0.72	1.61
11	1.00	2	0.61	1.26	0.98	1.13	1.27
12	1.30	2	0.51	1.31	1.64	0.87	2.07

Note. Reference group = non-mobile, focal group = smartphone-MO. *a* and *b* are estimated item parameters.

Table 12

Step 3 for APM Non-Mobile versus Smartphone-MO Using Items 2, 4, 5 as Anchors

				Non-Mobile		Smartphone-MO	
Item	$\chi2$	*df*	*p*	*a*	*b*	*a*	*b*
1	1.00	2	0.61	1.84	-0.72	2.45	-0.65
3	3.30	2	0.19	1.41	-0.48	2.34	-0.53
6	4.60	2	0.10	0.99	0.09	2.01	-0.04
7	0.50	2	0.77	1.53	0.48	1.24	0.54
8	1.70	2	0.44	0.78	1.25	1.26	0.82
9	2.90	2	0.24	1.42	0.38	2.20	0.08
10	1.10	2	0.57	0.64	1.96	0.88	1.24
11	0.10	2	0.95	1.26	0.98	1.38	0.95
12	1.30	2	0.53	1.31	1.65	1.06	1.62

Note. Reference group = non-mobile, focal group = smartphone-MO. *a* and *b* are estimated item parameters.

Table 13

Step 1 for APM Non-Mobile versus Smartphone-MF

				Non-Mobile		Smartphone-MF	
Item	χ2	*df*	*p*	*a*	*b*	*a*	*b*
1	0.50	2	0.77	1.83	-0.72	2.28	-0.62
2	2.60	2	0.27	2.04	-0.57	1.26	-0.75
3	0.80	2	0.66	1.41	-0.48	1.81	-0.50
4	0.10	2	0.96	1.87	-0.55	2.04	-0.53
5	1.60	2	0.46	2.17	-0.41	1.51	-0.39
6	0.20	2	0.89	0.99	0.09	1.12	0.15
7	0.30	2	0.86	1.52	0.49	1.60	0.58
8	0.80	2	0.68	0.78	1.25	1.01	0.90
9	4.80	2	0.09	1.42	0.38	0.74	0.35
10	4.70	2	0.09	0.65	1.95	1.36	0.90
11	1.80	2	0.41	1.26	0.98	1.16	1.35
12	0.20	2	0.91	1.31	1.64	1.52	1.48

Note. Reference group = non-mobile, focal group = smartphone-MF. *a* and *b* are estimated item parameters.

Table 14

Step 3 for APM Non-Mobile versus Smartphone-MF Using Items 2, 4, 5 as Anchors

				Non-Mobile		Smartphone-MF	
Item	χ2	*df*	*p*	*a*	*b*	*a*	*b*
1	2.30	2	0.32	1.82	-0.72	3.02	-0.56
3	3.10	2	0.22	1.43	-0.47	2.36	-0.47
6	1.40	2	0.51	1.00	0.09	1.46	0.03
7	1.10	2	0.57	1.49	0.49	2.05	0.36
8	2.90	2	0.24	0.78	1.25	1.30	0.61
9	2.80	2	0.25	1.42	0.38	0.97	0.18
10	7.80	2	0.02*	0.65	1.93	1.75	0.62
11	0.30	2	0.84	1.27	0.98	1.50	0.96
12	2.40	2	0.30	1.31	1.64	1.99	1.06

Note. Reference group = non-mobile, focal group = smartphone-MF. *a* and *b* are estimated item parameters.
*p < .05

Table 15

Step 1 for APM Smartphone-MO versus Smartphone-MF

				Smartphone-MO		Smartphone-MF	
Item	χ2	*df*	*p*	*a*	*b*	*a*	*b*
1	0.20	2	0.92	2.00	-0.71	2.26	-0.64
2	1.50	2	0.48	1.81	-0.63	1.24	-0.76
3	0.10	2	0.95	1.90	-0.55	1.79	-0.51
4	1.50	2	0.48	1.34	-0.63	2.02	-0.55
5	0.60	2	0.74	1.88	-0.34	1.49	-0.40
6	1.70	2	0.42	1.64	0.05	1.11	0.14
7	1.80	2	0.40	1.01	0.76	1.58	0.57
8	0.70	2	0.70	1.03	1.10	0.99	0.90
9	6.60	2	0.04*	1.80	0.19	0.73	0.34
10	3.00	2	0.22	0.72	1.62	1.35	0.91
11	0.20	2	0.93	1.13	1.27	1.15	1.36
12	1.90	2	0.40	0.87	2.07	1.51	1.49

Note. Reference group = smartphone-MO, focal group = smartphone-MF. *a* and *b* are estimated item parameters.
**p* < .05.

Table 16

Step 3 for APM Smartphone-MO versus Smartphone-MF Using Items 1, 3, 5 as Anchors

				Smartphone-MO		Smartphone-MF	
Item	χ2	*df*	*p*	*a*	*b*	*a*	*b*
2	1.00	2	0.62	1.82	-0.63	1.32	-0.76
4	1.70	2	0.43	1.33	-0.63	2.14	-0.56
6	1.10	2	0.59	1.63	0.05	1.19	0.09
7	2.30	2	0.31	1.01	0.76	1.68	0.50
8	0.90	2	0.64	1.03	1.10	1.06	0.81
9	5.70	2	0.06	1.81	0.19	0.78	0.28
10	4.00	2	0.14	0.72	1.62	1.43	0.81
11	0.10	2	0.97	1.12	1.27	1.22	1.23
12	1.70	2	0.42	0.87	2.07	1.59	1.36

Note. Reference group = smartphone-MO, focal group = smartphone-MF. *a* and *b* are estimated item parameters.

Table 17

Step 1 for Conscientiousness Non-Mobile versus Smartphone-MO

				Non-Mobile				Smartphone-MO			
Item	χ2	*df*	*p*	*a*	*b*1	*b*2	*b*3	*a*	*b*1	*b*2	*b*3
1	2.70	4	0.60	2.26	-1.53	-0.88	0.77	1.87	-1.56	-0.95	0.77
2	1.80	4	0.78	1.91	-2.28	-1.83	-0.35	1.85	-2.39	-1.85	-0.18
3	1.60	4	0.81	1.72	-0.96	-0.56	0.96	1.43	-1.07	-0.57	1.00
4	10.30	4	0.04*	2.84	-2.18	-1.26	0.11	3.76	-1.49	-1.10	-0.01
5	8.10	4	0.09	2.79	-1.77	-1.20	0.42	3.45	-1.34	-1.02	0.18
6	1.60	4	0.80	2.87	-1.93	-1.57	-0.14	2.62	-2.26	-1.65	-0.14
7	1.90	4	0.76	1.53	-2.21	-1.35	0.73	1.71	-1.86	-1.16	0.56
8	1.50	4	0.83	1.52	-2.49	-1.41	0.30	1.38	-2.37	-1.48	0.29
9	2.70	4	0.61	3.48	-1.64	-1.08	0.07	3.16	-1.86	-1.13	-0.06
10	4.40	4	0.36	3.96	-1.79	-1.24	0.19	4.22	-1.54	-1.13	0.04
11	4.60	4	0.33	2.18	-1.17	-0.63	0.20	2.25	-1.22	-0.68	0.41
12	4.40	4	0.36	2.68	-0.94	-0.60	0.19	2.08	-1.24	-0.74	0.30
13	5.10	4	0.28	1.61	-1.53	-1.09	0.05	1.63	-1.81	-1.02	0.03
14	1.00	4	0.91	2.82	-1.66	-1.19	-0.03	2.67	-1.56	-1.10	0.03
15	3.00	4	0.56	1.97	-1.93	-1.14	-0.29	1.88	-2.25	-1.14	-0.14
16	6.50	4	0.16	2.33	-1.55	-1.05	0.04	1.72	-1.99	-1.04	0.13
17	1.10	4	0.90	3.17	-1.43	-0.97	0.04	3.20	-1.32	-0.86	0.03
18	2.70	4	0.60	2.18	-1.77	-1.36	-0.06	2.04	-1.79	-1.33	-0.22
19	2.30	4	0.69	2.14	-1.55	-1.16	-0.05	2.04	-1.78	-1.16	-0.04
20	4.60	4	0.34	2.13	-0.86	-0.49	0.53	1.79	-1.06	-0.43	0.59

Note. Reference group = non-mobile, focal group = smartphone-MO. a, b_1, b_2, b_3, are estimated item parameters.
*p < .05

Table 18

Step 3 for Conscientiousness Non-Mobile versus Smartphone-MO Using Items 6, 9, 10, 14, 17 as Anchors

				Non-Mobile				Smartphone-MO			
Item	χ2	*df*	*p*	*a*	*b*1	*b*2	*b*3	*a*	*b*1	*b*2	*b*3
1	2.80	4	0.59	2.27	-1.53	-0.88	0.77	1.77	-1.61	-0.97	0.84
2	0.70	4	0.95	1.91	-2.28	-1.82	-0.35	1.76	-2.48	-1.91	-0.15
3	1.70	4	0.79	1.72	-0.95	-0.55	0.96	1.35	-1.10	-0.57	1.09
4	6.80	4	0.15	2.82	-2.19	-1.26	0.11	3.56	-1.54	-1.13	0.02
5	3.80	4	0.43	2.78	-1.77	-1.20	0.41	3.27	-1.38	-1.04	0.22
7	1.10	4	0.90	1.53	-2.21	-1.35	0.73	1.62	-1.92	-1.19	0.63
8	1.50	4	0.82	1.52	-2.48	-1.41	0.30	1.32	-2.45	-1.52	0.34
11	4.10	4	0.39	2.19	-1.17	-0.63	0.20	2.14	-1.25	-0.68	0.46
12	4.30	4	0.37	2.70	-0.93	-0.60	0.19	1.97	-1.28	-0.75	0.35
13	5.20	4	0.27	1.61	-1.53	-1.09	0.05	1.55	-1.87	-1.04	0.06
15	2.30	4	0.69	1.98	-1.92	-1.13	-0.28	1.79	-2.33	-1.16	-0.12
16	5.30	4	0.26	2.34	-1.54	-1.05	0.04	1.64	-2.06	-1.06	0.17
18	2.30	4	0.67	2.18	-1.76	-1.36	-0.06	1.94	-1.85	-1.37	-0.20
19	2.40	4	0.66	2.16	-1.54	-1.16	-0.05	1.94	-1.85	-1.19	-0.01
20	4.50	4	0.34	2.14	-0.86	-0.49	0.53	1.70	-1.08	-0.41	0.65

Note. Reference group = non-mobile, focal group = smartphone-MO. a, b_1, b_2, b_3, are estimated item parameters.

Table 19

Step 1 for Conscientiousness Non-Mobile versus Smartphone-MF

				Non-Mobile				Smartphone-MF			
Item	χ2	*df*	*p*	*a*	*b*1	*b*2	*b*3	*a*	*b*1	*b*2	*b*3
1	0.40	4	0.98	2.26	-1.53	-0.88	0.77	2.05	-1.61	-0.89	0.84
2	1.50	4	0.82	1.91	-2.28	-1.83	-0.35	1.86	-2.57	-2.01	-0.27
3	4.40	4	0.35	1.72	-0.96	-0.56	0.96	1.62	-1.06	-0.42	0.89
4	4.40	4	0.36	2.84	-2.18	-1.26	0.11	3.53	-1.71	-1.10	0.01
5	7.50	4	0.11	2.79	-1.77	-1.20	0.42	2.68	-1.76	-1.02	0.21
6	4.70	4	0.32	2.87	-1.93	-1.57	-0.14	2.21	-2.32	-1.64	-0.23
7	5.30	4	0.26	1.53	-2.21	-1.35	0.73	1.66	-1.95	-1.26	0.35
8	1.70	4	0.79	1.52	-2.49	-1.41	0.30	1.35	-2.82	-1.37	0.30
9	2.30	4	0.68	3.48	-1.64	-1.08	0.07	3.46	-1.70	-1.09	-0.09
10	2.60	4	0.63	3.96	-1.79	-1.24	0.19	3.44	-1.79	-1.26	0.07
11	4.40	4	0.35	2.18	-1.17	-0.63	0.20	2.59	-0.90	-0.53	0.30
12	1.70	4	0.79	2.68	-0.94	-0.60	0.19	2.49	-1.06	-0.63	0.30
13	3.20	4	0.52	1.61	-1.53	-1.09	0.05	1.77	-1.57	-0.95	0.16
14	2.30	4	0.68	2.82	-1.66	-1.19	-0.03	3.33	-1.46	-0.98	0.02
15	1.50	4	0.83	1.97	-1.93	-1.14	-0.29	1.61	-2.10	-1.18	-0.20
16	4.90	4	0.30	2.33	-1.55	-1.05	0.04	1.68	-1.97	-1.29	0.25
17	3.70	4	0.45	3.17	-1.43	-0.97	0.04	2.58	-1.48	-1.06	-0.07
18	5.70	4	0.23	2.18	-1.77	-1.36	-0.06	2.25	-1.85	-1.12	-0.08
19	1.10	4	0.89	2.14	-1.55	-1.16	-0.05	2.05	-1.61	-1.10	0.05
20	5.40	4	0.25	2.13	-0.86	-0.49	0.53	1.71	-0.90	-0.41	0.47

Note. Reference group = non-mobile, focal group = smartphone-MF. *a*, b_1, b_2, b_3, are estimated item parameters.

Table 20

Step 3 for Conscientiousness Non-Mobile versus Smartphone-MF Using Items 4, 6, 9, 10, 17 as Anchors

				Non-Mobile				Smartphone-MF			
Item	χ2	*df*	*p*	*a*	*b*1	*b*2	*b*3	*a*	*b*1	*b*2	*b*3
1	0.60	4	0.96	2.26	-1.53	-0.88	0.78	1.92	-1.63	-0.86	0.98
2	1.70	4	0.79	1.92	-2.27	-1.82	-0.35	1.74	-2.65	-2.06	-0.20
3	4.60	4	0.33	1.72	-0.95	-0.55	0.96	1.52	-1.04	-0.36	1.04
5	6.20	4	0.18	2.78	-1.77	-1.20	0.42	2.51	-1.78	-1.00	0.31
7	2.70	4	0.61	1.52	-2.21	-1.35	0.73	1.56	-1.99	-1.26	0.46
8	1.90	4	0.75	1.52	-2.49	-1.41	0.30	1.27	-2.92	-1.37	0.41
11	4.90	4	0.30	2.20	-1.17	-0.63	0.19	2.43	-0.87	-0.48	0.41
12	1.80	4	0.77	2.70	-0.93	-0.60	0.19	2.33	-1.04	-0.58	0.41
13	3.50	4	0.48	1.62	-1.52	-1.08	0.05	1.66	-1.59	-0.92	0.25
14	2.70	4	0.61	2.83	-1.66	-1.18	-0.03	3.11	-1.47	-0.96	0.11
15	1.50	4	0.82	1.98	-1.92	-1.13	-0.29	1.52	-2.15	-1.17	-0.12
16	3.90	4	0.42	2.34	-1.54	-1.05	0.04	1.58	-2.01	-1.29	0.35
18	5.80	4	0.21	2.19	-1.76	-1.35	-0.07	2.11	-1.88	-1.11	0.00
19	1.40	4	0.85	2.16	-1.54	-1.16	-0.05	1.92	-1.63	-1.09	0.15
20	5.70	4	0.23	2.14	-0.86	-0.49	0.52	1.60	-0.87	-0.34	0.59

Note. Reference group = non-mobile, focal group = smartphone-MF. a, b_1, b_2, b_3, are estimated item parameters.

Table 21

Step 1 for Conscientiousness Smartphone-MO versus Smartphone-MF

				Smartphone-MO				Smartphone-MF			
Item	$\chi2$	*df*	*p*	*a*	*b*1	*b*2	*b*3	*a*	*b*1	*b*2	*b*3
1	1.80	4	0.76	1.72	-1.64	-0.98	0.89	1.88	-1.69	-0.91	0.97
2	0.80	4	0.94	1.70	-2.55	-1.96	-0.14	1.71	-2.74	-2.13	-0.24
3	2.50	4	0.64	1.31	-1.11	-0.57	1.15	1.49	-1.09	-0.41	1.03
4	2.30	4	0.68	3.46	-1.57	-1.14	0.04	3.24	-1.80	-1.14	0.06
5	7.80	4	0.10	3.17	-1.41	-1.05	0.25	2.47	-1.85	-1.05	0.28
6	2.40	4	0.66	2.41	-2.41	-1.74	-0.10	2.03	-2.46	-1.72	-0.19
7	1.80	4	0.77	1.57	-1.97	-1.21	0.67	1.53	-2.06	-1.32	0.44
8	3.40	4	0.49	1.27	-2.52	-1.55	0.37	1.24	-3.02	-1.43	0.39
9	0.60	4	0.96	2.90	-1.97	-1.18	-0.01	3.19	-1.79	-1.12	-0.04
10	1.90	4	0.75	3.88	-1.62	-1.18	0.09	3.17	-1.90	-1.31	0.13
11	5.20	4	0.27	2.07	-1.28	-0.68	0.50	2.38	-0.92	-0.52	0.39
12	1.60	4	0.81	1.91	-1.30	-0.75	0.38	2.29	-1.09	-0.63	0.39
13	1.80	4	0.77	1.50	-1.91	-1.05	0.08	1.63	-1.66	-0.97	0.23
14	1.90	4	0.76	2.45	-1.65	-1.14	0.09	3.06	-1.53	-1.01	0.08
15	3.10	4	0.55	1.73	-2.39	-1.18	-0.10	1.48	-2.23	-1.23	-0.16
16	4.80	4	0.31	1.59	-2.11	-1.08	0.20	1.54	-2.09	-1.35	0.32
17	3.40	4	0.49	2.94	-1.38	-0.88	0.09	2.37	-1.55	-1.09	-0.02
18	4.50	4	0.34	1.88	-1.89	-1.39	-0.19	2.07	-1.95	-1.16	-0.03
19	1.20	4	0.88	1.88	-1.89	-1.21	0.00	1.88	-1.70	-1.14	0.12
20	3.30	4	0.51	1.65	-1.10	-0.41	0.69	1.57	-0.92	-0.38	0.57

Note. Reference group = smartphone-MO, focal group = smartphone-MF. a, b_1, b_2, b_3, are estimated item parameters.

Table 22

Step 3 for Conscientiousness Smartphone-MO versus Smartphone-MF Using Items 4, 5, 9, 10, 17 as Anchors

				Smartphone-MO				Smartphone-MF			
Item	χ2	*df*	*p*	*a*	*b*1	*b*2	*b*3	*a*	*b*1	*b*2	*b*3
1	2.20	4	0.71	1.72	-1.64	-0.98	0.89	2.07	-1.55	-0.83	0.88
2	0.50	4	0.98	1.71	-2.54	-1.95	-0.14	1.88	-2.51	-1.95	-0.22
3	3.30	4	0.51	1.31	-1.11	-0.57	1.15	1.64	-1.00	-0.37	0.93
6	2.60	4	0.64	2.42	-2.40	-1.74	-0.10	2.23	-2.25	-1.58	-0.18
7	1.50	4	0.82	1.57	-1.97	-1.21	0.67	1.67	-1.89	-1.21	0.40
8	3.50	4	0.47	1.28	-2.51	-1.55	0.37	1.36	-2.76	-1.31	0.35
11	7.30	4	0.12	2.07	-1.27	-0.68	0.50	2.61	-0.85	-0.48	0.35
12	3.10	4	0.54	1.91	-1.30	-0.75	0.38	2.51	-1.00	-0.58	0.35
13	2.70	4	0.62	1.50	-1.91	-1.05	0.08	1.79	-1.52	-0.89	0.20
14	3.20	4	0.53	2.46	-1.64	-1.14	0.09	3.34	-1.40	-0.93	0.07
15	3.20	4	0.52	1.73	-2.39	-1.18	-0.10	1.62	-2.04	-1.13	-0.15
16	4.40	4	0.35	1.59	-2.11	-1.08	0.20	1.69	-1.91	-1.23	0.29
18	5.40	4	0.25	1.87	-1.89	-1.39	-0.19	2.26	-1.79	-1.07	-0.03
19	2.00	4	0.74	1.88	-1.89	-1.21	0.01	2.06	-1.55	-1.05	0.10
20	4.10	4	0.40	1.65	-1.09	-0.41	0.69	1.72	-0.85	-0.36	0.51

Note. Reference group = smartphone-MO, focal group = smartphone-MF. a, b_1, b_2, b_3, are estimated item parameters.

Table 23

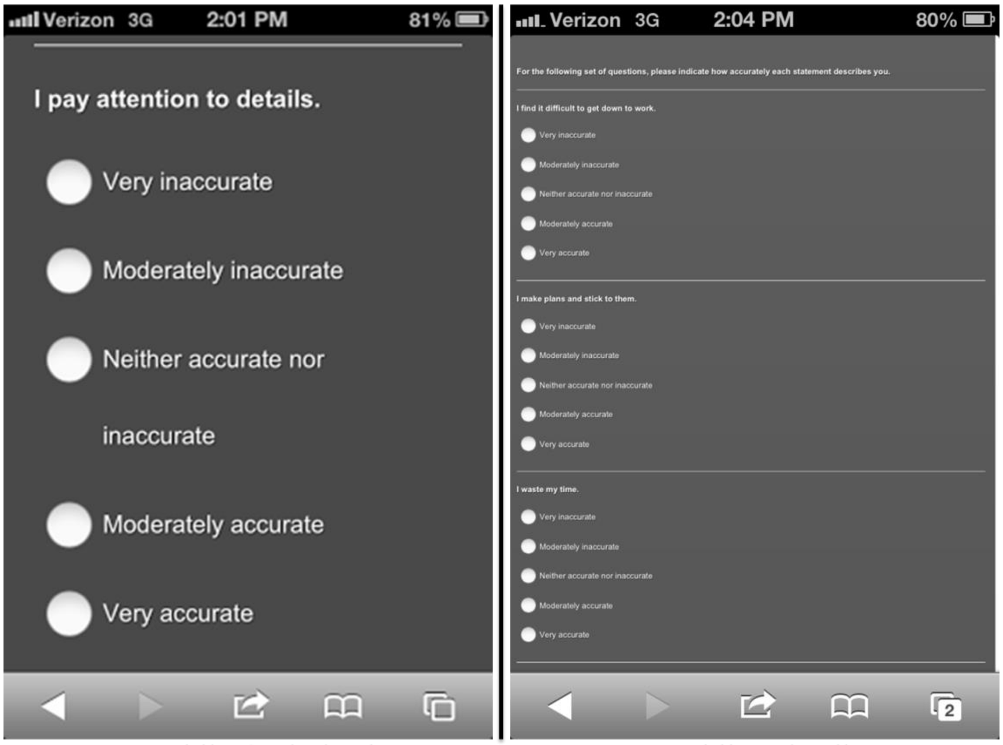
Mean and SD for Conscientiousness and the APM

Descriptives	Non-Mobile	Smartphone-MO	Smartphone-MF	Total
APM				
n	258	209	156	623
Mean	5.46	5.46	6.01	5.60
SD	3.01	2.99	3.32	3.09
Range	0-12	0-12	0-12	0-12
Conscientiousness				
n	261	217	214	692
Mean	82.09	81.81	81.61	81.85
SD	14.39	13.19	13.93	13.86
Range	27-100	37-100	35-100	27-100

Table 24

Changes in Sample Size by Group

Group	Starting Sample	Reduced Sample	Final Sample
Non-Mobile	286	271	261
Smartphone-MO	295	249	217
Smartphone-MF	362	252	215
Total	943	772	693



I pay attention to details.

○ Very inaccurate

○ Moderately inaccurate

○ Neither accurate nor inaccurate

○ Moderately accurate

○ Very accurate

Mobile-Optimized

For the following set of questions, please indicate how accurately each statement describes you.

I find it difficult to get down to work.

○ Very inaccurate

○ Moderately inaccurate

○ Neither accurate nor inaccurate

○ Moderately accurate

○ Very accurate

I make plans and stick to them.

○ Very inaccurate

○ Moderately inaccurate

○ Neither accurate nor inaccurate

○ Moderately accurate

○ Very accurate

I waste my time.

○ Very inaccurate

○ Moderately inaccurate

○ Neither accurate nor inaccurate

○ Moderately accurate

○ Very accurate

Mobile-Friendly

Figure 1. Screen captures illustrating the difference in survey website format for the smartphone-MO and smartphone-MF groups, viewed on an iPhone.

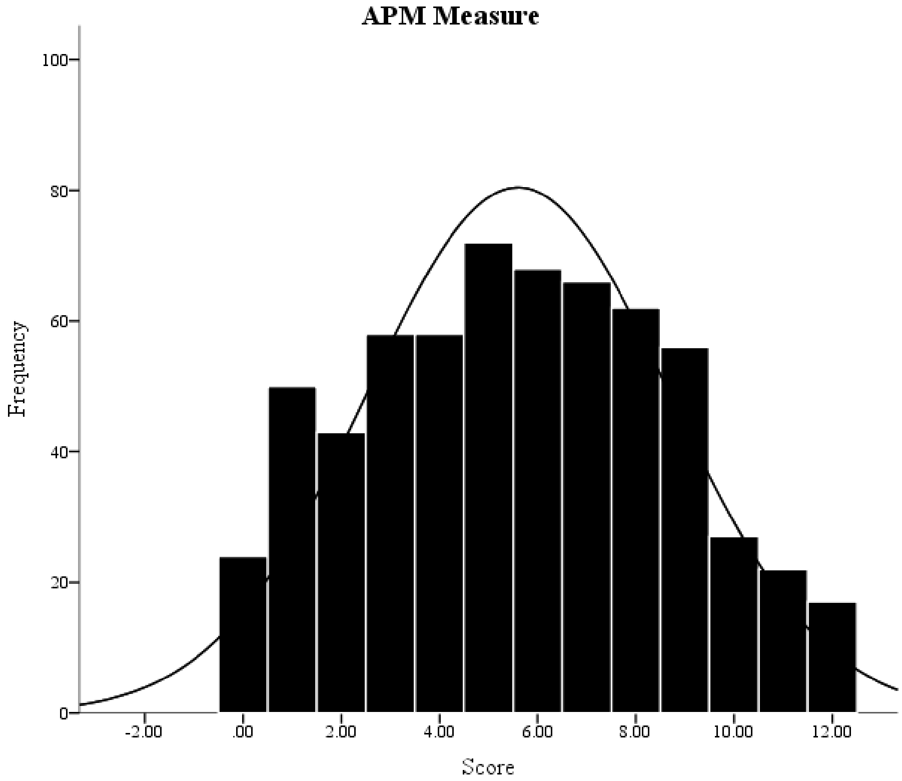


Figure 2. Histogram of the distribution of participant scale scores for the APM (*n* = 623). A normal distribution line is included.

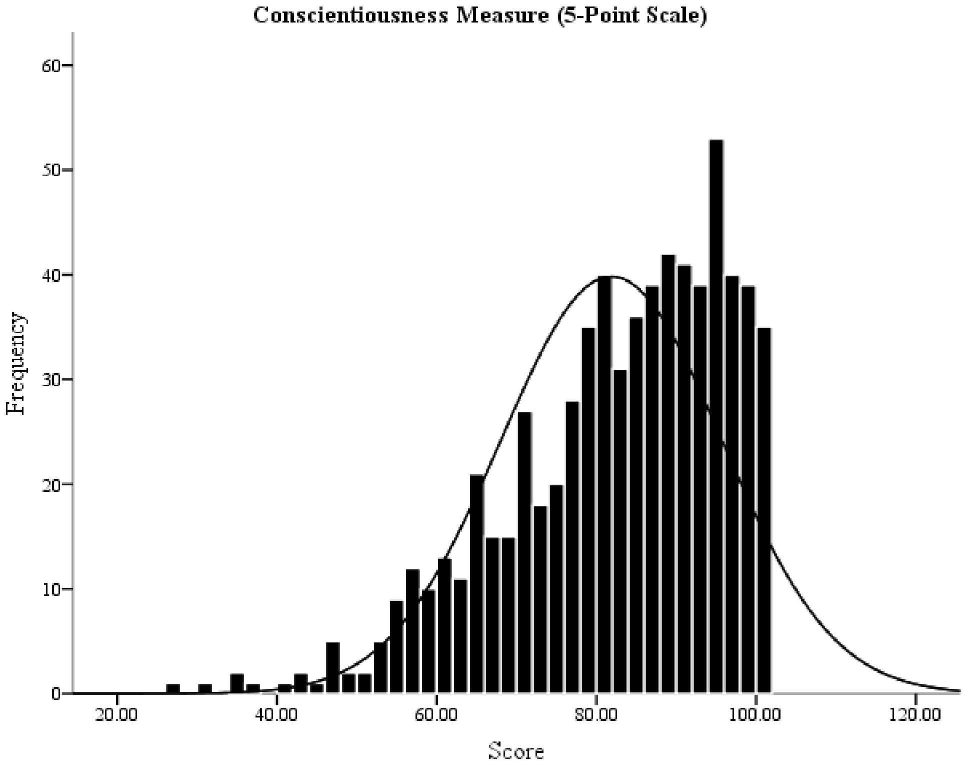


Figure 3. Histogram of the distribution of participant scale scores for the conscientiousness measure (*n* = 692). A normal distribution line is included.

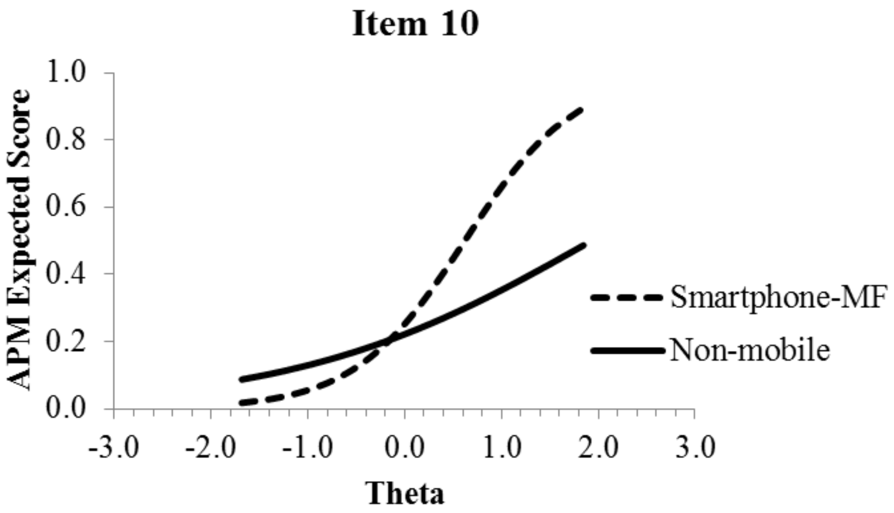


Figure 4. Expected item score plots for Item 10 of the APM, comparing the non-mobile and smartphone-MF groups.

APPENDICES

Appendix A

Raven's Advanced Progressive Matrices Example Item

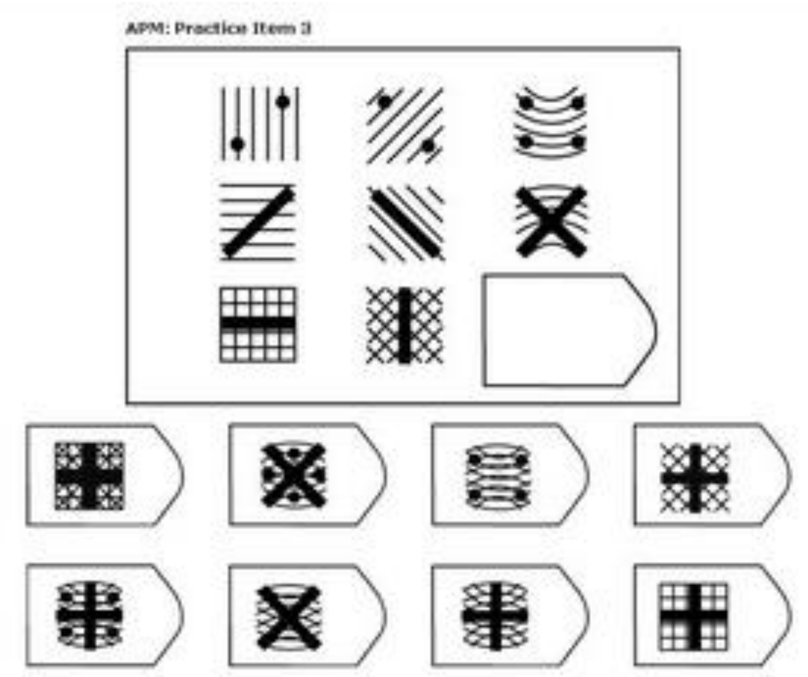


Image Source: http://www.ravensprogressivematrices.com

Appendix B

Conscientiousness items

Positively Keyed	Negatively Keyed
I am always prepared.	I waste my time.
I pay attention to details.	I find it difficult to get down to work.
I get chores done right away.	I do just enough work to get by.
I carry out my plans.	I don't see things through.
I make plans and stick to them.	I shirk my duties.
I complete tasks successfully.	I mess things up.
I do things according to a plan.	I leave things unfinished.
I am exacting in my work.	I don't put my mind on the task at hand.
I finish what I start.	I make a mess of things.
I follow through with my plans.	I need a push to get started.

5-point scale: (1) Very inaccurate, (2) Moderately inaccurate, (3) Neither accurate nor inaccurate, (4) Moderately accurate, (5) Very accurate

Appendix C

Qualifying questionnaire for participants

Question	Response Required for Participation
Are you currently employed or have you been employed in the past six months?	Yes
Do you own a cell phone with internet capabilities (e.g., smartphone, iPhone, Blackberry)?	Yes
Have you ever used your cell phone to go online?	N/A
Do you currently reside in the United States?	Yes

Note. Qualtrics (an online software survey builder) was used to create and administer the online survey. Participants first completed this short qualifying questionnaire which restricted the sample to those with work experience and access to both a computer and smartphone. After completing this questionnaire, those that qualified received a link to the main survey. The main survey used different links for each of the three conditions. Those in the non-mobile condition received a link to the survey, which functioned like a typical online survey. Those in the smartphone-MO condition received the same link, but the Qualtrics software auto-detected the use of a smartphone and reformatted the survey page so that it was mobile-optimized. Qualtrics does not have a specific method for creating a mobile-friendly version of the survey. (Most, if not all, survey software does not have an option to "turn-off" the reformatting feature for smartphones.) In order to get around this, the mobile-friendly survey was embedded in a website that did not automatically reformat for mobile devices. Those in the smartphone-MF condition were provided a link that redirected to the mobile-friendly website, which displayed the desktop version of the survey (i.e., the same format as viewed by the non-mobile device group).

Appendix D

Smartphone survey format manipulation check

Question	Response Options
At any point in the survey, did you enlarge (i.e., "zoom in") the text to better read or answer questions?	Yes/No
At any point in the survey, did you have to scroll left or right to better ready or answer questions?	Yes/No
Did you think the navigation buttons (i.e., the "next page" arrow on the survey, the multiple choice response options) were difficult to use?	Yes/No

Appendix E

DISSERTATION PROPOSAL:

Smartphones in Selection: Exploring Measurement Invariance using Item Response Theory

Brandy N. Parker

North Carolina State University

Abstract

The use of mobile devices (e.g., smartphones) by applicants when completing assessments is a growing phenomenon in the area of selection. Like the transition from paper-and-pencil to online testing, research is needed in order to understand whether measurement invariance holds across device types and website formats. The aim of this proposal is to examine smartphones and non-mobile devices on their equivalence of the psychometric properties for two measures used in selection. More specifically, this study will employ an experimental design and, using item response theory likelihood ratio tests, explore whether measurement invariance holds for both a cognitive ability measure and a personality measure across three formats: non-mobile, mobile-friendly, and mobile-optimized.

Smartphones in Selection: Exploring Measurement Invariance using Item Response Theory

Over the past several decades, technology has greatly impacted the field of I/O psychology. The area of selection in particular has seen dramatic changes as a result of technological developments. The computer brought forth a more efficient way of processing applicants (McBride, 1998); next followed unproctored internet testing, resulting in cost savings and expansion of the applicant pool (Tippins, 2009). With the continued advances in technology people can now carry computers in their pockets, as many cell phones come equipped with internet connectivity (i.e., smartphones), allowing potential applicants to browse and apply for jobs whenever and wherever they would like.

According to research conducted by the Pew Research Center's Internet and American Life project, 91% of American adults own a cell phone of some kind and 56% own a smartphone (Smith, 2013). Though currently there is no specific information regarding the percentage of cell phone owners who use their device to search or apply for jobs, there is evidence to suggest that this occurs. Some organizations have started tracking the operating system and browser types used by online applicants and they are finding that anywhere from less than 1% to 14% of applicants are using a mobile device (i.e., any portable device with a limited operating system and internet connectivity, such as a smartphone or tablet) to apply for jobs (for examples see Doverspike, Arthur, Taylor, & Carr, 2012; Impleman, 2013; Lawrence, Wasko, Delgado, Kinney, Wolf, 2013; Morelli, Illingworth, Moon, Scott, & Boyd, 2013).

Whether or not organizations are anticipating or prepared for applicants to use smartphones and other mobile devices, the number of mobile device applicants is only expected to grow. In 2011, Fallaw and Kantrowitz (2013) found that 9% of human resource professionals surveyed reported they had candidates request to complete application forms and/or assessments on their mobile device; this number increased to 19% in 2012 and 23% in 2013. Golubovich and Boyce (2013) saw an increase in mobile device use from 3.1% of applicants in 2009 to 14.3% in 2013. Additionally, human resource professionals are growing more interested in testing applicants via mobile device (Fallaw & Kantrowitz, 2013; Fallaw, Kantrowitz, & Dawson, 2012).

Like the transition from paper-and-pencil to internet testing, research is needed to determine if applicant test scores are comparable between mobile and non-mobile devices. This paper aims to add to the small but growing body of research around mobile device use in selection. More specifically, this paper will compare smartphones and non-mobile devices on their equivalence of the psychometric properties for both a cognitive ability measure and a personality measure.

Mobile Device Applicants

Organizations are seeing diversity of both race and gender with mobile device applicants. In their sample from a large organization in the restaurant/retail category, Golubovich and Boyce (2013) reported that higher proportions of African-American and Hispanic applicants were using mobile devices (including both smartphones and tablets) to apply for jobs than White applicants. This finding was consistent from 2009 through 2013.

Additionally, there were more female applicants using a mobile device to apply than male applicants. This trend was also observed in applicant data from four organizations in the hospitality industry (Impelman, 2013). Impelman (2013) found that with hourly positions, most mobile device applicants were African-America (50%); 66% of mobile device applicants were female. Though Doverspike and colleagues (2013) found that the majority of their applicant data was from Whites (regardless of device used), they did note slightly higher proportions of African-American and Hispanic applicants in the mobile device user category. They also found that the majority of mobile device applicants were female (59%).

Smartphone Use

These trends are not surprising given the research on smartphone ownership and internet use. According to findings from the Internet and American Life Project, minorities are less likely than Whites to have broadband internet; only 49% of African-Americans and 51% of Hispanics have high-speed broadband connection at home, compared to 66% of Whites (Zickuhr & Smith, 2012). Part of the reason for the low percentages can be attributed to cost, as the biggest demographic differences center around household income and education (Zickuhr & Smith, 2012). It seems the smartphone has helped to address this disparity. Of those who were surveyed, 64% of African-Americans and 60% of Hispanics reported that they own a smartphone (Smith, 2013). These two minority groups are also more active online than Whites with regard to accessing the internet using a phone; sixty percent of African-American cell phone owners and 66% of Hispanics cell phone owners reported that they use their phone to access the internet, compared to only 52% of Whites

(Duggan & Rainie, 2012). Furthermore, 38% of African-American reported that they go online mostly using their smartphone (Smith, 2013).

It is plausible that those of lower socioeconomic status and education use a smartphone as their primary means of accessing the internet; this certainly could include searching and applying for jobs. In the U.S., ethnicity and race are linked to socioeconomic status (APA Task Force on Socioeconomic Status, 2007). House and Williams (2000) found that race/ethnicity correlates with almost every indicator of a person's socioeconomic status. It can be more cost-effective to own a smartphone (which allows for making calls, accessing email, the internet, playing games, etc.) than to own a regular cell phone (or landline), a computer, and pay for broadband internet. It would logically follow that higher proportions of minorities are using their smartphones to apply for jobs, as can be seen in recent research on mobile device use in selection. Thus, offering testing via smartphone may help organizations increase the diversity of their applicant pools.

Mobile Devices in Selection

Research is just beginning on the use of smartphones and other mobile devices in selection. A literature search revealed that there have been no journal publications focused specifically on the use of mobile devices in job applicant testing; the limited research in this area has come from I/O practitioners, some collaborating with academics, presenting their findings at the Society for Industrial and Organizational Psychology (SIOP) conference. Though much of the research has focused on understanding mobile device use (e.g., Gutierrez & Meyer, 2013) and capturing the demographics of mobile device applicants (e.g., Golubovich & Boyce, 2013), a few studied have examined test invariance (e.g., Morelli et al., 2013) and performance differences (e.g., Doverspike et al., 2012) across mobile and non-mobile applicants.

Data collected by various organizations imply that applicants have been using mobile devices for at least the past few years; Golubovich and Boyce (2013) noted mobile applicant data from as early as 2009. Despite the limited research, organizations are moving forward with making applications accessible via smartphone. In their survey of HR professionals, Fallaw and Kantrowitz (2013) found that approximately 40% of respondents indicated they would allow applicant testing via mobile device (if the option existed). Organizations like Aon Hewitt ("Mobile Enhanced Assessments," n.d.) and PeopleAnswers ("PeopleAnswers Launches Mobile App," 2012) are already offering mobile-compatible assessments. But organizations should proceed with caution; the psychometric properties of a scale used across different mediums of administration, such as a computer and a smartphone, should be examined in order to determine whether the scale is functioning the same way, capturing the same attribute, across formats.

Measurement invariance (MI) is the degree to which, under different conditions or formats, measurement operations yield identical measures of the same construct (Horn & McArdle, 1992). If a test/scale is invariant, persons having equal standing on a latent trait should have equal probability of obtaining the same observed score, regardless of being from different samples or groups (Meade & Wright, 2012). If there is a lack of MI (i.e., differential functioning, DF), any findings of differences across groups or individuals cannot be reliably interpreted. In an area like selection, it is critical to know whether an applicant test is invariant across formats; hiring decisions are made, in part, based on individual scores. If DF is suspected for a measure used across different methods of administration, then the psychological constructs cannot be assumed to be identical (Meade, Michels, & Lautenschlager, 2007). To date, there have been very few studies that compared the psychometric properties of assessments completed on mobile and non-mobile devices (Illingworth et al., 2013; Lawrence et al., 2013; Mitchell & Blair, 2013; Morelli et al., 2012; Morelli et al., 2013).

Cognitive ability measures with mobile devices. There have been a few recent studies that utilized mobile (i.e., smartphones and tablets) applicant data from cognitive ability measures (Doverspike et al., 2012; Hawke, 2013; Impelman, 2013); however, findings focused only on performance differences between mobile and non-mobile applicants. Doverspike and colleagues (2012), examined performance differences on a general mental ability measure (GMA) comprised of both a verbal and numerical component. Over one million job applicants were included in their study, with applicants free to use the device of their choosing; approximately 1.7% of applicants used a mobile device. The authors found that mobile device users had significantly lower performance scores than non-mobile users on overall GMA, GMA verbal, and GMA numerical. Impelman (2013) reported similar findings. Using data from management position applicants across four organizations in the hospitality industry, he found that the 2.8% of applicants who completed a cognitive ability measure via mobile device performed worse than those using a non-mobile device.

Interestingly, differences between mobile and non-mobile applicants in cognitive ability scores were less pronounced among racial minorities; results were mixed for gender.

While these findings imply that mobile applicants are at a disadvantage when completing a cognitive ability assessment, there are no studies that have examined whether MI exists across mobile and non-mobile devices for cognitive ability measures. It is possible that the performance differences observed by both Doverspike et al. (2012) and Impelman (2013) could be attributed to a lack of MI. As stated by Vandenberg and Lance (2000), "violations of measurement equivalence assumptions are as threatening to substantive interpretations as is an inability to demonstrate reliability and validity" (p. 6).

Non-cognitive measures with mobile devices. Much of the available research on applicant testing via mobile devices has focused on personality and other types of non-cognitive measures. Unlike the research on cognitive ability measures, I/O psychologists have already begun to test for MI of non-cognitive measures across device types (Illingworth et al., 2013; Lawrence et al., 2013; Mitchell & Blair, 2013; Morelli et al., 2012; Morelli et al., 2013). Morelli and colleagues (2012) collected data from over 900,000 customer support job applicants on five personality constructs: conscientiousness, customer service, integrity, interpersonal skill, stress tolerance, and teamwork, using both a Likert-type scale and biodata. Using multiple-group confirmatory factor analysis (MGCFA), the authors found the measures of conscientiousness, customer service, integrity, and teamwork to be invariant across devices, except for construct means. These findings were later replicated and extended by Morelli et al (2013). Using data from 664,469 online applicants for a retail sales position, the authors conducted MGCFAs to examine MI for three measures: conscientiousness, curiosity, and customer service. Based on several iterations of model fit, there did not appear to be any differences between mobile and non-mobile users. The authors also found no practically significant performance differences across devices.

Lawrence and colleagues (2013) examined data collected from nearly 200,000 retail candidates on several personality and situational judgment measures: attention to detail, stress tolerance, productivity, likelihood for absenteeism, likelihood for turnover, service potential and sales potential. Eight percent of the applicants in their sample used a mobile device. Using MGCFA, they found no meaningful differences in model fit when comparing mobile and non-mobile devices, nor did they find meaningful performance differences.

Smartphone Formatting and Usability

Though the current research implies that personality measures are invariant across mobile and non-mobile devices, there is value in knowing the conditions under which MI will hold. Illingworth and colleagues (2013) explored whether non-cognitive measures were invariant across different device browsers and operating systems. Data came from 660,269 online applicants for a retail sales position. The authors used MGCFA to explore whether conscientiousness, openness, and customer service measures were invariant across five browser types and five operating systems. Results suggested that all three measures were invariant across all operating systems and browser types.

Research on the effects of small displays (like that of a smartphone) on information processing supports the idea that I/O psychologists need to explore MI of mobile devices if organizations are to move forward with their use. With smaller displays, textual information often flows across multiple screens, requiring the user to scroll in order to read the entire text (Albers & Kim, 2002). Even when websites are intentionally designed to better display information for a mobile platform, there is still difficulty in displaying all information, in a readable format, on a single screen (Sanchez & Branagahn, 2011). Sanchez and Branagahn (2011) hypothesized that the restrictions of a smaller screen would affect an individual's ability to reason using the information displayed on the screen. They had participants read several emails containing information that was necessary in order to correctly answer a short multiple choice test and found that compared to a full-size display, reasoning deficits occurred when using a small display.

One explanation for these deficits is the effect of scrolling. A small display usually necessitates scrolling in order to read all the textual information available. If an individual is trying to retain information, scrolling can be taxing because there is a level of stress in maintaining large amounts of information in short-term memory (STM; Albers & Kim, 2002). "Limitations of human STM determine how much the user can mentally process while moving from screen to screen. The handheld's small screen size requires people to hold more information in STM for longer periods of time so they can compare information" (Albers & Kim, 2002, p. 52).

In addition to scrolling, Sanchez and Goolsbee (2010) examined whether character size is responsible for these reasoning deficits. They explained that if characters are too small, it becomes difficult to distinguish the letters and numbers, leading to perceptual jumbling of the characters. The jumbling might increase processing load for the reader, which would take resources away from things like retaining information. They found that when reading from a small device, the interaction between screen size and characters size used to portray textual information can result in reduced reading performance and comprehension. When scrolling was kept to a minimum, factual recall was found to be equivalent to that of a full-size display. Sanchez and Branaghan (2011) found similar results when they had participants read emails. When participants used a small display with vertical (portrait) orientation, which necessitated scrolling, they performed worse on a multiple choice recall test than those using a large display; however, when the orientation was horizontal (landscape), performance decrements were eliminated. This was because scrolling was reduced.

Organizations realize the importance of having a website that is easy to view and use on a smartphone. According to Latitude (a research company), 61% of surveyed mobile device users (*N* = 909) said they had a better opinion of a brand when that company offered a good mobile experience (Latitude, 2012). Many IT and marketing companies are shifting focus to mobile users as more people browse websites and shop using their smartphones or tablets. Currently, most companies consider a website to be “mobile-friendly” if it is accessible from a mobile device (e.g., Gallizzi, 2013; “Mobile Friendly vs Mobile Optimized,” 2012). A mobile-friendly website looks identical across devices, but the smaller screen of the smartphone means users must scroll from left to right or zoom to better view the webpage. While a mobile-friendly website is functional, it is not ideal. To create a better user experience, many companies and organizations will create a mobile-optimized website. Websites that are mobile-optimized do not require zooming or scrolling left and right and will often have larger navigation buttons. Mobile-optimized websites can either be existing websites that auto-detect mobile devices and reformat accordingly ("Mobile Friendly vs Mobile Optimized," 2012) or websites that are specifically designed for a smartphone or tablet (Gallizzi, 2013). Mobile-optimized websites are considered easier to use and navigate.

Organizations that currently allow or are considering smartphone-based assessments should be aware of the differences in mobile website design. If selection assessments are not optimally formatted, it is possible that the aforementioned effects could result in DF. Though Illingworth and colleagues (2013) found that MI held for non-cognitive measures across browser type and operating system, there is more to consider. Of the available research on mobile devices in selection, it is unclear whether the assessments used in various studies were hosted on websites that reformatted specifically for a mobile device (i.e., mobile-optimized) or if they were simply accessible via mobile device (i.e., mobile-friendly). It is plausible that poorly formatted assessments, like those on mobile-friendly websites, would results in DF when compared with either mobile-optimized assessments or non-mobile assessments.

There is still much research to be done in order to understand the effects of using smartphones in applicant testing. Current mean-difference studies, in particular those looking at cognitive ability measures, have assumed MI across devices; thus there is a definite need to identify whether DF exists across devices for cognitive ability measures.

Furthermore, studies that have tested for DF have not explicitly examined the effect of smartphone website format and only two studies (Morelli et al., 2012 and Morelli et al., 2013) have differentiated among mobile device types. The current study seeks to address these gaps by examining whether MI holds for a cognitive ability measure and a personality measure across three device categories: non-mobile, smartphone with mobile-optimized website (smartphone-MO) and smartphone with mobile-friendly website (smartphone-MF). The following research questions will be investigated:

Research Question 1: For a cognitive ability measure, does MI hold across non-mobile, smartphone-MO, and smartphone-MF conditions?

Research Question 2: For a personality measure, does MI hold across non-mobile, smartphone-MO, and smartphone-MF conditions?

This study will make a unique contribution to the literature on mobile devices in selection by examining MI for a cognitive ability measure, which has not yet been done. By employing random assignment, this study will parse out the effects of respondent characteristics associated with device choice. The use of random assignment improves upon previous research because if any DF is detected, it can be attributed to the conditions of the study, rather than possibly to the characteristics of the mobile device users. Furthermore, this study will explore conditions (i.e., mobile website compatibility) under which MI may or may not hold. As stated by Meade et al (2007), one goal of MI research "should be to specify guidelines regarding when MI likely would be present or absent" (p. 326).

Method

Sample

Data will be collected from approximately 600 Mechanical Turk users. Mechanical Turk is a crowdsourcing website hosted by the company Amazon. Research by Behrend, Sharek, Meade, and Wiebe (2011) found that participants sourced from Mechanical Turk were more diverse and had more work experience than the traditional participant pool of university students. Additionally, the authors found that the reliability of the data from Mechanical Turk participants was as good as or better than university participants. Potential participants for this study will be restricted to the United States and will be asked if they currently hold a job or have applied for a job in the past six months; those who do not meet these requirements will not be accepted to participate in the study. Participants must also have access to both a smartphone and non-mobile device in order to participate in the study. All participants will receive a payment of $1.00 upon completion of the study.

Measures

Cognitive ability measure. I will assess cognitive ability using the 12-item short form of the Raven Advanced Progressive Matrices Test (APM; Arthur & Day, 1994). The original test, developed by John Raven, is a series of 36 matrix problems that increase in difficulty. For each item, participants are required to select the piece (out of eight options) that completes the pattern (see Appendix A for an example item). The original APM typically takes 40-60 minutes to administer. Arthur and Day (1994) wanted to create a shorter version of the APM that still provided a sound assessment of general intelligence. They created the 12-item APM short form, which demonstrated psychometric properties similar to the original 36-item test (Arthur & Day, 1994; Arthur, Tubre, Paul, & Sanchez-Ku, 1999).

Personality measure. I will assess conscientiousness using a 20-item scale (α = .88) from taken from the International Personality Item Pool (IPIP; see Appendix B). This scale will be administered as part of the IPIP 100-item measure of the Big Five personality constructs (i.e., extroversion, openness to experience, neuroticism, agreeableness, and conscientiousness). I chose conscientiousness because it is one of the most frequently used scale in selection (Meade et al., 2007) and it has been shown to be a valid predictor of job performance criteria (Barrick & Mount, 1991). Additionally, it has previously been examined for MI across formats; Meade et al (2007) found that it exhibited strong MI across paper-and-pencil and computer formats.

Design

This study will employ an experimental design. Participants will be randomly assigned to one of three conditions: non-mobile, smartphone-MO, or smartphone-MF. Non-mobile devices can include desktop or laptop computers and can be either PCs or Macs. For the purposes of this study, smartphones can include both iPhones and Android phones, as well as other cell phones that allows for internet browsing (e.g., Blackberry). Tablet devices (e.g., devices that do not also function as a phone, such as iPads) will not be permitted. In addition to asking participants to report their device, each participant's operating system and browser information will be collected in order to confirm the type of device used.

Qualtrics will be used to create the online surveys. Those in the non-mobile condition will receive a link to the survey, which will function like a typical online survey. Those in the smartphone-MO condition will receive the same link, but the survey software will auto-detect the use of a smartphone and reformat the survey page so that it is mobile-optimized. Qualtrics does not have a specific method for creating a mobile-friendly version of the survey. (Most, if not all, survey software does not have an option to "turn-off" the reformatting feature for smartphones.) In order to get around this, the mobile-friendly condition will require the survey to be embedded in a website that does not automatically reformat for mobile devices. Those in the smartphone-MF group will be redirected to a website that is mobile-friendly and the webpage will display the desktop version of the survey (e.g., the same format as viewed by the non-mobile device group).

Procedure

Participants will first complete a short qualifying questionnaire (Appendix C). Those that qualify will continue on to the actual study. Random assignment will be implemented at the end of the qualifying questionnaire; Qualtrics will display one of two survey links (one non-mobile and smartphone-MO, one for smartphone-MF) and a message indicating that the participant should use either their smartphone or non-mobile device to complete it. The survey will require participants to use the type of device assigned to them; if they do not, they will not receive payment.

At the start of the survey (prior to consent), participants will be told that the survey is meant to assess the quality of data collected using different formats. They will be asked to treat the survey as though it were a job application, taking time to answer each question to the best of their ability. In order to simulate the effects of completing an actual job application and assessment, participants will be told that the top 5% of scorers will receive an additional $0.50.

The survey will take participants approximately 40 minutes to complete. The Raven APM will take approximately 15 minutes; the conscientiousness measure (administered as part of the 100-item IPIP Big Five personality measure) will take approximately 20 minutes. To verify that the smartphone surveys were perceived as being mobile-optimized or mobile-friendly, participants will complete a few questions about how they interacted with the survey (Appendix D). Additionally, the survey will include text entry items to approximate an applicant completing bio-data type questions.

Proposed Analyses

The APM test and the conscientiousness measure will be examined for MI separately using an item response theory (IRT) approach. Prior to testing for DF, both measures will be tested for unidimensionality. I will conduct an exploratory factor analysis and examine the eigenvalues, looking for a clear delineation of one factor. I will then examine IRT model fit using the program MODFIT (Stark, 2001). MODFIT provides graphs of the predicted (or expected) and empirical item characteristic curve (ICC) for each item, and statistical tests of item fit (i.e., χ^2). For the graphs, I will look for whether the predicted ICC falls within the 95% error bars of the empirical ICC. This is done for each item, though with the polytomous data there are ICC graphs for each response option within each item. The χ^2 tests of fit are reported for each item and those that are significant are indicators that the model may not fit the item (though this will be considered in tandem with the predicted/empirical ICC graphs.

Using the IRT likelihood ratio (LR) test, individual items will be examined for DF. The IRT LR test is considered a more desirable method than MGCFA when examining the equivalence of a single scale because more information is available for MI testing (Meade & Lautenschlager, 2004). Specifically, IRT LR test examines both the *a* parameter (the slope of an ICC) and *b* parameter (item location) of each item for DF. Like MGCFA, the LR test uses maximum likelihood estimation to estimate item parameters; however, the IRT approach uses a log-linear model (rather than linear) to describe the relationship between observed item responses and the underlying trait. The LR test generates a fit function, which is an index of how well the model fits the data. This same test can be used on both dichotomous data (i.e., the APM test using a three-parameter model) and polytomous data (i.e., the conscientiousness measure using a graded response model).

The LR test functions similarly to MGCFA, in that it uses nested models: a baseline model to evaluate the comparison model. The most common approach is all others as anchors (AOAA). In AOAA approach, the baseline model has all estimated item parameters for like items constrained to be equal across the two groups. That is to say, the item parameters for Group A’s Item 1 are equal to those of Group B’s Item 1. Each item is tested separately for DF using the comparison models, in which the item parameters for all items are again constrained to be equal across the two groups with the exception of the item of interest; the item parameters of the item being tested are free to vary across the two groups.

The fit function (or likelihood value) produced by both the baseline model and the comparison model are compared and a G^2 value, distributed as χ^2, is calculated. Significant G^2 values indicate that an item exhibits DF.

The issue with the AOAA approach is that all the items, some of which may exhibit DF, are treated as anchors. This problem is inherent in both the MGCFA and AOAA approach and including items with DF as anchors can lead to errors in detecting DF (see Meade & Wright, 2012 for review). An alternative is to use the free baseline approach, in which only a single item serves as an anchor in the baseline model and the comparison model has the single item being tested constrained in addition to the anchor item. But again there is the problem of knowing whether the selected anchor item exhibits DF (see Lopez Rivas, Stark, and Chernyshenko, 2009 for research on selecting anchor items with the free baseline LR test). Meade and Wright (2012) tested several different approaches for selecting anchor items and testing for measurement invariance using IRT. Based on their findings, the authors proposed a series of steps which incorporates both the AOAA approach and the free baseline approach. Table 1 (this Appendix) outlines four steps (adapted from Meade & Wright, 2012) for testing for MI using IRTLR test.

First, I will conduct LR tests with the AOAA approach using IRTPRO. In this step I will identify which items are considered most likely free of DF based on the lack of significant G^2 values. Second, looking at only the non-significant items from Step 1, I will examine the *a* parameters for each item, rank-ordering them. Meade and Wright (2012) recommended selecting the five items with the largest *a* parameters to serve as anchors, which was 25% of the total items in the measure they used. The conscientiousness measure contains 20 items, so I plan to select 5 items to serve as anchors (as long as there are at least 5 items identified from Step 1). The AMP test contains only 12 items, so I plan to select between 3 and 5 items (25% - 42% of the total items) to serve as anchors, based on the results of Step 1.

Next (Step 3), I will conduct LR tests with the free baseline approach, using the anchor items identified in Step 2. In each of these tests I will again evaluate DF significance levels, flagging items with significant G^2 values as exhibiting DF. Finally, in Step 4 I will compute the DF effect size indices using the output from Steps 1 and 3. These effect size indices are important because they provide information on the extent to which items and scales function differently (Meade, 2010); the significance findings from the LR tests (DF items) may not be of practical importance. I will report on six effect size indices: unsigned item difference in the sample (UIDS), signed item difference in the sample (SIDS), expected score standardized difference (ESSD), signed test difference in the sample (STDS), unsigned expected test score difference in the sample (UETSDS), and expected test score standardized difference (ETSSD). Table 2 (adapted from Meade, 2010; this Appendix) provides descriptions for each of these indices. UIDS, SIDS, and ESSD are included because they are item-level effect size indices. STDS, UETSDS, and ETSSD are test-level effect size indices that Meade (2010) suggested should be included regardless of the level of analyses. I will use Meade's (2010) VisualDF program to compute the six effect size estimates.

The IRT LR test only allows for comparisons between two groups: a focal group (typically the minority group in traditional studies examining DF) and a referent group. In order to test for DF across the three groups I will complete the analyses (previously outlined) three times: smartphone-MO versus non-mobile, smartphone-MF versus non-mobile, and smartphone-MF versus smartphone-MO. This will be done for both the AMP test and the conscientiousness measure for a total of six comparisons.

Appendix E, Table 1

Recommended Best-Practice Steps in Conducting IRT Invariance Analyses

Step	Description
1	Conduct invariance analyses via the AOAA approach.
2	Of the non-significant items, rank-order the items by the largest *a* parameters and select approximately 5 items to serve as anchors.
3	Conduct LRTs with a free baseline approach using the anchor items identified in Step 2, evaluating DF significance levels.
4	Compute DF effect size indices using output from Steps 1 and 3.

Note. IRT = item response theory; AOAA = all others as anchors; LR = likelihood ratio; DF = differential functioning. Adapted from "Solving the Measurement Invariance Anchor Item Problem in Item Response Theory," by A. W. Meade and N. A. Wright, 2012, *Journal of Applied Psychology, 97,* p. 1030.

Appendix E, Table 2

IRT Differential Functioning Effect Size Index Descriptions

Index		Description
SIDS	Signed item difference in the sample	Average difference in ESs across focal group respondents as compared to reference group respondents. DF across respondents is allowed to cancel in cases of nonuniform differences in ESs.
UIDS	Unsigned item difference in the sample	The average difference in ESs across focal group sample respondents had differences been uniform in nature. Comparing UIDS and SIDS gives an indication of the extent to which differences in ESs cancel across different respondents.
ESSD	Expected score standardized difference	An ES version of Cohen's *d*. Mean ESs are computed for the focal group respondents using both focal and reference item parameters. The difference between these means are divided by the pooled SD of the two sets of ESs. The metric can be interpreted using the guidelines given by Cohen (1988).
STDS	Signed test difference in the sample	The difference in observed summed scale scores expected, on average, across all focal group respondents, due to DF alone. Allows cancellation of DF across both items and persons.
UETSDS	Unsigned expected test score difference in the sample	The hypothetical difference in expected scale scores that would have been present if scale-level DF had been uniform across respondents. Allows cancellation of DF across items but not persons.
ETSSD	Expected test score standardized difference	An ETS version of Cohen's *d*. The metric can be interpreted using the guidelines on effect size given by Cohen (1988).

Note. In all cases, focal group expected scores are computed using item parameters estimated in both the focal group sample and the reference group sample. Uniform differences in expected scores mean that one group is favored. Nonuniform differences in expected scores indicate that for some respondents, expected scores will be higher for the focal group and for other respondents expected scores will be higher for the reference group. ES = expected score; DF = differential functioning; ETS = expected test score. Adapted from "A Taxonomy of Effect Size Measure for the Differential Functioning of Items and Scales," by A. W. Meade, 2010, *Journal of Applied Psychology, 95,* pp. 732-733.