DOES THE CONFIGURABLE APPROACH TO PERSONALITY TESTING IMPACT MEASUREMENT CHARACTERISTICS? A MEASUREMENT EQUIVALENCE/INVARIANCE ANALYSIS

by

Anne-Marie Winter Shumaker

A thesis submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Master of Arts in
Industrial/Organizational Psychology

Charlotte

2015

Approved by:

_____
Dr. Eric Heggestad

_____
Dr. Linda Shanock

_____
Dr. Scott Tonidandel

ABSTRACT

ANNE-MARIE WINTER SHUMAKER. Does the configurable approach to personality testing impact measurement characteristics? A measurement equivalence/invariance analysis. (Under the direction of DR. ERIC HEGGESTAD)

Computerized personality testing has allowed organizations using scores from personality tests within hiring systems to configure their tests to deliver only items that are directly related to the job. Although configurable personality testing may seem to be a better approach at first glance, research is needed to investigate the psychometric nature of the these tests to determine if restructuring the items yields test scores that are equivalent to those that would be obtained from the full-length assessment. Practitioners and researchers need to be assured that when they use configurable personality tests, they can trust that the results are equivalent to their full-length counterparts. The current study used a within- and between-person design to examine the measurement equivalence of personality scales across different configurations of tests. Results provide some initial evidence for the viability of configurable personality testing. Additional research is needed with larger samples and different personality measures before the configurable personality testing can be recommended in actual employee selection contexts.

ACKNOWLEDGMENTS

To begin, I would like to acknowledge Dr. Eric Heggestad as my thesis advisor and mentor through the years. This effort could not have been completed without his continued support and commitment to me and to the study. In addition, a special thanks to my additional committee members, Dr. Linda Shanock and Dr. Scott Tonidandel, for their counsel and assistance throughout this process.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

In the past, organizations interested in assessing applicant personality characteristics often used "off the shelf," paper-and-pencil, type measures in which applicants were handed a booklet of items and asked to mark their responses. Not customized to the job or the organization, these measures typically included a lot of items, many of which were associated with scales that assessed traits not related to the job. For example, if an organization were to use the California Psychological Inventory because they were interested in using scores from four of the 20 scales to assess job applicants, those applicants would have to respond to all 434 items of the inventory. Clearly, a notable drawback to these measures is that they are inflexible and rigidly structured.

The advent of computer-based testing, however, has opened the door to increased flexibility in personality testing. Personality tests can now be administered and scored quickly and efficiently without having to have any direct contact with the applicant. What's more, test administrators can now choose to give applicants only those items from scales that are directly job-relevant. For example, assume that there exists a battery that is made up of 12 trait scales. Through job analytic methods it might be determined that trait scales A, D, and F should be predictive of performance in Job 1 while trait scales A, C, and E should be predictive of performance in Job 2. In such a situation, applicants could be presented with the items from the scales that are relevant to the job for which they are applying and not from the other non-job-relevant scales. We refer to this approach to personality testing as configurable testing in that the test is configured for a particular job.

Despite the potential advantages of configurable personality testing, we are unaware of any research which examines the psychometric nature of tests administered in this manner. To the extent that different "mixtures" of items within an assessment change the measurement properties of a scale, configurable approaches to personality testing would not be viable. Let us be more specific by considering an example. Let's say, for instance, that in one testing situation the items from scales A, B, C, and D are administered together (such that the test includes an item from Scale A, then an Item from Scale B, etc., as is common with these types of tests) and that in another testing situation only the items from scales A and B are administered. The question is whether or not the items of scales A and B represent constructs A and B in the same way in both of these testing situations. If they do not, and the capability of a set of items to represent a construct is dependent on the other items included in an assessment, then configurable personality testing is not viable.

The purpose of the present research is to explore the measurement equivalence of personality scales across different configurations of tests. When some scales are *removed* from a test to help make it more job-relevant, are the *remaining* scales equivalent to their counterparts in the full-length test? To date, there is no empirical data in the literature that has addressed this question directly. Other analyses and reviews address the possible differences that could arise based on how tests are structured or how items are ordered, but they are not designed to address the *specific* differences that could occur in configurable testing (Knowles, 1988; Leary & Dorans, 1985; Schell and Oswald, 2010). This research will provide an initial indication as to the viability of configurable personality testing in applied situations. The available evidence, presented below,

suggests that changing the configurations of the test will not preclude us from observing measurement equivalence.

Advantages of Configurable Personality Tests

We believe that the configurable approach to personality testing might provide an advantage to the more traditional fixed questionnaire approach. They are shorter and more job-relevant, so applicant reactions could be higher. Also, hiring managers do not have extraneous information about the applicant, so the quality of hiring decisions could be improved.

Applicant Reactions

How an applicant perceives a selection tool can have implications for how positively they view the company (Murphy, 1986), how they discuss the company with others (Smither, Reilly, Millsap, Pearlman, & Stoffey, 1993), and the likelihood that they will accept an offer of employment from the company (Macon, Avedon, Paese, & Smith, 1994). We believe that there are two reasons why configurable personality tests should result in more positive applicant reactions than traditional personality tests. First, a test configured to a particular job will be shorter, requiring less time from applicants (Ackerman & Kanfer, 2009). Second, given that a configurable test will only contain scales that are job-relevant, applicants should perceive the measure as more job relevant and, consequently, result in more positive reactions than traditional measures (Bauer, Truxillo, Sanchez, Craig, Ferrara, & Campion, 2001; Ployhart & Ryan, 1997; Rynes & Connerley, 1993; Smither et al., 1993; Steiner & Gilliland, 1996). Research by Hausknecht, Day, and Thomas (2004) demonstrated that perceived favorability of

selection tools increased when the relationship between the content of the selection tool and the duties of the job were more closely linked.

Quality of Hiring Decisions

Research has shown that having more information than what is actually needed about an applicant can influence judgment processes in selection decisions (Beehr & Gilmore, 1982; Kulik & Clark, 1994). Beehr and Gilmore (1982), for instance, found that perceived applicant physical attractiveness inflated raters' perceptions of the applicant's job abilities. Research by Kulik and Clark (1994) suggested that the features of an applicant's prototypicality and favorability can compensate for one another, such that negative features of the applicant only result in a disadvantage for them when the applicant is perceived as nonprototypical. Providing any non-job-relevant information, especially if it is negative, can impact hiring decisions; therefore, we should work to minimize the influence of extraneous information (Bolster & Springbett, 1961; Carlson & Mayfield, 1967; Springbett, 1958).

By using personality inventories that assess scales or items that are not job-relevant, the door could be left open for job-irrelevant information to influence a hiring manager's decision-making process. Configurable personality inventories would guard against the use of such extraneous information in that only job-relevant scales would be included on the assessment.

Scale Equivalence Across Configurations

The advantages of configurable testing would be worthless if the scales do not operate in the same way across different configurations. That is, if the items measuring a construct represent that construct to a different extent when those items are paired with

differing sets of items (representing different constructs), then configurable testing cannot be recommended for use in hiring systems. To date, there is no empirical data in the literature that has addressed this issue directly. Other analyses and reviews, discussed below, address 1) the possible differences that could arise based on how tests are structured or they are 2) not designed to address the *specific* differences that could occur in configurable testing (Knowles, 1988; Leary & Dorans, 1985; Schell and Oswald, 2010).

Leary and Dorans (1985) conducted a conceptual review of the literature from the 1950s to the 1980s examining the ordering of items within a test. Specifically, they examined studies of power vs. speed tests, studies of achievement vs. aptitude tests, and studies that used different item arrangement strategies. Most of the studies that they examined placed items on a test randomly to create multiple forms. They then administered these different forms and examined whether any between-group differences emerged. Although they concluded that there was evidence of context effects, they carefully noted that the evidence was not substantial enough to suggest that the effects of rearranging test items or sections of tests invalidated the test. It should be noted, however, that their analyses were rather unsophisticated with respect to how we would examine invariance of measurement characteristics today.

Positing that personality tests are unique from other kinds of objective tests in that even small item effects could accumulate to impact the overall results on the test, Knowles (1988) investigated the effects of item position within personality tests. He looked at how much of an item's response is due to the content of the item preceding it. He found that an item's correlation with the other items on the test had a positive linear

relationship with its serial position on the test, such that items that were presented later on in the test had higher correlations with the other items than when the same item was presented toward the beginning of the test. As such, the content of already completed items can affect the responses of subsequent items.

Perhaps the most direct evidence for whether scales will operate differently in distinct scale configurations comes from a recent study by Schell and Oswald (2010). Again examining a personality test (the 50-item International Personality Item Pool version of the Big Five personality instrument), these authors used three different item order strategies: 1) presenting the items in random order, 2) presenting the items grouped by trait (i.e., all of the items for Trait 1 then all of the items for Trait 2, etc.), and 3) an "item/factor rotation" approach in which one item from each trait was presented with a repeating pattern (i.e., item from Trait 1, item from Trait 2, item from Trait 3, item from Trait 1, item from Trait 2, item from Trait 3, etc.). Using a sample size of 397 participants, they found that the internal consistency of the scales was unaffected by the item order. Further, using state of the art measurement equivalence/invariance analysis, they found evidence for measurement invariance across the different versions of their test. These findings indicate that the relationships between the items and the constructs were generally the same regardless of how the test was put together. They did find, however, there was variation in the correlations between the scales depending on which item order was used. In general, the correlations were weaker when the items were grouped by factor.

Although the Schell and Oswald (2010) data suggest that item ordering results in only minimal changes in the construct validity of the scales, it must be recognized that

the exact same items were used in each of the conditions that they included; only the item ordering was changed across the conditions. This would not be the case in configurable testing where some scales would be removed altogether from the test. Each configurable test would have different combinations of scales. Having items from different combinations of scales could potentially change the test such that the items could be responded to in different ways.

Taken together, these three studies provide mixed evidence regarding how the ordering of items can impact the measurement characteristics of a scale. Although the Leary and Dorans (1985) review and the Schell and Oswald (2010) study indicate that changes in item ordering should have little impact on the item characteristics, the research by Knowles (1988), and to some extent that by Schell and Oswald (2010), suggests that changing the ordering of items could impact the way people respond to the items. This mixed evidence and potential for measurement inequivalence directly imply a need for further study. What's more, only the work by Schell and Oswald (2010) used state of the art analytic techniques for examining the relationships between the items and the constructs. It should also be noted that Knowles (1988) compared effects at the item level while Schell and Oswald (2010) focused on scale-level analyses. Using these different levels of analysis also leaves room for potential measurement differences that were not apparent.

Practitioners and researchers need to have assurance that when they use only some items from a larger inventory they will get the same result as if they used the full inventory. Simply "hoping" that the measure will be the same is not enough. We need to establish that the tests are in fact psychometrically equivalent using advanced analytical

techniques available today in order to claim that the scales are actually measuring what we claim they should be measuring.

The current study uses state of the art measurement equivalence/invariance (ME/I) analyses to determine the psychometric and structural equivalence of a traditional personality inventory to two different configurations of scales from that inventory. There are multiple advantages to conducting an ME/I analysis over the traditional confirmatory factor analysis approaches. ME/I analysis is the best approach for identifying where the measurement differences exist due to the multiple steps necessary to conduct the procedure (Bollen; 1989; Diefendorff, Silverman, & Greguras, 2005; Vandenberg & Lance, 2000). Also, when conducting a confirmatory factor analysis, there may be substantial measurement inequivalence across groups. We cannot compare mean group differences across measures that are not equivalent. The tests of ME/I can provide support for measurement equivalence and thus justify the comparison of mean group differences.

Because of the increasingly restrictive models used in ME/I analysis (described in detail below), if differences exist between the scale configurations, the source of those differences can be identified. Should we find evidence for measurement equivalence, then additional research into the viability of configurable testing would be warranted (e.g., comparisons of criterion-related validity; various other possible configurations). In contrast, if we fail to find evidence for measurement equivalence, then the results would suggest that configurable personality testing may not be a viable practice and that practitioners should continue to use more traditional fixed questionnaire approaches. As we expect that configurable personality testing may be associated with more positive

(applicant) reactions, we will also evaluate participants reactions to the different configurations of the test administered in this study.

Research Questions

This study will focus on exploring two primary research questions to help us determine the viability of configurable personality testing.

Research Question 1 – Will a configurable test show measurement equivalent/invariance with the full-length standard test?

Research Question 2 – Will peoples' reactions be different on a configurable test than on the full-length standard test?

METHOD

Participants

We collected data on 146 respondents (N = 50 in Version 1, N = 49 in Version 2, and N = 47 in Version 3). Participants were recruited from undergraduate psychology courses at a large southeastern university to participate in the two-part study. Participants returned for the second session three weeks after they completed the first session. Students who participated in the research study earned 2 research credits toward their course grade. Demographic information on the participants were as follows: 118 females and 28 males; average age of 23.5; 31 freshman, 26 sophomores, 44 juniors, and 44 seniors (1 participant did not respond to this item); 79 White, 34 African-American, 9 Hispanic, 10 Asian-Pacific Islander, and 13 Other (1 participant did not respond to this item). Participants were not screened on race, gender, age (except the minimum age of 18), or the number of jobs previously held.

Procedure and Experimental Conditions

A two-time period design was used in this study in order to make within-person comparisons (e.g., comparing Time 1 results with the same individuals' Time 2 results). Between-person comparisons will also be conducted using the versions *within* each time period.

Time 1. Upon arrival to the laboratory, participants were seated at individual computers and informed consent was obtained (groups consisted of up to 30 individuals). Participants were told that they would begin by completing a personality measure and that it was important for them to answer the items as honestly and as carefully as possible. They were also told that the session would last the entire hour scheduled and

that rushing would not get them out any earlier. What's more, they were told that if they found that they were getting bored they could take a brief break, but that they could not use the computer (i.e., the internet), their cell phones, or talk to others. All items were administered on the computer using a customizable, online, data collection system called Qualtrics.

Participants began by completing the full, 300-item, IPIP-NEO (The International Personality Item Pool Representation of the NEO-PI-R^{TM}) with items ordered using a repeating item/factor rotation approach, similar to that used by Schell and Oswald (2010). A researcher monitored the participants to ensure that they are responded carefully. Upon completing the measure, a brief demographic questionnaire was administered. Next, a reactions measure was administered. After that, participants completed a set of items unrelated to the study. These "filler items" were included to keep all participants busy for the fully allotted time period.

Time 2. Roughly three weeks after the completion of the first session, participants returned to the laboratory for the second session (again in groups of up to 30 individuals). Participants were instructed that it was once again important for them to answer the items as honestly and as carefully as possible. To begin this session, participants completed one of the three versions of the personality measure as follows:

- Version 1: Participants completed all scales from the IPIP-NEO, the exact same measure that they completed at Time 1. The participants who took this version served as a control group. This measure included 30 facet scales, 6 representing each of the Big Five Personality traits.

- Version 2: Participants completed only the items measuring the following six facet scales: Achievement-Striving (Conscientiousness), Self-Discipline (Conscientiousness), Anxiety (Neuroticism), Self-Consciousness scales (Neuroticism), Morality (Agreeableness) and Modesty (Agreeableness).
- Version 3: Participants completed only the items measuring the following six facets scales: Achievement-Striving (Conscientiousness), Self-Discipline (Conscientiousness), Friendliness (Extraversion), Assertiveness (Extraversion), Imagination (Openness to Experience) and Intellect (Openness to Experience).

Note that only the Achievement Striving and Self-Discipline scales (Conscientiousness scales) are common to Versions 2 and 3.[1]

Participants were randomly assigned to a version of the personality measure. Upon completion of the personality measure, participants completed the same reactions measure used in Time 1. After completing the reactions measure, participants completed a set of "filler items" to keep them busy during the duration of the study. At the conclusion of the one-hour session, participants were debriefed and thanked for their participation.

Measures

IPIP-NEO (Version 1). The IPIP-NEO (The International Personality Item Pool Representation of the NEO-PI-R™) (Goldberg, 1999; Goldberg, Johnson, Eber, Hogan,

[1] The facet scales used in Versions 2 and 3 were chosen based on their relevance to the testing situation and to the college sample. The facet scales were selected using a panel of graduate student subject-matter experts (SMEs) who were asked to go through all items for each of the facets on the full IPIP-NEO and determine the facets that they believed were most relevant for a school context (university setting) for college students. The items were given in the same order as the full IPIP-NEO. For Version 2, the panel of SMEs determined that one item from the Morality facet of the Agreeableness trait was irrelevant to the college-aged sample. That one item ("At school I would never cheat on my taxes") was omitted from the study. For Version 3, the panel of SMEs determined that all items in the facets for these traits were relevant to the college-aged sample; thus, no items were omitted from the facets given in Version 3.

Ashton, Cloninger, & Gough, 2006) has scales that are constructed to be analogs to the commercial NEO-PI-R (Srivastava, 2011; Costa & McCrae, 1992). The IPIP-NEO scales are in the public domain and do not require permission for use. Although two shorter versions of this inventory are available, this longer, 300-item version was selected because it was the only inventory of the three that includes six facets for each of the Big Five scales. These facets allowed us to make more distinct comparisons between the larger version of the personality inventory and the shorter versions created for this study. Evidence of convergent validity of the IPIP-NEO with other scales of the Big Five is provided in Goldberg et al. (2006) and in Goldberg (1999). The mean correlation between the 30 facet scales of the IPIP-NEO measure with the NEO-PI-R (Costa & McCrae, 1992) is .73 (.94 after correcting for attenuation due to unreliability) (Goldberg, et al., 2006). This measure served as a common standard for comparison purposes. The response scale used for this study was the standard 1-5 Likert scale used for the IPIP-NEO (Very Inaccurate = 1, Moderately Inaccurate = 2, Neither Accurate Nor Inaccurate = 3, Moderately Accurate = 4, Very Accurate = 5).

Reactions Measure. We selected and/or adapted items from other previously used measures to assess the reactions to the IPIP-NEO. These reactions included satisfaction, lack of concentration, job relatedness, and perceived fairness. Eleven items were adapted from Tonidandel, Quiñones, and Adams (2002) to measure Questionnaire Satisfaction (see Items 1 – 3 in Appendix A) and Perceived Fairness (see Items 12 – 19 in Appendix A). Four items were adapted from Bauer, Truxillo, Sanchez, Craig, Ferrara, and Campion's (2001) Selection Procedural Justice Scale (SPJS) to measure Job Relatedness (see Items 8 – 11 in Appendix A). Four items were adapted from Arvey, Strickland,

Drauden, and Martin's (1990) Test Attitude Survey to measure Lack of Concentration (see Items 4 – 7 in Appendix A). Responses to these 19 items were made on a 5-point Likert-type scale (*strongly agree* to *strongly disagree*). At Time 1, the Cronbach's alphas for the Questionnaire Satisfaction, Lack of Concentration, Job Relatedness, and Perceived Fairness scales were .91, .89, .76., and .86, respectively. At Time 2, the Cronbach's alphas were .90, .88, .87 and .84 for same scales, respectively.

Demographic Questionnaire. A brief demographic questionnaire was included to gather information on the participants' age, sex, race/ethnicity, and years of work experience.

RESULTS

Comparison of Scale Means

For the key analyses presented here, we only focused on those facet scales that were included in both Versions 2 and 3. Thus, we focus on 10 facets, two from each trait (Time 2, Version 2 included two facets for N (Neuroticism), A (Agreeableness) and C (Conscientiousness) and Version 3 included two facets for E (Extraversion), O (Openness to Experience) and C (Conscientiousness)). The specific facets examined are presented in Table 1. The table also presents the means, standard deviations and alpha internal consistency reliability coefficients for each of the facets on each of the two assessment times. These results are important to help us see if the means appear to be fairly stable across time. To more formally test the differences in the means, standardized mean differences effect sizes (Cohen's d) between Time 1 and Time 2 scores for each facet are also presented in the table. Cohen gives the standards for small, medium, and large effect sizes such that $d = .2$ is considered "small," $d = .5$ is considered "medium," and $d = .8$ is considered "large" (Cohen, 1988). The largest effect size measured was Friendliness (facet of Extraversion) in Version 3 with a Cohen's d value of -.25 (Table 1). This corresponds to a small size effect, suggesting little difference between the Time 1 and Time 2. Overall, we see from these results that giving a configurable test in Time 2 yielded little change in the means of the scales.

We also looked at the between-person effects at Time 2 for the two facets of Conscientiousness, Achievement-Striving and Self-Discipline, as these were the only facets included in all 3 Versions and across Time periods. Results are presented in Table 2. Five of the six between-person effect sizes at Time 2 are considered small by Cohen's

standards; the one exception was for the comparison of the Achievement-Striving facet between Versions 2 and 3, which was a medium effect. The generally small effect size differences observed between the Versions for Achievement-Striving and Self-Discipline indicate that scale means are quite similar across conditions, even between Versions that include different facets. Thus, the configuration of certain sets of items from different facets had little effect on mean scores on the scales.

To more formally test the differences in these means, repeated measures analyses of variance (ANOVA) were conducted for each of the facets of each trait (2 per trait) for all 3 Versions of the test. The two independent variables analyzed were Time (Time 1 vs. Time 2) and Version. As previously stated, Version 1 included facets for all 5 of the traits (N, E, O, A and C). Version 2 only included facets from N, O and C. Version 3 only included facets from E, A and C. As such, the Version variable had only *two* levels of analysis for the facets of N, E, O and A traits as each of these 4 traits were only given in Version 1 and in ONE of the other Versions (either Version 2 or Version 3) at Time 2. However, because the facets of Conscientiousness were given in all 3 Versions at Time 2, the Version variable in the ANOVA had *three* levels in the analyses for the Conscientiousness facets. The results for each facet are presented in Table 3. Main effects for Time were found for Anxiety (Neuroticism facet), Assertiveness (Extraversion facet), and Friendliness (Extraversion facet). The means for Anxiety in Version 1 and Version 2 were higher at Time 1 than they were at Time 2. The means for Assertiveness were lower at Time 1 than at Time 2 for both Version 1 and Version 3. No main effects for Version were found. There was a statistically significant Time by Version interaction for Friendliness (a graphic representation of this interaction is depicted in Figure 1). As

shown in Figure 1, the mean for Friendliness in Version 1 at Time 1 was higher than the mean for Friendliness in Version 3 at Time 1. The means, however, shifted at Time 2 such that Friendliness in Version 3 at Time 2 was *higher* than Friendliness in Version 1 at Time 2. Overall, the fact that only *one* interaction was found implies that regardless of how the test was configured, participants' mean responses on the facet scales were for the most part similar.

Comparison of Correlations

A comparison of test-retest correlations was conducted between facets to assess internal stability and patterns of change between Versions and across Time. Results are presented in Table 4. All participants were given the same assessment at Time 1, regardless of Version, and we see that all of the facets significantly correlated with each other across time. For example, Anxiety (N) in Version 2 Time 1 significantly correlated with Anxiety (N) in Version 2 Time 2 at $r = .89$. Conscientiousness was the only trait with facets recurring in all Versions at Time 2. The test-retest correlations were fairly similar across the Versions for both of the facets of Conscientiousness (Achievement-Striving and Self-Discipline). Achievement-Striving had significant test-retest correlations from Time 1 to Time 2 of $r = .87$ for Version 1, $r = .87$ for Version 2, and $r = .78$ for Version 3. The same is true for the Self-Discipline facet of Conscientiousness which had even more similar correlations across time with $r = .83$ for Version 1, $r = .87$ for Version 2, and $r = .84$ for Version 3. These findings indicate that overall, regardless of how the test is configured, the facets remained stable.

Measurement Equivalence/Invariance Analysis

Determining ME/I is accomplished by specifying increasingly restrictive confirmatory factor models. Vandenberg and Lance (2000) outline seven steps for conducting a thorough ME/I analysis. The first four steps are tests of measurement invariance examining the relationships between both measured variables and latent factors/variables. The last three steps are tests of structural invariance examining the latent variables themselves (Byrne, Shavelson, & Muthén, 1989; Vandenberg & Lance, 2000).

Initially, we wanted to examine our data within-person (Version 1 compared between Time 1 and Time 2, Version 2 compared between Time 1 and Time 2, etc.) and between-persons (Version 1 compared to Version 2 within Time 1, Version 2 compared to Version 3 within Time 1, etc.). We tried to compare Version 1 in Time 1 with Version 1 in Time 2 using all of the facets as indicators of the five traits. We were unable to generate model fit. We then compared Version 2 in Time 1 with Version 2 in Time 2 using the items from only those facets that were included at Time 2 as indicators. The model did not fit. We also looked at the comparison of Version 3 in Time 1 with Version 3 in Time 2 using the items from those facets that were included at Time 2 as indicators. Again, we were unable to generate model fit.

We then attempted to do some between-groups comparisons (e.g., Version 1 compared to Version 2 within Time 1). In this analysis, the facets were used as indicators for the five traits. Again, we were unable to generate model fit. The same was true when we looked at reduced models, comparing the Versions at Time 2 (where there was only on a small number of facets in common between the versions).

Given that we could not get a model to fit that would allow us to examine whether the changes that we made to the measures created measurement invariance, we decided to conduct an ME/I analysis for demonstration purposes only. For this demonstration we decided to focus on a model of Conscientiousness at Time 1, with the six facets as indicators of a Conscientiousness latent trait, and to compare Version 1 with Version 2, Version 1 with Version 3 and Version 2 with Version 3. Note that in this demonstration the results should indicate invariance as these are comparisons of groups that were created by random assignment taking the same test. The results are depicted in Table 5 for Version 1 compared to Version 2, Table 6 for Version 1 compared with Version 3, and Table 7 for Version 2 compared with Version 3.

Model Zero (Conscientiousness)

The 6 facets of Conscientiousness were used as indicators for Conscientiousness in the ME/I analysis (Self-Efficacy, Orderliness, Dutifulness, Achievement-Striving, and Cautiousness). The data used for the facets of Conscientiousness were from all participants who completed the test at Time 1 (all 3 versions of the test were the same at Time 1). The goodness of fit statistics for this model indicated a questionable level of model fit ($\chi^2 = 23.72$ ($p < .05$), RMSEA = 0.11, TLI =0.96, SRMR = 0.04, and CFI = 0.98); however, we continued through the steps of ME/I analysis for demonstration purposes.

Version 1 Time 1 with Version 2 Time 1 (Conscientiousness)

We compared Version 1 and Version 2 within Time 1 using the seven steps of ME/I analysis. The first step in the ME/I analysis, called configural invariance, involves specifying a model in each condition so that the sets of indicator variables define the

same number of latent factors. If one finds that the observed indicator variables represent different numbers of latent factors, then further tests of ME/I are unnecessary. Configural invariance provides a "baseline" evaluation that is needed to see if there are differences between the groups (Vandenberg & Lance, 2000, p. 18). If configural invariance is supported between the two groups such that the same number of latent factors is defined by the indicators, then one begins the second step of examining metric invariance.

For Step 1 of the ME/I analysis, the fit statistics comparing Version 1 to Version 2 within Time 1 clearly suggest a lack of model fit ($\chi^2 = 42.76$ ($p < .05$), RMSEA = 0.17, TLI = 0.91, SRMR = 0.06, and CFI = 0.94). Results of this analysis are presented in Table 5. The chi-square test statistic is significant ($\chi^2 = 42.76$, $p < .05$) indicating that there may be a difference between the two groups; however, and the Comparative Fit Index (CFI) equaled 0.94, which is above the recommended cutoff of .90 and considered acceptable. The Root Mean Square Error of Approximation (RMSEA) equaled 0.17 (above the recommended .10 threshold and thus outside of acceptable range), the Tucker-Lewis Index (TLI) equaled 0.91 (above the recommended .90 cutoff and considered acceptable), and the Standardized Root Mean Residual (SRMR) equaled 0.06 (below the recommended .08 threshold considered acceptable). At this point, interpretations are ambiguous: it could be that the six facets do not define the Conscientiousness latent variable very well or it could be that there are a different number of factors underlying the facets in Version 1 than there are in Version 2. With this lack of fit, no further tests of ME/I should be conducted. However, we will continue to Step 2 for demonstration purposes.

Step 2, metric invariance, examines whether or not the factor loadings of the items (the lambda weights) are equivalent across the two assessments. In this step, one must first specify the invariant factor pattern between the measures. Then, one must constrain for configural invariance while also constraining the loadings of like items within that invariant factor pattern to be equal. If one finds a difference in the degree of fit between Step 1 and Step 2 (i.e., the overall fit of the model gets notably worse), then the item indicators are loading differently on the factors for the different assessments. One wants to find no change in the fit of the indicators on the factors so that one can assert that the two assessments fit equally well (essentially predicting the null hypothesis). If the null hypothesis is rejected in this step, then the item indicators are loading differently on the factors for different assessments. If the factor loadings are equivalent, one continues on to Step 3 to the test for scalar invariance.

The fit statistics for Step 2, metric invariance, in the comparison of Version 1 to Version 2 are shown in Table 5. What we look for here is a change in fit from Step 1. If the fit is similar to Step 1, then we know that the added constraints did not hurt the model and we can conclude that there is metric invariance. If the fit of Step 2 is worse than at Step 1, then we know that constraining the factor loadings to be equivalent harmed the fit of the model. As such, we could conclude that there is not metric invariance – that the factor loadings differ across the two groups. Looking at the fit statistics presented in Table 5, we can see that the fit of Step 2 ($\chi^2 = 45.70$ ($p < .05$), RMSEA = 0.14, TLI = 0.94, SRMR = 0.08, and CFI = 0.95) is not substantially worse than that of Step 1. Beyond simply comparing the fit statistics, we also calculated the Chi-square difference test. This test, also shown in Table 5, was not statistically significant, indicating that

constraining the factor loadings to be equivalent across the two groups did not change model fit. The results of the analysis of Step 2 collectively would indicate that the model holds metric invariance – i.e., that the factor loadings are the same within the two groups. As such, we continued to Step 3.

Step 3, scalar invariance, examines whether the intercepts are equivalent between the two groups. By rejecting the null hypothesis in this step, one can conclude that the respondents are not responding the same way to the items. For example, it could be that the respondents in one group are using the response scale in a way that is different from respondents in the other group. The assessments may show the same latent structure and indicator variables; however, respondents may be rating themselves higher (leniency bias) or lower (severity bias) on some of the items.

We conducted Step 3, scalar invariance, analysis comparing Version 1 and Version 2. The fit statistics for this comparison are shown in Table 5. What we look for here is a change in fit from Step 2 to Step 3. If the fit is similar to Step 2, then we know that the added constraints did not hurt the model and we can conclude that there is scalar invariance. If the fit of Step 3 is worse than at Step 2, then we know that constraining the intercepts to be equivalent harmed the fit of the model. As such, we could conclude that there is not scalar invariance – that the intercepts differ across the two groups. Looking at the fit statistics presented in Table 5, we can see that the fit of Step 3 ($\chi^2 = 154.81$ ($p < .05$), RMSEA = 0.27, TLI = 0.76, SRMR = 0.34, and CFI = 0.73) is substantially worse than that of Step 2. For example, the TLI in Model 2 was 0.94 but it was 0.76 in Model 3, indicating much less adequate fit. We also calculated the Chi-square difference test. This test, also shown in Table 5, was significant, indicating that constraining the intercepts to

be equivalent across the two groups significantly changed the model fit. Overall, we can conclude that there is not scalar invariance and the respondents are not rating themselves equally on the items. As ME/I analysis uses progressively restrictive models, and the Versions were found to be non-equivalent within Step 3, we did not continue to the Step 4.

Version 1 Time 1 with Version 3 Time 1 (Conscientiousness)

We compared Version 1 and Version 3 within Time 1 using the seven steps of ME/I analysis, just as we did comparing Version 1 with Version 2 above. The results are shown in Table 6. As expected, the configural model (Step 1) did not fit (χ^2 = 30.92 (p < .05), RMSEA = 0.12, TLI = 0.94, SRMR = 0.07, and CFI = 0.97). The added constraints of metric invariance (Step 2) did not hurt the fit (χ^2 = 35.18 (p > .05), RMSEA = 0.10, TLI = 0.96, SRMR = 0.10, and CFI = 0.97). However, there was not scalar invariance (χ^2 = 128.95 (p < .05), RMSEA = 0.24, TLI = 0.78, SRMR = 0.33, and CFI = 0.75). The model fit was particularly poor in the analysis of scalar invariance (Step 3), suggesting that the respondents are not rating themselves equally on all the items. Because the versions were found to be in-equivalent in Step 3, and because of the progressively restrictive models used in ME/I analysis, we did not continue to Step 4, invariant uniqueness.

Version 2 Time 1 with Version 3 Time 1 (Conscientiousness)

Finally, we compared Version 2 and Version 3 within Time 1 using the seven steps of ME/I analysis, just as we did with the comparisons of the other Versions above. Results from the analysis are presented in Table 7. Once again, the configural model (Step 1) did not fit (χ^2 = 41.38 (p < .00), RMSEA = 0.16, TLI = 0.90, SRMR = 0.07, and

CFI = 0.94). The added constraints of metric invariance (setting the factor loadings to be equivalent) did not hurt the fit (χ^2 = 47.61 (p < .00), RMSEA = 0.14, TLI = 0.92, SRMR = 0.10, and CFI = 0.94), indicating that the model holds metric invariance. Once again, the results of the test of scalar invariance (Step 3) indicated poorer fit (χ^2 = 150.97 (p < .05), RMSEA = 0.27, TLI = 0.74, SRMR = 0.34, and CFI = 0.70). This again suggests that the respondents are not rating themselves equally on all the items. Because the versions were found to not be equivalent in Step 3, and because of the progressively restrictive models used in ME/I analysis, we did not continue to Step 4, invariant uniqueness.

Participant Reactions

Means and standard deviations of the four scales in the Reactions Measure (Questionnaire Satisfaction, Lack of Concentration, Job Relatedness, and Perceived Fairness) are presented in Table 8 for both the Time 1 and Time 2 administrations. All means for all scales of the Reactions Measure were lower at Time 2 than at Time 1. To more formally test the differences in these means from Time 1 to Time 2, multivariate analyses of variance (MANOVA) was conducted on the scales of the Reactions Measure across administrations. No main or interaction effects were found; however, we conducted ANOVAs on the 4 scales of the Reactions Measure to look more closely at the results. The results of these analyses are presented in Table 9. The two independent variables analyzed were Time (Time 1 vs. Time 2) and Version (Versions 1, 2, and 3 of the assessment). Each scale of the Reactions Measure was given in all three versions of the test. The results revealed main effects for Time, such that the means were higher at Time 1 than at Time 2 for each of the four scales of the Reactions Measure. No main

effects for Version were found showing that the means of the Reactions Measure in each version were not statistically significantly different from one another at each time period. No statistically significant Time by Version interactions were found.

DISCUSSION

Historically, organizations choosing to assess applicants on their personality characteristics have used measures that are not customized to specific jobs or to the specific organization. The measures used are typically rather long and tend to include both items for personality traits that are related to the job in question and items for traits that are not directly related to the job. The widespread use of computerized/online personality testing has opened the door for organizations to use more flexible personality measures. Organizations can now configure their personality tests, such that only those items for trait scales that are directly related to the job are administered. Although configurable personality testing may seem to be a better approach at first glance, research is needed to investigate the psychometric nature of the these tests to determine if restructuring the items yields test scores that are equivalent to those that would be obtained from the full-length assessment. The current study sought to explore the measurement equivalence of personality scales across different configurations of tests using measurement equivalence/invariance (ME/I) analyses in order to provide some evidence into the viability of configurable personality testing. We were concerned with *scale* level analyses rather than *item* level analyses in this study as interpretations of personality are not made at the item-level and companies won't make decisions about applicants on the basis of personality *items*.

Analysis of Means, SDs, Effect Sizes, and Correlations

Our results indicated that the means and standard deviations of the facet scales remained fairly consistent across versions of the test administered in this study and across time (see Table 1). The majority of the effect size differences across versions and over

time were considered to be small by Cohen's standards, i.e., less than .20 (Cohen, 1988). The largest effect size differences were for the Friendliness and Assertiveness facets of Extraversion in Version 3 between Time 1 and Time 2. This could indicate that the participants responded to the items within these facets differently in the shorter version of the test (Version 3 at Time 2). A similar effect size difference was also observed for Anxiety (facet of Neuroticism) in Version 1 between Time 1 and Time 2 (Cohen's d = .23). As the test was exactly the same at both administration times for participants in the Version 1 condition, the results for this Anxiety facet suggest that normal fluctuations in responses can lead to effect sizes of this magnitude. As such, we can generally conclude from these effect sizes that the configurable personality measure did not yield results that varied significantly from the original, full-length, version of the measure.

The results of our ANOVA showed that the only statistically significant Time by Version interaction occurred at the facet of Friendliness (a facet of Extraversion; see Table 3 and Figure 1). The mean for Friendliness in Version 1 at Time 1 was higher than the mean for Friendliness in Version 3 at Time 1. However, at Time 2 the opposite pattern was found: the mean for Friendliness was higher in Version 3 than in Version 1. If configurable personality testing *does* influence test scores, then we would have expected to see more Time by Version interactions across the various facets. The fact that only *one* interaction was found indicates that in most cases, the score changes from Time 1 to Time 2 were the same, regardless of Version.

The results of the test-retest correlations from Time 1 to Time 2 indicate stability in scores over time (Table 4). Once again, this is good news for configurable personality testing as these results suggest that the scores on facet of the trait in the original

personality test were significantly correlated with scores on the same facet in the configurable versions of the test. Importantly, the size of the correlations of the facets over time in Version 1 (where participants took the same test on both occasions) were very similar to those in Versions 2 and 3 (which were the configurable versions of the test).

Measurement Equivalence/Invariance Analyses

The results of the analysis of the means, standard deviations, effect sizes, and correlations collectively suggest that different configurations of the items do not negatively impact the viability of the facet scales. However, in order to truly see where differences exist between the original full-length personality inventory and the shorter, configurable versions, ME/I analyses are needed. We ran into some difficulties with the ME/I analysis (unable to generate a Model Zero), most likely due to the small size of our sample. As such, we decided to conduct the ME/I analyses on Conscientiousness for demonstration purposes only.

As a reminder, the same sets of items were given in Versions 1, 2, and 3 at Time 1. As such, we expected to see that the measurement of Conscientiousness would be *equivalent* between the Version conditions at Time 1. The results, however, were not consistent with this expectation. Using the data from all three versions administered at Time 1, we were unable to generate model fit for the Model Zero to begin our ME/I analyses. Although we tried various modifications to our model to generate a Model Zero, we were unable to do so. Our inability to find an acceptably fitting model is most likely due to the small sample size; we had a total sample size of N = 146 (N = 50 for Version 1, N = 49 for Version 2, and N = 47 for Version 3). Alternatively, the results

could suggest deficiencies in the measurement properties of the IPIP-NEO – that the facets do not define the factors. A study examining the construct validity of the shorter (50-item) IPIP-NEO determined that even though a five-factor model could be generated by the facets, the results did not produce very good fit when item-level data were analyzed (Lim & Ployhart, 2006). Although the overall results of validation studies for the IPIP-NEO have been favorable (Goldberg, 1999; Goldberg et al., 2006; Lim & Ployhart, 2006), more research should be conducted to further validate that the facets as well as the *items* of the longer (300-item) IPIP-NEO.

Despite the lack of fit of a general model, we proceeded with the ME/I analyses for demonstration purposes. Although there was evidence of configural and metric invariance in all between-Version comparisons at Time 1 (Version 1 with Version 2, Version 1 with Version 3, and Version 2 with Version 3), the data indicated that there was not scalar invariance. This finding indicates that the intercepts for the Conscientiousness items were not equivalent across the groups, suggesting that the respondents were not rating themselves equally on some of the items depending on the Version of the test they received. The respondents appear to be either rating themselves higher or lower on some of the items. These results are unexpected as the test delivered to all of the participants was the same in each version and participants were randomly assigned to each version. Given the overall lack of fit for these models it is hard to make any strong conclusions, but future research on configural testing should investigate the ME/I of different combinations of items and, perhaps, for a different measure.

Reactions Measure

Applicant perceptions should be important to organizations that are requiring their applicants to complete selection tools (Macon, Avedon, Paese, & Smith, 1994; Murphy, 1986; Smither, Reilly, Millsap, Pearlman, & Stoffey, 1993). As stated earlier, we assert that there are two reasons why configurable personality tests should result in more positive applicant reactions than traditional personality tests. The first reason is that tests that are configured to a specific job will be shorter and thus require less time for the applicant to complete. The second reason is that configurable personality tests should appear to be more job-related with higher face validity. It was also important to this study to understand the reactions of individuals to the different configurations of the test we administered. As organizations may be considering giving different configurations of personality inventories to applicants, positive results regarding satisfaction, concentration, feelings of job-relatedness, and feelings of perceived fairness are important for positive reactions toward the organization. Although we thought that the shorter test that was configured to be more "job-relevant" to the role of "student" would lead to more positive reactions, our results did not support that expectation. Interestingly, the means for all four scales of the Reactions Measure we administered were *lower* at Time 2 than at Time 1 (Table 8). This may have been due to asking the students to take the same (or very similar) measure for a second time, even though there was a 3-week interval between the Time 1 and Time 2 administrations. Perhaps the students experienced testing fatigue from having to come to the lab setting for a second time to take yet *another* test. However, the results of the MANOVAs (Table 9) showed that while the means of the reactions measures were higher at Time 1, the differences between

Time 1 and Time 2 were not statistically significant. Although we did not find the reactions to be higher for the configurable versions, the results do suggest that one could give a configurable version of the personality test and expect similar reactions to that of the original (full) test.

Limitations and Future Research

Our recommendation is that further tests using ME/I analysis be conducted with larger sample sizes, a different demographic of respondents, and potentially a different personality measure, to fully understand if configurable personality tests are statistically equivalent to their full-length personality assessment counterparts. The smaller sample size in this study was likely the reason for our inability to generate a Model Zero for our ME/I analyses. A larger sample size could allow future researchers to generate a Model Zero and investigate whether configurable personality tests are in fact equivalent to their full-length personality assessment counterparts, or perhaps more accurately determine *where* they differ.

A different demographic of respondents may also yield different results. Our respondents consisted of a college student sample. Although we selected the items to ensure they applied to the job of "student," and we added the phrasing "At school" before each item, an older sample of respondents who have potentially had more experience applying for jobs and/or taking personality inventories for jobs might yield different results. Future researchers might also consider giving administering this study to an actual sample of job *applicants*, were some applicants take the full length test and others that a configurable test that only includes the scales that are job-relevant.

We chose to use the IPIP-NEO as it has scales that are constructed to be analogs to the commercial NEO-PI-R (Srivastava, 2011; Costa & McCrae, 1992). The IPIP-NEO scales are free to use and do not require permission for use. Although evidence of the validity of the IPIP-NEO with other scales of the Big Five is available (Goldberg et al., 2006; Goldberg, 1999), it is not considered to be the "benchmark" assessment for measuring the Big Five personality traits. We would recommend further studies be conducted perhaps using the commercial NEO-PI-R or other highly respected measures of the Big Five personality traits to determine if the results would yield a Model Zero for ME/I analyses.

Conclusion

Our results provide initial evidence in support of configurable personality testing. Different configurations did not change the scale means and the correlations of the facets over time were strong despite changes in the configuration. However, additional research is needed, especially to evaluate the equivalence of scores from full length and configured versions of the test. We believe that the configurable approach to personality testing might provide an advantage to the more traditional fixed questionnaire approach with respect to the amount of time required to complete the assessment, the potential for more positive reactions from respondents, and the potential for an improved quality of decision making by hiring managers. Practitioners and researchers need to be assured that when they use configurable personality tests, they can trust that the results are equivalent to their full-length counterparts.

REFERENCES

Ackerman, P.L., & Kanfer, R. (2009). Test length and cognitive fatigue: an empirical examination of effects on performance and test-take reactions. *Journal of Experimental Psychology,, 15(3),* 163-181.

Arvey, R.D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43,* 695-716.

Bauer, T.N., Truxillo, D.M., Sanchez, R., Craig, J., Ferrara, P., & Campion, M.A. (2001). Development of the Selection Procedural Justice Scale (SPJS). Personnel Psychology, 54, 387-419.

Beehr, T.A., & Gilmore, D.C. (1982). Applicant Attractiveness as a Perceived Job-Relevant Variable in Selection of Management Trainees. *Academy of Management Journal, 25*(3), 607-617.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Bolster, B.I., & Springbett, B.M., (1961). The reaction of interviewers to favorable and unfavorable information. *Journal of Applied Psychology, 45*(2), 97-103.

Carlson, R.E., & Mayfield, E.C. (1967). Evaluating the interview and employment application data. *Personnel Psychology, 20*(4), 441-460.

Cohen, J. (1988) Statistical Power Analysis for the Behavioral Sciences (second ed.). Lawrence Erlbaum Associates.

Diefendorff, J.M., Silverman, S.B., Gregarus, G.J. (2005). Measurement equivalence and multisource ratings for non-managerial positions: Recommendations for research and practice. *Journal of Business and Psychology, 19*(3), 399-425.

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.

Goldberg, L.R., Johnson, J.A., Eber, H.W., Hogan, R., Ashton, M.C., Cloninger, C.R., & Gough, H.C. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84-96.

Hausknecht, J.P., Day, D.V., & Thomas, S.C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57,* 639-683.

Knowles, E.S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology, 55*(2). 312-320.

Kulik, C.T., & Clark, S.C. (1994). Category-based and feature-based cognitive processes: The role of unfavorable information. *Journal of Applied Social Psychology, 24*(21), 1891-1918.

Leary, L.F., & Dorans, N.J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research, 55,* 387-413.

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

Macon, T.H., Avedon, M.J., Paese, M., & Smith, D.E. (1994). The effects of applicants' reactions to cognitive ability tests and an assessment center. *Personnel Psychology, 47,* 715-738.

Murphy, K.R. (1986). When your top choice turns you down: Effect of rejected job offers on the utility of selection tests. *Psychological Bulletin, 99,* 133-138.

Ployhart, R.E., & Ryan, A.M. (1998). Applicants' reactions to the fairness of selection procedures: The effects of positive rule violations and time of measurement. *Journal of Applied Psychology, 83,* 3-16.

Rynes, S.L., & Connerly, M.L. (1993). Applicant reactions to alternative selection procedures. *Journal of Business and Psychology, 7,* 261-277.

Schell, K.L., & Oswald, F.L. (2010). Item grouping and item randomization effects in personality measurement. *Manuscript in preparation.*

Smither, J.W., Reilly, R.R., Millsap, R.E., Pearlman, K., & Stoffey, R.W. (1993). Applicant reactions to selection procedures. *Personnel Psychology, 46*, 49-76.

Springbett, B.M. (1958). Factors affecting the final decision in the employment interview. *Canadian Journal of Psychology, 12*(1), 13-22.

Srivastava, S. (2011). *Measuring the big five personality factors.* Retrieved [*May 5, 2011*] from http://www.uoregon.edu/~sanjay/bigfive.html.

Steiner, D.D., & Gilliland, S.W. (1996). Fairness reactions to personnel selection techniques in France and the United States. *Journal of Applied Psychology, 84,* 134-141.

Tonidandel, S., Quiñones, M.A., & Adams, A.A. (2002). Computer-adaptive testing: the impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology, 87*(2), 320-332.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-70.

Table 1: Descriptive statistics

		Time 1				Time 2				
Version 1	**Facet**	**N**	**M**	**SD**	**alpha**	**N**	**M**	**SD**	**alpha**	**Cohen's d-value**
	Anxiety (N)	50	2.92	.66	.76	50	2.76	.74	.83	.23
	Self-Consciousness (N)	50	2.73	.71	.80	50	2.66	.75	.85	.10
	Friendliness (E)	50	3.71	.68	.84	50	3.67	.63	.82	.06
	Assertiveness (E)	50	3.17	.76	.86	50	3.20	.71	.84	-.04
	Imagination (O)	50	3.40	.64	.77	50	3.35	.55	.69	.08
	Intellect (O)	50	3.31	.73	.82	50	3.40	.69	.82	-.13
	Morality (A)	50	4.24	.51	.77	50	4.19	.44	.69	.10
	Modesty (A)	50	3.34	.57	.70	50	3.33	.55	.70	.02
	Achievement-Striving (C)	50	3.87	.63	.81	50	3.79	.66	.86	.12
	Self-Discipline (C)	50	3.39	.83	.88	50	3.44	.73	.85	-.06
Version 2										
	Anxiety (N)	49	2.81	.86	.87	49	2.70	.85	.89	.13
	Self-Consciousness (N)	49	2.58	.79	.86	49	2.59	.76	.86	-.02
	Friendliness (E)	49	3.75	.67	.85					
	Assertiveness (E)	49	3.34	.67	.82					
	Imagination (O)	49	3.29	.71	.82					
	Intellect (O)	49	3.63	.70	.85					
	Morality (A)	49	4.23	.51	.75	49	4.23	.55	.82	.00
	Modesty (A)	49	3.15	.54	.67	49	3.18	.57	.74	-.05
	Achievement-Striving (C)	49	3.94	.69	.82	49	3.96	.65	.84	-.03
	Self-Discipline (C)	49	3.61	.81	.90	49	3.52	.86	.91	.11
Version 3										
	Anxiety (N)	47	3.04	.67	.80	47				
	Self-Consciousness (N)	47	2.95	.70	.81	47				
	Friendliness (E)	47	3.60	.81	.90	47	3.80	.77	.92	-.25
	Assertiveness (E)	47	3.06	.71	.86	47	3.22	.70	.86	-.23
	Imagination (O)	47	3.32	.74	.85	47	3.45	.79	.90	-.17
	Intellect (O)	47	3.50	.68	.82	47	3.56	.63	.82	-.09
	Morality (A)	47	4.22	.52	.73					
	Modesty (A)	47	3.39	.62	.76					
	Achievement-Striving (C)	47	3.91	.59	.80	47	3.91	.67	.89	.00
	Self-Discipline (C)	47	3.36	.80	.87	47	3.51	.90	.93	-.18

Note: Means, Standard Deviations, and Reliabilities are for each Facet

Table 2: Standardized mean difference effect sizes (Cohen's d) between-person (time 2)

	Version 1 with Version 2	Version 1 with Version 3	Version 2 with Version 3
Achievement-Striving (C)	-.26	-.18	-.50
Self-Discipline (C)	-.10	-.09	.01

Table 3: Analysis of variance at the facet level

		Time		Condition (Version)		Time x Condition (Version)	
Scale	Facets	*F*	η^2	*F*	η^2	*F*	η^2
Neuroticism							
	Anxiety	9.58**	.09	.29	.00	.34	.00
	Self-Consciousness	.42	.00	.62	.01	.66	.01
Extraversion							
	Friendliness	3.94*	.04	.01	.00	10.86**	.10
	Assertiveness	5.49*	.06	.09	.00	2.49	.03
Openness to Experience							
	Imagination	.99	0.01	.01	.00	3.72	.04
	Intellect	2.86	.03	1.61	.12	.13	.00
Agreeableness							
	Morality	.48	.01	.02	.00	.48	.01
	Modesty	.09	.00	2.73	.03	.31	.00
Conscientiousness							
	Achievement-Striving	.43	.00	.45	.01	.97	.01
	Self-Discipline	.96	.01	.54	.01	3.04	.04

Note: *p < .05; **p < .01; The Condition (Version) variable in these analyses had two levels in the analyses of the facets for Neuroticism, Extraversion, Openness and Agreeableness, as each of these facets was given in Version 1 and one of the other versions. This variable had three levels in the analyses of the Conscientiousness facets, as these facets were included in all three versions of the test.

Table 4: Correlations for each facet within version

	Version 1 Time 1 with Version 1 Time 2	**Version 2 Time 1 with Version 2 Time 2**	**Version 3 Time 1 with Version 3 Time 2**
	Pearson *r* Correlation	Pearson *r* Correlation	Pearson *r* Correlation
Anxiety (N)	.80**	.89**	
Self-Consciousness (N)	.76**	.84**	
Friendliness (E)	.86**		.90**
Assertiveness (E)	.85**		.82**
Imagination (O)	.75**		.82**
Intellect (O)	.80**		.79**
Morality (A)	.78**	.70**	
Modesty (A)	.74**	.80**	
Achievement-Striving (C)	.87**	.87**	.78**
Self-Discipline (C)	.83**	.87**	.84**

*Sig. (2-tailed) = < .05

**Sig. (2-tailed) = < .01

Table 5: Results of measurement invariance tests of conscientiousness for version 1 compared with version 2 at time 1

Model	*df*	χ^2	RMSEA	TLI	SRMR	CFI	Δdf	$\Delta\chi^2$
1. Configural Invariance	18	42.76*	.17	.91	.06	.94		
2. Metric Invariance	24	45.70*	.14	.94	.08	.95		
1 vs. 2							6	2.94
3. Scalar Invariance	34	154.81*	.27	.76	.34	.73		
2 vs. 3							10	109.11**

Note: * $p < .05$

Table 6: Results of measurement invariance tests of conscientiousness for version 1 compared with version 3 at time 1

Model	*df*	χ^2	RMSEA	TLI	SRMR	CFI	Δdf	$\Delta\chi^2$
1. Configural Invariance	18	30.92*	.12	.94	.07	.97		
2. Metric Invariance	24	35.18	.10	.96	.10	.97		
1 vs. 2							6	4.26
3. Scalar Invariance	34	128.95*	.24	.78	.33	.75		
2 vs. 3							10	93.77**

Note: * $p < .05$

Table 7: Results of measurement invariance tests of conscientiousness for version 2 compared with version 3 at time 1

Model	*df*	χ^2	RMSEA	TLI	SRMR	CFI	Δdf	$\Delta\chi^2$
1. Configural Invariance	18	41.38**	.16	.90	.07	.94		
2. Metric Invariance	24	47.61**	.14	.92	.10	.94		
1 vs. 2							6	6.23
3. Scalar Invariance	34	150.97*	.27	.74	.34	.70		
2 vs. 3							10	103.36**

Note: * $p < .05$; ** $p < .00$

Table 8: Reactions measure means and standard deviations

		Time 1		Time 2	
		Mean	SD	Mean	SD
Questionnaire Satisfaction					
	Version 1	3.49	.72	3.05	.85
	Version 2	3.31	1.09	3.13	1.05
	Version 3	3.54	.85	3.26	.98
Lack of Concentration					
	Version 1	3.33	1.07	2.87	.98
	Version 2	3.31	1.08	3.13	1.06
	Version 3	3.45	.91	3.45	.97
Job Relatedness					
	Version 1	2.49	.70	2.43	.92
	Version 2	2.77	.70	2.65	.77
	Version 3	2.80	.70	2.41	.80
Perceived Fairness					
	Version 1	3.27	.67	3.00	.72
	Version 2	3.26	.69	3.09	.67
	Version 3	3.27	.64	3.10	.66

Table 9: Analysis of variance of the reactions measures

	Time		Condition (Version)		Time x Condition (Version)	
Scale	*F*	η^2	*F*	η^2	*F*	η^2
Questionnaire Satisfaction	25.48**	.15	.55	.01	1.62	.02
Lack of Concentration	7.25**	.05	1.94	.03	3.04	.04
Job Relatedness	9.08**	.06	1.74	.02	2.44	.03
Perceived Fairness	24.89**	.15	.10	.00	.73	.01

Note: *p < .05; **p < .01; Each scale of the Reactions Measure was given in each of the 3 Versions of the test at Time 1 and Time 2.

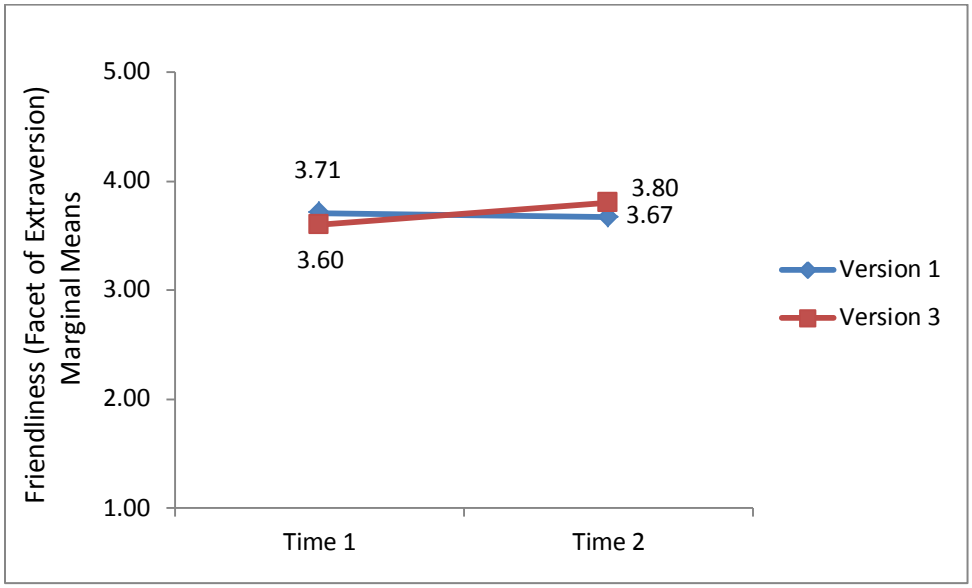

Figure 1. Graphic representation of the significant time x condition (version) interaction for the friendliness (facet of extraversion) personality scores.

APPENDIX A: USER REACTIONS

	Reverse Scored	Construct
1. I liked taking this questionnaire.		Questionnaire Satisfaction
2. This questionnaire was appealing to me.		
3. I did not like this questionnaire at all.	x	
4. It was hard to keep my mind on this questionnaire.		Lack of Concentration
5. I found myself losing interest and not paying attention to the questionnaire.		
6. When responding to the questionnaire, I was bored.		
7. I get distracted when responding to questionnaires of this type.		
8. Doing well on this questionnaire means a person can do the job well.		Job Relatedness
9. A person who scored well on this questionnaire will be a good performer on the job.		
10. It would be clear to anyone that this questionnaire is related to the job.		
11. The content of this questionnaire was clearly related to the job.		
12. The questionnaire was not a good indicator of my personality.	x	Perceived Fairness
13. The questionnaire was an unfair measure of a person's true personality.	x	
14. The questionnaire obtains accurate information about each person's personality.		
15. I have strong doubts that the questionnaire really measures a person's personality.	x	
16. This questionnaire should not be used to assess people's personality for jobs.	x	
17. I feel another procedure should have been used to assess my personality.	x	
18. My performance on this questionnaire was influenced by things that should not have been considered.		
19. Under the circumstances, the questionnaire was fair.		