Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Research in Personality 40 (2006) 359–376
JOURNAL OF RESEARCH IN PERSONALITY
www.elsevier.com/locate/jrp

Personality assessment: Does the medium matter? No ☆

Siang Chee Chuah *, Fritz Drasgow, Brent W. Roberts

Department of Psychology, University of Illinois in Urbana-Champaign, 603 East Daniel Street, Champaign, IL 61820, USA

Available online 3 March 2006

Abstract

The equivalence of an Internet administration of personality tests with two other administration formats was assessed using Item Response Theory (IRT) and various other statistical methods. The analyses were conducted on measures of Neuroticism, Extroversion, Agreeableness, and Conscientiousness. A total of 728 participants took part in the study. Participants were randomly assigned to one of three administrative conditions: paper-and-pencil, proctored computer lab, and unproctored Internet. Analyses with IRT, factor analysis, criterion-related validity, and mean differences supported the equivalence of Internet and traditional paper-and-pencil administrations of personality tests.

Keywords: Measurement equivalence; Personality assessment; Internet testing; Computer-based testing; Proctored testing; Big 5; IRT; DIF

1. Introduction

The proliferation of the Internet presents an opportunity to deliver questionnaires though a medium that offers tremendous benefits. It offers potential access to huge populations from diverse backgrounds and geo-political locations (Schmidt, 1997; Smith & Leigh, 1997). Tests can be administered at virtually any location where there is a computer and

☆ This research was partly supported by a grant from the National Institute of Aging R01 AG21178.
* Corresponding author.
E-mail address: chuah@uiuc.edu (S.C. Chuah).

0092-6566/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jrp.2005.01.006

Internet access. In that sense, the comfort of one's home may become a test center. Data are transmitted, literally at the speed of light, from the respondent to the test administrator through a network of computers connected by fiber-optic cables. Because the data are in a digital format, they may be quickly converted into a format suitable for data analysis (Davis, 1999). In addition, the resources needed to create a website to administer questionnaires can be relatively inexpensive (Schmidt, 1997).

However, with any new tool there lies the potential for abuse. As Butcher (1987) points out, people tend to be less critical when evaluating the legitimacy of computerized results. Computers are often seen as objective or impartial evaluators of test results. He argues that there is no assurance as to the quality of the computerized system. Without the appropriate input by trained professionals to develop or evaluate the adequacy of the system, the legitimacy of a computerized test is suspect.

A key concern for test developers and administrators who want to deploy their tests over the Internet is the equivalence of scores between paper-and-pencil administrations and Internet administrations. Before tests scores can be directly compared, equivalence between test formats needs to be established (Standards for Educational & Psychological Testing, 1999). Equivalence is important for several reasons. For example, the validity and reliability of the scale was most likely established using paper-and-pencil administration samples; however, we cannot simple assume that the Internet form has similar reliability and validity. Lacking empirical verification of cross-medium equivalence, the instrument needs to be reevaluated to assess its validity and reliability. Additionally, archival data is often used to establish population norms. These data are commonly collected using the paper-and-pencil version of the instrument. Norms are important because they allow one to interpret one's results in comparison with other studies across a broad range of test situations. If scores from an Internet are to be interpreted relative to norms obtained from a paper-and-pencil administration, equivalence is required.

This paper tests the equivalence of paper-and-pencil and Internet administrations by looking at the psychometric qualities and criterion validity of a personality measure across administration media. Item response theory (IRT) was used to assess the psychometric qualities of three different administrative methods of a personality questionnaire. Specifically, we examined the measurement equivalence of a set of adjectives from the Big 5 domain (Goldberg, 1992). We also assessed equivalence by mean of multi-group factor analysis: A single factor loading matrix was simultaneously fit to data from all three administrations. We assessed criterion-related validity by comparing correlations between the Big 5 and the Health Behavior Checklist (HBC; Vickers, Conway, & Hervig, 1990) across administrative media. Previous research has already established the relationship between personality and health behaviors (Bogg & Roberts, 2004; Booth-Kewley & Vickers, 1994). The results and implications of the findings are discussed.

2. Previous research

A number of factors can influence responses to questionnaires administered in diverse test modalities. Previous studies comparing the equivalence of computer administered tests to paper-and-pencil tests have suggested that specific features of computerized tests may affect the outcome. A meta-analysis by Richman, Weisband, Kiesler, and Drasgow (1999) found that non-cognitive tests suffered from social desirability distortions under certain conditions. Social desirability distortion refers to the tendency of respondents "to answer questions in a

more socially desirable direction than they would under other conditions or modes of administration" (Richman et al., 1999, p. 755). Richman et al., found that the inability of respondents to reconsider previous responses increased social desirability distortion. They also found that earlier studies reported larger effect sizes (for social desirability distortions) for computer administered tests as compared to more recent studies. Based on the assumption of poorer display capabilities of older computer systems, they argue that the difference in effect sizes could be partially explained by differences in the presentation format of the tests. This lends support to the idea that the closer in similarity a computer administered test is to its paper-and-pencil counterpart, the more similar its measurement properties should be.

Taken in the context of previous research on computerized tests, there are several reasons why Internet assessments may be different from their paper-and-pencil counterparts. Internet administration can allow research participants to answer the test or survey at a time and place of their choice. But with such unproctored administration, it is impossible to identify, with certainty, the identity of the test taker. Various tracking tools such as computer cookies, IP addresses and passwords can be circumvented by an informed and persistent test taker. This affords the test taker a heightened level of anonymity. In a study comparing paper-and-pencil and computerized administration, Richman et al. (1999) found that anonymity and taking the test alone (as opposed to taking it in a group) decreased social desirability distortion. Joinson (1999) found lower scores for social anxiety and social desirability items among Internet respondents. This suggests that respondents may be more forthcoming on an Internet questionnaire. If this premise holds true, it provides an opportunity for computer based assessments to provide measurement that is superior to their traditional paper-and-pencil alternatives for non-cognitive tests. Therefore, it is important that researchers look at how anonymity and proctoring may affect respondents.

Individual differences may also affect responses to computerized administration. The argument can be made that computer familiarity might impact examinee responses. A lack of familiarity with the computer equipment might impair the respondent's ability to respond accurately. However, empirical research on the impact of computer familiarity and anxiety has not been conclusive. Studies by Powers and O'Neill (1992) and Kirsch et al. (1998) found no relationship between computer familiarity and performance on a computerized test. They suggest that the initial tutorial to familiarize respondents with the workings of the program was sufficient to alleviate any effect a lack of familiarity with computers might have. This might be interpreted as an indication as to how intuitive or simple computer use has become. Consequently, computer familiarity might not be an issue for the current generation of computer users.

Conversely, other individual difference variables such as self-monitoring appear to have an impact on respondents. Rosenfeld et al. (1991) found that respondents high on self-monitoring reported lower job satisfaction scores in a computerized administration of the questionnaire in comparison to a paper-and-pencil administration. Conversely, they found that respondents low on self-monitoring reported higher job-satisfaction scores in the computerized administration. Rosenfeld et al. (1991) suggested that the respondents to the computer administration were more candid with their responses because the perceived increase in anonymity offered by the computer decreased the need for faking good. This suggests that the greater anonymity offered by an Internet administration may increase the validity of tests with items that are high in social desirability.

There have been a small number of studies comparing paper-and-pencil and Internet administrations of questionnaires (e.g., Davis, 1999; Pasveer & Ellard, 1998). However,

many of these studies used samples recruited from the Internet or some other non-random assignment for their studies. This is a critical flaw in an experiment because without random assignment, the equivalence of the groups administered items in alternative formats cannot be established using classical test theory methods that depend on sample equivalence. By using different samples and different test media, non-zero effects from sampling and media may cancel out. According to Buchanan and Smith (1999), Internet participants may make an active effort to locate experiments, and often do so for intrinsic rewards, like interest or curiosity. However, most traditional samples have other motivations, such as course credit or monetary rewards. These differences in motivation may alter the results of experiments. For example, people may be more forthright in their responses to the computerized media. However, samples drawn from the Internet might be less well adjusted due to self-selection. Therefore, finding no differences across test media and subject samples on a scale such as neuroticism does not necessarily mean that there were no differences due to the administrative format. To legitimately compare scores from the Internet to traditional paper-and-pencil scores, we must first ensure that the samples are drawn from the same population.

Another concern is the equivalency of the samples in terms of demographics. Some people have argued that samples recruited via the Internet are similar to traditional samples (i.e., college undergraduates). For example, Smith and Leigh (1997) looked at several demographic variables and found that samples recruited from the Internet were similar to introductory psychology subject samples. Conversely, Gosling, Vazire, Srivastava, and John (2004) reported that Internet-recruited samples were more diverse than traditional samples as published in a top tier psychology journal, but were nonetheless not representative of the general population. However, as Schmidt (1997) points out, the average Internet user is male, between his late-teens and early thirties, and has an above average socioeconomic and educational status. Therefore, we would not expect that a sample drawn from the Internet would reflect the demographic and socioeconomic characteristics of the population at large. In sum, the existing research assessing the equivalence of paper-and-pencil and Internet administration of personality measures has not controlled the medium or the samples taking the test. Consequently, equivalency cannot be determined using classical test theory approaches.

3. The present study

In the present study, participants were recruited from an Introductory Psychology subject pool and randomly assigned to one of three test administration conditions: paper-and-pencil, proctored computer lab and unproctored Internet. This was done rather than recruiting participants from the Internet because the purpose of the study was to assess the equivalence of the medium of administration, rather than the equivalence of samples drawn from the Internet community to traditional samples.

The three administration conditions were designed to examine factors that may contribute to differences in measurement properties. Specifically, the experimental design takes into account two factors that could potentially cause differential item functioning (DIF): proctoring and media. Traditional paper-and-pencil administrations are usually proctored. Internet administrations on the other hand, use computer administration and are not usually proctored. By including a proctored computer administration, we can disentangle proctoring effects from media effects. The proctored computer administra-

tion was conducted in a campus computer lab and was proctored; therefore it differed from the paper-and-pencil assessment only in the medium of administration. Analogously, the proctored computer lab version used the same software for administrating items as the unproctored Internet version and only differed in that respondents were administered the questionnaire in a proctored environment. In the event that differences are found between administration types, this experimental design allows us to identify which of the two factors—proctoring or administration medium—was responsible for the differences.

In addition to traditional statistical analysis this study contributes to the issue of equivalency by applying IRT methodology in the analysis. The application of IRT to the issue of Internet equivalency is important because it addresses several shortcomings in classical test theory (see Embretson & Reise, 2000). Unlike classical test theory, a subject's latent trait score is not dependent on the specific items of the test and—when the assumptions of the model hold—the item response functions do not depend on the subpopulation of respondents. With IRT, groups do not need to be identical when assessing equivalency of tests or test items, in contrast to classical test theory. IRT's invariance of item and ability parameters enables us to directly compare parameters from an item administered via different media. Moreover, IRT allows an examination of the measurement properties of individual items and consequently differences in measurement properties across media can thereby be identified. Thus, we are able to compare each item in its original paper-and-pencil format to the item in its Internet incarnation.

Previous research has shown that health behaviors and conscientiousness are positively related (Bogg & Roberts, 2004; Booth-Kewley & Vickers, 1994). We were therefore able to use criterion-related validity as a means of testing equivalence. By comparing the relationship between personality variables and health behaviors across media we attempted to establish that personality scales have equal predictive validity across administration media.

In summary, participants were randomly assigned to one of the three test administration conditions. Measurement equivalence was assessed using IRT methodology in addition to other traditional statistical procedures such as mean level comparisons, and factor structure. Criterion-related validity across media was also assessed to determine whether predictive validity was consistent across media.

4. Method

4.1. Participants

728 undergraduate students from a large Midwestern university enrolled in psychology courses participated in the study. Participants were randomly assigned to either the paper-and-pencil ($N = 266$, Male 40%, Female 60%), the proctored computer lab ($N = 222$, Male 42%, Female 58%), or the unproctored Internet ($N = 240$, Male 45%, Female 55%) administration condition. Respondents received course credit for their participation in the experiment.

4.2. Measures

The present study analyzed responses to adjectives selected to assess the Big 5 personality factors (neuroticism, extroversion, openness, agreeableness, and conscientiousness).

These items were drawn from Goldberg's (1992) 100 unipolar adjectival markers of the Big 5. To limit the undue influences of multidimensionality on our analyses, the 10 highest loading adjectives were selected for each of the personality dimensions, except for neuroticism, for which 13 adjectives were selected. The adjectives are listed in Appendix A. Participants in the study rated the adjectives according to how accurately the adjectives described their personalities using a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree).

The HBC was also administered as a means of assessing the equality of criterion-related validity across media. The HBC assesses preventive health behaviors, accident control, risk taking behavior, and substance risk. It is an established scale which has been administered in a number of previous studies (Vickers et al., 1990). Reliability estimates across samples for the HBC scales were adequate, ranging from .63 to .79 (mean = .69).

4.3. Test administration conditions

The paper-and-pencil sessions were administered in traditional proctored sessions. Respondents were administered the questionnaires in a large auditorium in groups of 50 or more, with a researcher present at all times. The proctored computer administration was conducted in the psychology department's computer laboratory facilities. There are between 10 and 18 computers in each computer laboratory. Consequently, 10–18 students were scheduled for each session and again a researcher was present at all times. Students in the unproctored Internet condition were contacted via e-mail. Instructions for the experiment and a unique username and password were included in the e-mail. The Internet respondents were given one week to respond to the questionnaires. They were specifically informed that they were able to leave and return to questionnaires as frequently as they wanted to within the experiment period. This feature, in conjunction with the unique username and password, allowed us to resolve the problem of multiple submissions from the same person. Obviously, a researcher was not present when students in this condition completed the questionnaire.

The computerized version of the questionnaire was adapted from the traditional paper-and-pencil version. Great care was taken to make sure that the computerized version was as similar in appearance to the paper-and-pencil version as possible. The computerized version was also designed to emulate the features of paper-and-pencil tests. Respondents were able to skip items and return to them later. Previous responses were also displayed so that respondents could review and revise their answers.

Students in the proctored computer lab and unproctored Internet administration conditions received the same computerized version of the questionnaires. The questionnaires were developed using Active Server Pages (ASP). This language uses server-side scripting to create dynamic websites. This means that the program can be deployed on any computer platform that is able to read standard HTML (Hypertext Markup Language). Most computer platforms have browsers that enable them to read HTML, including IBM compatible, Macintosh, and UNIX platforms. This enabled us to deploy the same version of the questionnaires in both the proctored computer and Internet-based administrations. The website required a unique username and password to ensure the security of the data.

Students in all the test administrations were informed by the department's centralized subject pool system about the location and time for the experiment. The centralized subject pool system allowed us to randomly assign participants to the various administrations conditions.

However, because of scheduling issues.[1] the numbers of participants in each of the administration conditions are not perfectly equal, even though they were conducted concurrently.

4.4. Statistical analyses

IRT is a model driven theory of measurement. The model must adequately fit data for IRT to provide meaningful insights. We decided to use a two-parameter logistic model for the data. Previous research supports the use of the two-parameter logistic model for personality data (Reise & Waller, 1990) for samples of the size collected here (Drasgow, 1989). A dichotomous IRT model was used because polytomous IRT models require larger sample sizes than were available in the current study (Bock, 1972; Levine, 1984; Samejima, 1969). The additional item parameters that need to be estimated in polytomous IRT models, such as Samejima's Graded Response model, increase parameter estimation error and reduce the power of the statistical analysis to find differences between the media of administration. Consequently, a two-parameter logistic model was adopted for the analysis of the data.

The two-parameter logistic model requires the use of dichotomous responses. The item scores were dichotomized with the middle point scored as a negative response (i.e., false or 0). Therefore, for the purposes of the analysis we dichotomized the responses such that 1, 2, and 3 were rescored as 0 and 4 and 5 were rescored as 1.

Before applying the two-parameter logistic IRT model, however, some assumptions should be checked. First, items on an assessment must be unidimensional for analysis by this model. A test or scale is unidimensional if responses to all the items can be accounted for by a single dominant trait. Principal axis factoring (PAF) is well suited to assessing unidimensionality in this situation because each personality scale has a relatively small number of items per scale. However, because the items were dichotomized, tetrachoric correlations should be used because the dichotomized variables reflect an underlying continuous scale. Other methods, such as Dimtest (Stout, 1987), generally require a larger number of items within a scale to assess unidimensionality adequately.

To assess measurement equivalence across media in the IRT framework, three differential item functioning methodologies were used: Mantel–Haenszel (Holland & Thayer, 1988), Sibtest (Stout & Roussos, 1995), and Lord's χ^2 (Lord, 1980). Differential item functioning (DIF) refers to items that do not function the same for people with equal standings on the latent trait but sampled from different groups. In other words, DIF can cause people with the same standing on the latent trait to have different raw scores when they are administered the assessment via different media.

The SIBTEST computer program (Stout & Roussos, 1995) calculates the simultaneous item bias (SIB) statistic (see Shealy & Stout, 1993, for explanation) and the

[1] Participants in the subject pool have the right not to participant in an experiment. It should be noted that none of the participants in the traditional paper-and-pencil test administration choose to withdraw from the experiment. We believe that after the students made a commitment in time and effort to come to the test location, they were reluctant to decline participation in a very innocuous experiment. Subjects scheduled for the proctored computer administration had to be scheduled in smaller groups per session because of the limited number of computers. Therefore, the experiments were scheduled across a broader range of times in the day. Interestingly, there were more absentees for time slots earlier in the morning. Again, no student declined participation after coming to the psychology building. The non-participation rate for the unproctored Internet condition was intermediate between absentee rate of the two proctored conditions.

Mantel–Haenszel statistic (see Holland & Thayer, 1988, for explanation). It should be noted that the SIBTEST program does not rely on any IRT model parameters, and therefore, the assumptions of the two-parameter logistic model are not required here. SIBTEST detects DIF by comparing the responses of individuals in the reference and focal groups that have been allocated to bins using their scores on a "matching subtest" (Stout & Roussos, 1995).[2] The matching subtest is a subset of items that are known or believed to not have DIF. Given that we do not have a priori knowledge as to which items in the test have no DIF, we used an iterative process to identify non-DIF items (Candell & Drasgow, 1988). Here, individual-item DIF analysis was performed on all items to flag potentially DIF items. This was done by assuming that all items were non-DIF except for one item which was under study. The other non-studied items were used as the matching subtest and a SIB statistic was calculated. This process was repeated for each item in the test. Subsequently, flagged items were removed from the set of items used for matching. The process of identifying potential DIF items was repeated for the remaining items and then was repeated until no further DIF items were identified. Consequently, items on the matching subtest pool had no DIF according to SIBTEST. Because we did not have an a priori hypothesis about the direction of the DIF for the items, a two-tailed hypothesis test, SIB-p ($z = 1.96$, $p < .05$), was used in our analysis.

Stark's (2001) ITERLINK program also uses an iterative procedure to link latent trait metrics and calculate Lord's χ^2 DIF statistic. We began by estimating item parameters using BILOG (Mislevy & Bock, 1991), with the default settings, for the two-parameter logistic model (except 40 quadrature points were used for numerical integration). Item parameters and their covariances were estimated separately for each group using BILOG. Linking constants were then computed by assuming no item had DIF. After applying the linking constants to the item parameter estimates, transformed item parameters were then compared using a DIF index (Lord's χ^2) to determine which items displayed DIF. Items displaying DIF were removed, new linking constants were computed, the scales were re-linked, and the DIF index was recomputed for all items. The process was then repeated and scales re-linked until the same DIF items were identified on consecutive trials.

In terms of a practical understanding of the effect of DIF, an effect size index would be useful in determining if statistically significant differences are more than trivial. To compute the effect size, we used Stark, Chernyshenko, and Drasgow (2004) index of differential test functioning effect size DTFR. DTFR is the expected total test score difference due to differential test functioning (DTF). To place the results on a commonly accepted metric, DTFT is divided by the standard deviation of the focal group's observed scores. This allows us to interpret the effect size using a *d*-score metric (Cohen, 1992).

Another means of assessing measurement equivalence across media is multi-group factor analysis. Equivalence was assessed using AMOS 4.0 by simultaneously fitting a model across groups with a single factor loading matrix. By comparing the fit between the constrained model and an unconstrained model where the factor loadings were freely estimated for each group, we assessed the equivalence of the factor loadings across test administrations. Small differences in fit between the unconstrained and the constrained models suggest that the constrained model provides an adequate representation. This

[2] Designation of which group is the "focal" group and which is the "reference" group is simply a matter of terminology. The focal group is typically a subpopulation of interest to the researcher, and the reference group serves as the standard for comparison.

would suggest that the different groups based, in this case on administration media, have equivalent factor loadings.

4.5. Website security

Several researchers have pointed out that web-based data collection may be compromised by multiple entries or disingenuous responses (Davis, 1999; Schmidt, 1997). Unless one is planning to passively or randomly recruit participants from the Internet, security protocols are a key component in controlling access to Internet surveys. This curtails access from people outside the sample being studied and prevents multiple submissions from a single respondent. To resolve these issues, a unique username and password were required to access to our Internet-based questionnaire.

The use of a unique username and password means that the identity of the respondent may be traced. This loss of anonymity may affect the degree of socially desirable responding (Richman et al., 1999). However, tracking a respondent's identity is important under practical assessment situations. In many academic institutions, course credit is given for participation in experiments. In this case, confidentiality of the participant's responses can be guaranteed for respondents to Internet surveys, but not complete anonymity. Similarly, in non-academic settings, it is often useful to be able to tie together information about a person to other sources of data. For example, a person's personality profile may be linked to his or her job performance.

5. Results

5.1. Unidimensionality

Tetrachoric correlations were first calculated and PAF was used to assess the number of factors for each scale. The results suggested that the 'openness' dimension was not unidimensional, nor were there sufficient items to form separate sub-dimensions. Therefore, we were unable to assess the equivalence of 'openness' across media. The PAF also suggested that two items in the 'neuroticism' scale loaded on a second factor. Because the purpose of this study was to assess the equivalence of test media, and not to establish the validity of the scales, the two items, 'emotional' and 'fretful', were dropped from the analysis to satisfy the unidimensionality requirement. All other scales were found to be unidimensional according to guidelines provided by Drasgow and Lissak (1983).

5.2. Differential item functioning

The SIBTEST (Stout & Roussos, 1995) and ITERLINK (Stark, 2001) programs were used to computer differential item functioning (DIF) statistics; Table 1 lists the items with DIF significant at $p < .05$. Fig. 1 illustrates the item response function for one item ("untalkative" from the Extroversion scale) across the three media. If DIF is present there should be significant deviations between the response functions across media. As we can clearly see the item response functions overlap considerably indicating that there is no DIF. However, when using a .05 Type I error rate we expect to find a number of false positives due to the multiple significance tests. We can calculate the number of expected number of DIF items by multiplying the alpha level ($\alpha = .05$) by the number of comparisons we

Table 1
DIF results for big 5 personality traits

	Paper vs Internet	*p*	Paper vs computer	*p*	Computer vs Internet	*p*
SIB statistic						
Neuroticism	Unenvious	.047	Imperturbable	.000	Imperturbable	.042
	Unexcitable	.008				
Extroversion			Complex	.005		
Agreeableness					Sympathetic	.042
Conscientiousness	Inefficient	.018	Systematic	.041		
			Unsystematic	.006		
Mantel–Haenszel						
Neuroticism	Unexcitable	.037	Imperturbable	.000		
			Unexcitable	.039		
Extroversion						
Agreeableness						
Conscientiousness	Inefficient	.019	Systematic	.041		
			Unsystematic	.006		
Lord's χ^2						
Neuroticism						
Extroversion						
Agreeableness						
Conscientiousness			Systematic	.001		
			Neat	.025		
			Disorganized	.011		
			Unsystematic	.036		

Note. Paper-and-pencil, $N = 266$; Computer, $N = 222$; and Internet, $N = 240$.

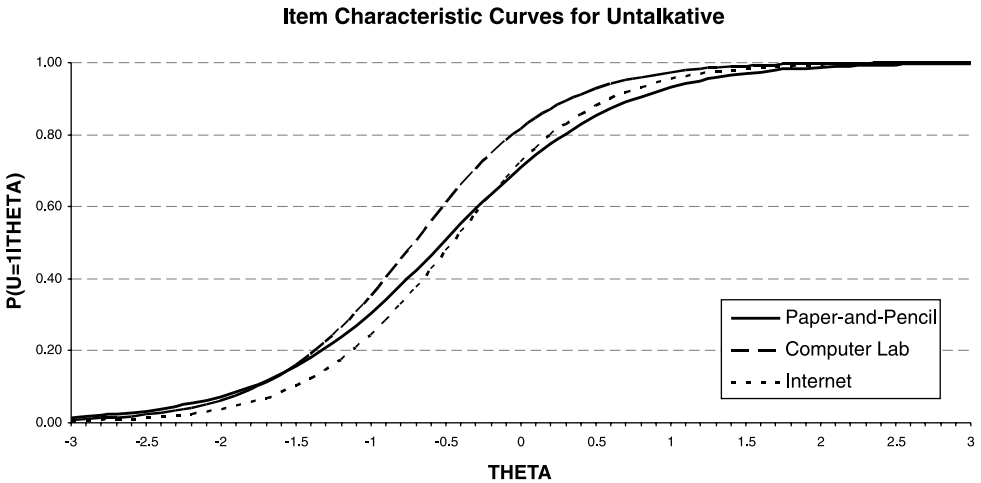

Fig. 1. Item characteristic curves for item untalkative, from the extroversion scale.

made. Moreover, an approximate confidence interval can be constructed from the binomial distribution by assuming independence across comparisons. For a dimension with 10 items, there were 30 comparisons made. If there is a .05 chance of false rejection, 1.5 false positives are expected; between 0 and 4 items would be falsely rejected 95% of the time

Table 2
Number of items showing DIF at $p < .05$

	SIB statistic	Mantel–Haenszel	Lord's χ^2	Probable range
Neuroticism	4	3	0	0–4
Extroversion	1	0	0	0–4
Agreeableness	1	0	0	0–4
Conscientiousness	3	3	4	0–4
	Paper vs Internet	Paper vs Computer	Computer vs Internet	Probable range
SIB statistic	0	4	0	0–5
Mantel–Haenszel	2	4	0	0–5
Lord's χ^2	3	4	2	0–5

Note. The top part of the table provides the number of significant DIF items aggregated across response formats and the lower part of the table aggregates across personality scales. Paper-and-pencil, $N = 266$; Computer, $N = 222$; and Internet, $N = 240$.

when no DIF is present. The number of items showing DIF is summarized in Table 2. These results are consistent with the number of false rejections that we expected to find when no DIF was truly present.

From a substantive perspective, we attempted to interpret the distribution of items identified as having DIF to see if there were any meaningful patterns. Specifically, results were examined in regard to the two issues being addressed in this experiment, and administration medium. Based on the design of the experiment, DIF would logically be identified between the paper-and-pencil version and the two electronic versions (Internet and proctored computer lab), or between the Internet version and the two proctored administrations (paper-and-pencil and proctored computer lab). However, if we look at Table 2, which shows the distribution of items identified as having DIF, DIF occurred most frequently in the comparison between the paper-and-pencil and the proctored computer lab administrations. If there was truly DIF between the paper-and-pencil version and the electronic versions, we would expect to find a similar number of DIF items between the paper-and-pencil and the Internet administration. Because this was not found, it appears that there were no real differences across modality.

Additionally, if we look at the effect size indices for DTF in Table 3, we can see that there are negligible effect sizes for DTF. The DTF effect sizes (d_{DTF}) range between −.20 and .19, which according to Cohen (1992) are small or negligible. This suggests that the few

Table 3
DTF and mean level ability estimate effect size values for the personality scales

	Paper (F) vs Internet (R)		Paper (F) vs Computer (R)		Computer (F) vs Internet (R)	
	d_{DTF}	d	d_{DTF}	d	d_{DTF}	D
Neuroticism	−0.15	−0.16	−0.20	−0.26	0.05	0.01
Extroversion	−0.17	−0.16	−0.17	−0.16	0.01	0.00
Agreeableness	−0.03	−0.02	0.08	−0.01	−0.14	−0.02
Conscientiousness	0.16	0.01	0.19	0.07	−0.02	0.01

Note. R, reference group; F, focal; d_{DTF}, DTF effect size; d, effect size between mean level ability estimates; A positive DTF effect size indicates bias against the focal group.

items that were found to be statistically significant did not have much of an impact on the overall scale equivalence.

Fig. 2 presents the test characteristic curve for the Extroversion scale. The test characteristic curve is the sum of the item response functions and gives the expected observed score as a function of the latent trait. Fig. 2 clearly shows that the differences across the three administrative media are nugatory.

5.3. Factor equivalence

Next we created item parcels from each personality scale for the purposes of the analysis. This was done so that each personality dimension had multiple indicators, and that each indicator would be more closely approximate a normal distribution. Each personality scale consists of 10 items, except Neuroticism which consisted of 11 items The 2 items eliminated from the Neuroticism scale for the IRT analysis were also dropped here. 3 parcels were created from each scale with 2 parcels consisting of the mean score of 3 items each and 1 parcel with the mean score of 4 items (2 parcels of 4 items and 1 parcel with 3 items for Neuroticism).

An analysis of the statistical significance of the difference in fit of the constrained and free models indicated that the difference in χ^2 was non-significant ($\chi^2_{[16]} = 24.1$, $p > .05$), yielding a χ^2 to degrees of freedom ratio of 1.51. Additionally, the three separate factor solutions for the paper-and-pencil, proctored computer and Internet administrations yielded RMSEA values of .092, .096, and .106, respectively. The simultaneous constrained factor solution yielded an RMSEA of .054. The simultaneous constrained factor solution offers a substantial gain in parsimony with no decrement in fit. In sum, based on the fit statistics, we believe our factor analysis results demonstrate the cross-media equivalency of personality assessment. In fact, the consistency between the IRT analysis and the factor-equivalence analysis provides strong evidence for the equivalency across media.

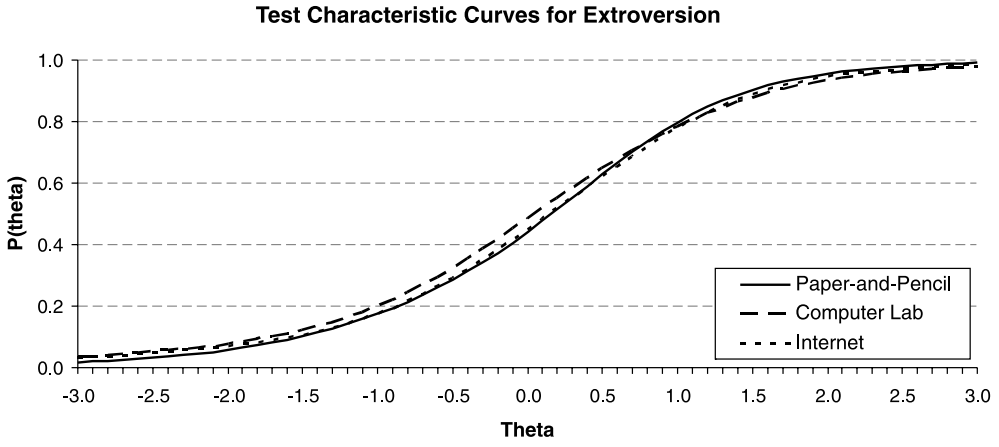

Fig. 2. Test characteristic curves for the extroversion scale.

Table 4
Mean level of ability estimates on the personality scales

	Paper-and-pencil		Computer		Internet	
	Mean	*SD*	Mean	*SD*	Mean	*SD*
Emotional stability	0.11	0.81	−0.10	0.80	−0.03	0.80
Extroversion	0.10	0.92	−0.05	0.95	−0.05	0.86
Agreeableness	0.01	0.82	0.00	0.82	−0.02	0.79
Conscientiousness	−0.04	0.84	0.03	0.87	0.04	0.89

Note. Paper-and-pencil, $N = 266$; Computer, $N = 222$; and Internet, $N = 240$.

5.4. Mean level analysis

The BILOG (Mislevy & Bock, 1991) program calculates an estimate of each respondent's latent trait value for each dimension. Because they were randomly assigned to conditions, we would expect no differences if the medium of administration has no effect. One-way analysis of variance was used to compare ability estimates across conditions. Mean ability estimates did not differ for three of the four personality dimensions; see Table 4.Using a Bonferroni adjustment for multiple comparisons ($\alpha/4 = .05/4 = .0125$), there were no significant differences. Similarly, the effect size for differences between the mean ability estimates are relatively small if use Cohen's (1992) standard for evaluating effect sizes. The effect size (d) values range between −.26 and 0.07, see Table 3.

5.5. Criterion-related validity

The equivalence of correlations between health behaviors and personality was assessed using a statistical test for the difference between independent correlations (Bruning & Kintz, 1987). The BILOG estimates of ability across each personality variable were correlated with the four health behavior dimensions (preventive health behaviors, accident control, risk taking behavior, and substance risk) across each media; Table 5 presents these results. The results suggest that the relationship between personality variables and health behaviors are equivalent across media. The only statistically significant difference between the Internet and paper-and-pencil administration ($p = .04$) was between emotional stability and substance risk. However, given the large number of correlations tested (48), finding a single significant difference is likely to be a Type I error.

6. Discussion

The APA Taskforce on Internet Testing (Naglieri et al., 2004) has emphasized the importance of establishing measurement equivalence when a test is converted for administration on the Internet. Consequently, researchers have been reluctant to fully exploit the benefits of Internet administered questionnaires because they have had misgivings about the equivalence of this new medium with traditional paper-and-pencil assessments. We have attempted to address this issue using a controlled experiment, an IRT analysis, a multi-group factor analysis, and an evaluation of criterion-related validity.

Table 5
Correlations between personality scales and health behavior scales across media

	Preventive health behaviors	Accident control	Risk taking behavior	Substance risk
Emotional stability				
Paper-and-pencil	0.03	0.10	0.01	−0.08
Computer	−0.08	0.00	0.08	0.11
Internet	0.01	0.07	−0.01	−0.06
Extraversion				
Paper-and-pencil	0.21*	0.14*	0.04	0.15*
Computer	0.11	0.10	−0.04	0.08
Internet	0.18*	0.04	0.11	0.08
Agreeableness				
Paper-and-pencil	0.19*	0.18*	−0.13*	−0.11
Computer	0.12	0.12	−0.14*	−0.07
Internet	0.18*	0.08	−0.11	−0.18*
Conscientiousness				
Paper-and-pencil	0.24*	0.25*	−0.19*	−0.11
Computer	0.22*	0.34*	−0.18*	−0.11
Internet	0.20*	0.22*	−0.18*	−0.28*

Note. $N = 262$ for paper-and-pencil; $N = 215$ for Computer; and $N = 230$ for Internet.
* $p < .05$, two-tailed.

Our experiment was designed with three conditions: a traditional paper-and-pencil, an Internet-based, and a proctored computer lab condition. The third condition was included in an attempt to determine, in the event that there were differences between the paper-and-pencil and Internet-based condition, if either computerization or unproctored administration was the source of the disparity. The proctored computer condition used software that was identical to the Internet-based version and supervision that was virtually identical to the paper-and-pencil administration.

Subjects were not recruited from the World Wide Web because we wanted to randomly assigned individuals to the three different conditions. Experimenter control over subject assignment was important to avoid self-selection effects, and enable a true experimental design.

No significant systematic differences were found in measurement properties across administration conditions. There were a small number of items flagged with statistically significant p value DIF statistics (see Table 1). However, these were well within the range of expected false positives expected when using an $\alpha = .05$ confidence level that was not adjusted for multiple comparisons.

Additionally, a majority of the false positives were found between the proctored computer condition and the other two conditions; surprisingly, there were few instances of DIF items between the Internet and paper-and-pencil administrations. This provides further support that the flagged items were false positives. If the Internet and paper-and-pencil administrations were truly different, we would expect more DIF items between these two administrations, and we should have seen more agreement between the proctored computer administration and one of the other test administrations. According to the theoretical framework, Internet and paper-and-pencil administrations are conceptually distinct

because of the differences in administration media and supervision whereas the proctored computer lab administration represents a midpoint between the Internet and paper-and-pencil administrations. However, few differences between the Internet and paper-and-pencil administrations were found. Therefore, we feel confident in discounting the significant differences we found.

This conclusion was further reinforced by the results of the multi-group factor analysis. Goodness of fit statistics supported the hypothesis of equivalence across media. We further assessed the equivalency of test administration by comparing the correlations between the personality dimensions and health behaviors across media. The lack of statistically significant differences across media, with one single exception, supports our assertion that the assessments of personality tests are equivalent across media. Taken in total, evidence from the IRT, factor-equivalence analysis, ANOVA, and criterion validity, constitutes strong evidence for the equivalence of Internet administrations of personality questionnaires with paper-and-pencil administrations.

6.1. Limitations

The findings from this investigation should be interpreted with caution. Previous studies examining Internet-based questionnaires have reported contradictory evidence both supporting (Davis, 1999; Pasveer & Ellard, 1998) and refuting (Ployhart, Weekley, Holtz, & Kemp, 2003) their equivalence to paper-and-pencil questionnaires. This discrepancy may best be explained by the theory of social desirability distortion (see Richman et al., 1999). Richman et al. (1999) point out that the features of a computerized test administration can have significant effects on responses. These distortions or dissimilarities with paper-and-pencil test administration were minimized in the present study. For example, the proctored computer version shared features with the paper-and-pencil version such as anonymity and the ability to backtrack. Little is known about the specific features of the Internet-based questionnaires utilized in previous research, and differences in the execution of the software or samples may explain their results. We should therefore emphasize that at least one Internet-based personality questionnaire (i.e., the Goldberg scales we studied) appears equivalent to its paper-and-pencil version when the Internet questionnaire resembles the paper-and-pencil version in its ability to review answers and skip items. Interestingly, in this experiment, having a unique username and password which identified the respondent did not appear have a meaningful effect.

Another limitation of this study is that all the participants were college undergraduates. Further research should replicate this study with samples from the general population or in an industrial setting. College undergraduates may have different motivations for taking the questionnaire than a person applying for a job, for example. Job applicants have real world consequences, and thus may be more motivated to 'fake good' to present a more positive impression to the prospective employer. Further research using alternative samples as well as different personality scales is needed before we close the subject of the equivalence of Internet-based questionnaires.

6.2. Implications

The findings of this study are important to a broad range of parties. However, we believe they are most important to three groups: researchers, test publishers, and test tak-

ers. The first group includes researchers using personality tests, such as industrial and organizational psychologists and personality psychologists, among others. The ability to deploy tests over the Internet is a considerable convenience and cost saving for these researchers. It alleviates the need for proctoring tests and manual data entry. Additionally, the infrastructure required to deploy Internet-based questionnaires is very affordable. The only cost would be an inexpensive computer to host the survey and associated software. The unobtrusive nature of Internet questionnaires is also advantageous when studying real world samples. For example, managers at a company may object to pulling people from their work to answer a survey. However, they may have fewer objections if their employees can access the survey during a break from work.

The second group to benefit from this study consists of test publishers. There are obvious monetary savings from Internet administration and scoring. Internet-based questionnaires also allow all the test materials to be maintained at a central location. Any changes can be made conveniently and nearly instantaneously. Additionally, moving to a computerized format allows the test publisher to include useful features such as randomizing the order of items and customizing an assessment for a particular individual. Computerizing the test also makes the change towards adaptive measurement easier.

Lastly, test takers are likely to appreciate the convenience of Internet-based tests. Any Internet equipped computer can administer the test. This means that the test takers can take the test from the comfort of their home at convenient times.

In so far as personality assessment is concerned, we believe that this study provides some evidence that the differences between computerized and paper-and-pencil measures of personality may be small, at least when administered to college students. With computers becoming cheaper and more widespread, Internet-based deployment make sense. However, particular features of an Internet questionnaire can change the psychometric qualities of the responses (Richman et al., 1999). Further research is necessary to clarify which features have the strongest impact and to generalize these findings beyond the conditions studied here.

Appendix A

Big 5 personality items

	Neuroticism	Extroversion	Openness	Agreeableness	Conscientiousness
1	Unenvious	Extraverted	Intellectual	Kind	Organized
2	Unemotional	Talkative	Creative	Cooperative	Systematic
3	Relaxed	Assertive	Complex	Sympathetic	Through
4	Imperturbable	Verbal	Imaginative	Warm	Practical
5	Unexcitable	Energetic	Bright	Trustful	Neat
6	Anxious[a]	Introverted[a]	Unintellectual[a]	Cold[a]	Disorganized[a]
7	Moody[a]	Shy[a]	Unintelligent[a]	Unkind[a]	Careless[a]
8	Temperamental[a]	Quiet[a]	Unimaginative[a]	Unsympathetic[a]	Unsystematic[a]
9	Envious[a]	Reserved[a]	Uncreative	Distrustful[a]	Undependable[a]
10	Irritable[a]	Untalkative[a]	Simple[a]	Harsh[a]	Inefficient[a]
11	Jealous[a]				
12	Emotional[a]				
13	Fretful[a]				

[a] Reverse scored Items.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Butcher, J. N. (1987). The use of computers in psychological assessment: An overview of practices and issues. In J. N. Butcher (Ed.), *Computerized psychological assessment* (pp. 3–14). New York, NY: Basic Books.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29–51.

Bogg, T., & Roberts, B. W. (2004). Conscientiousness and health behaviors: A meta-analysis of the leading behavioral contributors to mortality. *Psychological Bulletin, 130*, 887–919.

Booth-Kewley, S., & Vickers, R. R., Jr. (1994). Associations between major domains of personality and health behavior. *Journal of Personality, 62*, 281–298.

Bruning, J. L., & Kintz, B. L. (1987). *Computational handbook of statistics*. Glenview, IL: Scott, Foresman.

Buchanan, T., & Smith, J. L. (1999). Using the internet for psychological research: Personality testing on the World Wide Web. *British Journal of Psychology, 90*, 125–144.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12*, 253–260.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

Davis, R. N. (1999). Web-based administration of a personality questionnaire: Comparison with traditional methods. *Behavior Research Methods, Instruments and Computers, 31*, 572–577.

Drasgow, F., & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology, 68*, 363–373.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement, 13*, 77–90.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4*, 26–42.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions. *American Psychologist, 59*, 93–104.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel–Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Joinson, A. (1999). Social desirability, anonymity, and Internet-based questionnaires. *Behavior Research Methods, Instruments, and Computers, 31*, 433–438.

Kirsch, I., Jamieson, J., Taylor, C., & Eignor, D. (1998). *Computer familiarity among TOEFL examinees*. (TOEFL Research Report 59). Princeton, NJ: Educational Testing Service.

Levine, M. V. (1984). *An introduction to multilinear formula score theory* (Measurement Series 84-4). Champaign, IL: University of Illinois, Department of Educational Psychology.

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

Mislevy, R. J., & Bock, R. D. (1991). *BILOG users' guide*. Chicago, IL: Scientific Software.

Naglieri, J., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., & Velasquez, R. (2004). Psychological testing on the Internet: New problems, old issues. *American Psychologist, 59*, 150–162.

Pasveer, K. A., & Ellard, J. H. (1998). The making of a personality inventory; Help from the WWW. *Behavior Research Methods, Instruments and Computers, 30*, 309–313.

Ployhart, R. E., Weekley, J. A., Holtz, B. C., & Kemp, C. (2003). Web-based and paper-and-pencil testing of applicants in a proctored setting: Are personality, biodata, and situational judgment tests comparable? *Personnel Psychology, 56*, 733–752.

Powers, D. E., & O'Neill, K. (1992). *Inexperienced and anxious computer users: Coping with a computer-administered test of academic skills. The Praxis series: Professional assessments for beginning teachers.* Princeton, NJ: Educational Testing Service.

Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement, 14*, 45–58.

Richman, W. L., Weisband, S., Kiesler, S., & Drasgow, F. (1999). A meta-analytic study of social desirability response distortion in computer-administered and traditional questionnaires and interviews. *Journal of Applied Psychology, 84*, 754–775.

Rosenfeld, P., Giacalone, R., Knouse, S., Doherty, L., Vicino, M., Kantor, J., & Greaves, J. (1991). Impression management, candor, and microcomputer-based organizational surveys: An individual differences approach. *Computers in Human Behavior, 7*, 23–32.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monographs*, 34 (Suppl. 17).

Schmidt, W. C. (1997). World-wide web survey research: Benefits, potential problems, and solutions. *Behavior Research Methods, Instruments & Computers, 29*, 274–279.

Shealy, R., & Stout, W. (1993). An item response theory model for test bias and differential test functioning. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197–240). Hillsdale, NJ: Erlbaum.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology, 89*, 497–508.

Stark, S., 2001. Iterlink.[On-line]. Available: <http://io.psych.uiuc.edu/irt/downloads>.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589–617.

Stout, W., & Roussos, L. (1995). *SIBTEST user manual*. Urbana, IL: University of Illinois.

Smith, M. A., & Leigh, B. (1997). Virtual subjects: Using the internet as an alternative source of subjects and research environment. *Behavior Research methods, Instruments, and Computers, 29*, 496–505.

Vickers, R. R., Jr., Conway, T. L., & Hervig, L. K. (1990). Demonstration of replicable dimensions of health behaviors. *Preventive Medicine, 19*, 377–401.