Abstract

HUFF, KYLE C. The Effects of Mode of Administration on Timed Cognitive Ability Tests. (Under the direction of Dr. Joan J. Michael.)

Although widely used, there exists very little published research on the equivalence of web-based cognitive ability tests used for employee selection to their original paper-and-pencil versions. This issue is even further complicated by the limited research into the effects of proctoring on these types of tests. To investigate this issue, data were analyzed from the Wonderlic Personnel Test (WPT) and the Wonderlic Personnel Test-Quicktest (WPT-Q). Using the Differential Functioning of Items and Test (DFIT) procedure, data from 325 paper-and-pencil WPT administrations were compared to 325 web-based proctored administrations of the test. To check for the effects of proctoring, 108 proctored administrations of the WPT-Q were compared to 104 unproctored administrations again using the DFIT procedure. The results indicated that although the differences in administration produced low levels of differential item functioning (DIF), there was enough DIF to warrant conducting new validation studies when changing the mode of administration.

The Effects of Mode of Administration on Timed Cognitive Ability Tests

By

Kyle Huff

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

Doctor of Philosophy

March 27, 2006

Industrial/Organizational & Vocational Psychology

Approved By:

_____
Dr. Joan J. Michael
Chair of Advisory Committee

_____ _____
Dr. Mark A. Wilson Dr. John Fleener

_____
Dr. Paul Mulvey

Date filed with the Department: _____

Dedication

I would like to dedicate this dissertation to my friends and family. Without their support and encouragement, I could never have made it this far. I would also like to dedicate this to the memory of my cousin, Corporal Justin Huff, USMC, whom I will always remember as a fine man, a good Marine, and a great person.

Biography

Kyle C. Huff was born on March 24, 1975 in St. Louis, MO. Kyle's family resides in Dunwoody, GA. After completing high school in Dunwoody, he attended the Georgia Institute of Technology. He graduated in 1998 with a BS in Management and two certificates: one in Social/Personality Psychology and the other in Industrial/Organizational Psychology.

In the fall of 1998, he went on to attend Georgia College and State University in Milledgeville, GA. While there, he supported himself working as a research assistant conducting research on a variety of topics. He graduated from there with a MS in Psychology in 2000.

In the spring of 2000, Kyle was accepted into the Industrial/Organizational Psychology program at North Carolina State University. He has spent the last 6 years as a full time graduate student.

Acknowledgments

I would like to thank Wonderlic Inc for their generous support of this research. Without them providing the data and test materials, this research would never have been possible. I want to thank my advisor, Joan J. Michael, for her guidance on this project. I would also like to thank Adam Meade, Bart Craig, and Phillip Braddy for their assistance with the IRT analysis. Finally, I would like to thank the rest of my committee, John Fleenor, Paul Mulvey, and Mark Wilson for their patience and help in the final stages of my draft proposal.

Table of Contents

List of Tables

List of Figures

List of Appendices

The Effects of Mode of Administration on Timed Cognitive Ability Tests

The introduction of the computer to industrial/organizational psychology and human resources has had a dramatic impact, and the results so far have been encouraging. In general, reduced workload and increased productivity and efficiency have resulted (Beaty, Fallon, & Barrett, 2002). This, of course, lets the human resource departments become more strategic in nature. Unfortunately, computer technology has traditionally not been well researched in the field. It is only recently that computers and the Internet have received increased attention. Despite the recent advances in our understanding of the effects of the computer in I/O psychology, much still needs to be done.

The primary aim of this paper was to expand what little is known about web-based cognitive ability testing. To accomplish this goal, several different aspects of computers in psychological testing were covered. First, it was important to understand the role that computers have had in psychological testing. Next, there is a discussion on computer-based testing. Finally, the issues involved with web-based testing have been explored.

Evolution of Computer-Based Testing

Despite the relatively recent popularity of web-based assessments, the area enjoys a rich history that extends back several decades. The first use of computers in psychological assessment served as an alternative to hand scoring, which is tedious, time consuming, and error prone. Having computers score the test results had several

advantages. Although psychologists who used this service enjoyed saving time, probably the greatest advantage was the reduction in scoring error (Gregory, 2000).

As time progressed, the computer's impact on psychological testing continued to evolve. The first computer-administered test appeared in the 1970s. Eventually, computers conducted the entire testing process, from administration to report generation, thereby resulting in even more advantages in terms of timesavings, cost, and error reduction. The popularity of computer-based testing continued to grow so that now, at a minimum, large test publishers offer computer-based report generation (McBride, 1998; Gregory, 2000).

The next logical step in this evolution was moving psychological testing to the Internet. Web-based psychological testing offers even more advantages than computer-based testing. However, the popularity of this type of testing grew very quickly, raising concern about the lack of empirical research. In fact, there was so much concern that the American Psychological Association formed a task force in the fall of 2000 to study web-based psychological testing (Naglieri, Drasgow, Schmidt, Handler, Priftera, Margolis, & Velasquez, 2004).

At each stage in the evolution, new problems were realized. For example, equivalence or invariance issues between computer administered tests and the traditional paper and pencil administrations were realized early. Later, when tests were moved to the Internet, issues such as proctored versus unproctored tests arose. Throughout the current study, several of these issues have been identified and explored in more detail; however, it should be noted that the issues identified in this study are by no means exhaustive.

Computer-Based Testing Concerns

Probably the greatest impact of computers in the testing field came with computer-based tests (CBT). As previously mention, CBT has several advantages over traditional paper-and-pencil tests. One advantage is that the same computer that administers the test can also score, interpret, and generate reports, thereby saving time, paper, and money. Additionally, CBT allows test developers to realize the application of computer adaptive testing. Computer adaptive tests are procedures that allow for accurate and efficient measurement of a construct. Adaptive testing offers an advantage of shorter tests, which save time, while increasing efficiency and power. Third, a potential exists to expand the content and presentation styles. This potential can include incorporating multimedia applications into tests as well as testing in ways that previously were not feasible with paper-based tests. Finally, the computer has the potential to measure facets that cannot be measured with paper-and-pencil tests or to measure existing facets in new ways (McBride, 1998; Gregory, 2000).

As the advantages of computer-based testing became more apparent, major US organizations have moved, or are in the process of moving, from paper-and-pencil tests to computer-based tests for employee selection. Various problems have been encountered as this stage emerged. First, the question arose concerning whether computerized versions of tests are as reliable or valid as the paper-and-pencil tests. Research has demonstrated that the psychometric properties of power tests appear to exhibit small or even non-existent differences between paper-and-pencil and computerized versions. Power tests are tests that allow plenty of time for a subject to complete all items, but the items are difficult enough so that very few examinees, if any, get a perfect score.

However, speed tests, which are timed tests that contain items of uniform low difficulty, but have time constraints that make it make it difficult for most people to complete all items, can vary on both psychometric characteristics and construct composition (Buchanan, 2000; Gregory, 2000; McBride, 1998). A third type of test, a power test with a time limit is a combination of the two previous types of test. Examples of this type of test include the *Graduate Record Exam* and the *Scholastic Achievement Test*. Research has demonstrated that a well-developed power test with a time limit seems to maintain the original properties of the paper-and-pencil version even in a computerized version (Mead & Drasgow, 1993).

A second concern relates to the technology itself. Computer systems available for purchase today are overall better than the computers manufactured just a few years ago. Variations in operating systems, processing time, hard drive seek time, and displays can all have an impact on testing (Buchanan, 2000; Gregory, 2000; McBride, 1998).

Web-Based Testing

Web-based testing (also called Internet-based testing) has some distinct properties that make it different from computer-based testing. The most noticeable difference between the two administration modes is that a web-based test is administered through a web browser, whereas a computer-based test is not. Another major difference that could exist between the two methods is the Internet connection. A test that is administered over a high-speed Internet connection would enable the web-based test to appear very similar to a computer-based test. However, the slower this connection is, the greater the difference. Slower connection speeds could cause the test taker to have to wait for longer periods of time for the test items to be displayed, answers to be sent to the server, and

feedback to be sent to the test taker. Since a web-based test is administered through a web browser and possible differences in connection speed exist, Internet-based tests should be considered as a different administration mode from computer-based tests.

Advantages of web-based tests. Despite these differences, all of the advantages and disadvantages that exist for computer-based tests also exist for web-based tests. However, web-based tests have additional advantages over computer-based tests. The most obvious advantage is that it is possible for an individual to complete a web-based test at nearly any time and at any place that they wish. In addition to this, almost any computer lab can become a testing center. These two advantages are probably the most cited reasons for engaging in web-based testing. Also, prior research has shown that web-based tests can have increased psychometric properties over their paper-and-pencil counterparts. Additional advantages include researchers being able to collect data much more rapidly, conveniently, at lower costs, and from populations that are traditionally difficult to reach. New tests can also be made available throughout the world almost instantly. As a result of the minimal costs associated with web-based testing as well as the scalability of web-based systems, additional test administrations have lower costs than those for any other administration mode. Furthermore, the direct input of answers by test-takers provides more accurate scoring than do traditional paper-and-pencil tests. Finally, test norms can be continuously and immediately updated (Buchanan & Smith, 1999; Bridgeman, Lennon, & Jackenthal 2003; Lievens & Harris, 2003; Naglieri, Draasgrow, Schmit, Handler, Prifitera, Margolis, & Velasquenz 2004; Polyhart, Weekly, Holtz, & Kemp, 2003; Oswald, Carr, & Schmidt, 2001).

The advantages that can come from the use of a web-based test can be seen in the following example. A government agency implemented an on-line selection system. Their goal was to select current employees for participation in an information-technology (IT) training course. The application process had four stages. During stage one, interested employees filled out an application on-line. Those who possessed the minimum requirements for the training were sent an e-mail inviting them to take an on-line psychological test for stage two. Four hundred fifty applicants took the on-line test at stage two. From that group, seventy-six individuals were selected for stage three. In stage three, applicants took a proctored computer-based test similar to the first one. Applicants who successfully passed this stage were invited for interviews (Beaty, Fallon, & Barrett, 2002).

In, the end, the government agency selected sixty employees to participate in the IT training. If the test had not been given on-line, it was estimated that it would have taken six weeks to administer the test to every applicant. With the web-based system, the proctored testing was finished in nine days. It was estimated that by using this system there was a 500% savings in time for the organization (Beaty, Fallon, & Barrett, 2002).

Web-based testing issues. In addition to all of the advantages, there are a number of issues that are associated with web-based testing. One of the major issues associated with this type of testing, and a major question addressed by this paper, is that of administration mode, proctored vs. unproctored and computer-based vs. web-based vs. paper-and-pencil. Another important issue associated with web-based testing is that web-based delivery systems have to deal with a broader range of computer hardware and software settings than do computer-based tests. As a result of the broad range of

hardware and software settings, the amount of information that is displayed on the screen and the legibility of that information may have an impact on test scores. Also, test security and cheating have received much attention recently. Another popular issue is the question of whether various groups have as equal access to the Internet as do other groups. Finally, there is the possibility of normative issues arising with web-based tests. These normative issues come into effect when someone from a different population than that for which the test was designed takes the test. (Bridgeman, Lennon, & Jackenthal, 2003; Naglieri, Draasgrow, Schmit, Handler, Prifitera, Margolis, & Velasquenz, 2004; Polyhart, Weekly, Holtz, & Kemp, 2003; Epstein, Klinkenberg, Wiley, & McKinley, 2001; Lievens & Harris, 2003). Research studying these issues has lagged behind the development of new technologies and web-based testing trends.

Of the previous issues, equivalence of web-based tests to their paper-and-pencil or computer-based counterparts was selected as the primary topic of this study. In general, most of the research that has been conducted on the equivalence of web-based testing occurs in the personality arena with very little testing occurring for web-based ability tests. Thus, it appears that there is a need to demonstrate the equivalence of web-based ability tests.

One of the few studies to demonstrate the equivalence of the web-based version with a paper-and-pencil version on web-based cognitive ability tests was conducted with the *Wonderlic Personnel Test* (Dembowski & Callans, 2000). In this study, participants were divided into two groups, each taking a web-based version and a paper-and-pencil version. Additionally, each participant completed two alternate forms of the tests. One of these forms was completed in the web-based administration and the other was

completed in the paper-and-pencil administration. The entire design was counter balanced. Overall, the researchers were able to demonstrate equivalence, not only between the two alternate forms, but also between the administration modes.

However, despite the finding of equivalence, the Dembowski & Callans (2000) study had a major issue. The analysis consisted of t-tests between *Wonderlic Personnel Test* forms and method of administration, a Wilcox Signed Rank test to compare distributions, and a Spearman Rho to compute the correlation between forms. While the analysis of the data revealed encouraging results, these types of studies have been criticized, most notably by Epstein, Klinkenberg, Wiley, and McKinnley (2001) for using limited statistical comparisons. Another important limitation of this study is that it did not test to see if the web-based version of the *Wonderlic Personnel Test* is equivalent to it's paper-and-pencil counter-part in an unproctored setting.

Further complicating the issue of establishing the equivalence of web-based measures, is this question of whether or not a test is administered in a proctored or in an unproctored environment. Proctored web-based testing occurs when the test-taker completes the test form in the presence of a test administrator. Unproctored testing occurs when a test-taker completes the test form in any location with Internet access and without direct supervision of a test administrator (Polyhart, Weekly, Holtz, & Kemp, 2003). Sinar and Reynolds (2004) have noted that more and more web-based testing is being administered in unproctored settings.

The question was then raised whether a test administered in an unproctored mode actually fits the definition of a standardized test. Namely, a standardized test can be defined as a systematic procedure for measuring a sample of behavior. There are, at

least, three parts to this "systematic procedure". First, item content is chosen from the behavioral domain that the test is supposed to measure. Second, the procedures for administration are standardized such that each time the test is administered, directions for taking the test and recording the answers are identical, time limits (if applicable) are the same, and distractions are kept to a minimum. Finally, the scoring of the test is objective. These three requirements are stipulated to try to insure that the test is free from contaminants (Cascio, 1998; Gregory, 2000). Obviously, tests that are administered in an unproctored setting violate the second part of the definition of a test because there is no control in place for the distractions that could be present in an unproctored environment.

The issue of an unproctored environment is probably more critical for cognitive ability tests than for personality tests, and probably even more so for timed tests than for tests that are not timed. There are several factors in the test taker's environment that can cause contamination of test scores. These include temperature, humidity, illumination, and noise. Noise in particular is a factor that must be controlled in testing (Gregory, 2000). Noise has been shown to cause decreases in performance on tasks, especially when the noise is unpredictable, intermittent, and loud (Boggs & Simon, 1968). However, the effect of noise on performance on psychological tests is an area that has had very little research (Gregory, 2000). It seems likely that a person taking an unproctored web-based test could be exposed to noise in the environment, such as music, people talking, or a television being on in the background. However, it seems unlikely that there would be more noises that are unpredictable, intermittent, and loud in these scenarios. Therefore, the effects of an unproctored setting on web-based tests needs to be studied.

Potosky and Bobko (2004) conducted a study comparing paper-and-pencil versions of a timed cognitive ability test to a web-based unproctored version. The researchers used repeated measures under simulated high-stakes testing conditions to see if the tests were equivalent. Their results were mixed. First, the researchers correlated the scores on the two versions of the test. For the cognitive ability test, they found a moderate cross mode correlation. Additionally, Potosky and Bobko (2004) found significantly different means between the two administration modes. Because of this, the moderate correlation was probably a reflection of the sampling of similar behavioral domains and not necessarily evidence of measurement equivalence.

In a study to test the equivalence of proctored versus unproctored environments, Oswald, Carr, and Schmidt (2001) conducted research on web-based personality and cognitive ability tests versus their paper-and-pencil counter parts. Using a 2 x 2 between subjects factorial research design (proctored vs. unproctored and paper-and-pencil vs. web), the researchers used confirmatory factor analysis to compare the equivalence of the tests in the four conditions. They found that for the personality test, measurement equivalence between paper-and-pencil and web-based was demonstrated only for the proctored setting and not for the unproctored setting. For the ability measures, equivalence was demonstrated between the paper-and-pencil and web-based tests in both proctored and unproctored settings.

While to date, the research by Oswald, Carr, and Schmidt (2001) seems to be the most comprehensive of the studies demonstrating equivalence in cognitive ability tests, there are still some problems associated with the study. First, the researchers did nothing to try to reduce or to control practice effects. Second, the researchers made no attempt to

gather any information about the environment in which the participants took the test in the unproctored setting. Finally, there was no attempt to measure the usability of the web-based tests.

After reviewing the available research, it is apparent that more research needs to be conducted on the equivalence of web-based cognitive ability tests to paper-and-pencil versions. Very few studies have investigated this issue. Unfortunately, the few studies that have been conducted to date show inconsistent results.

Usability

Usability as a construct has several definitions. One definition offered by Nielsen (2003) is that usability is a quality attribute of interfaces that assesses how easy it is to use a particular user interface. Nielsen further divides usability into five components: learnability, efficiency, memorability, errors, and satisfaction. On the other hand, an alternate definition by Lundby & Mack (2003) defines usability as having three components: efficiency, effectiveness, and user-satisfaction.

Nielsen (2003) reported that usability could affect employee productivity since intranet systems (performance management systems, web-based employee surveys, etc.) with low usability can cause the employees to waste time trying to use the system. Additionally, Nielsen reported that the usability of a website has been known to affect people's perceptions of a website as well as to affect their use of that website. If a website is difficult to use, difficult to read, or wastes people's time, then they leave.

If these known effects of poor usability that are reported by Nelson (2003) are applied to web-based testing, the results could be the same. If a person has difficulty figuring out how to take the test, to answer the questions, to read figures, or to read the

text of the test, then they could leave before they finish the test. However, there is another possible test taking scenario that poor usability could cause. Poor usability could cause people to complete the form incorrectly, thereby introducing much error into the assessment process.

The importance of measuring usability of a computer-based or a web-based test in a proctored testing environment using a relatively uniform standard hardware and software configurations is apparent; however, measuring usability in an unproctored environment in which computers could have a much broader range of hardware and software settings is even more important. In one study that investigated this issue, Sinar and Reynolds (2004) found that people who took an unproctored test at home rated the test differently in terms of user friendliness than did people who took a proctored test or an unproctored test outside of their home.

Another study investigating the effects of usability in computer-based and Internet-based assessment (Bridgemen, Lennon, and Jackenthal, 2003) found that screen size and resolution impacted verbal scores on SAT questions. Specifically, it was found that participants who had more information on the screen tended to have higher test scores on the verbal portion of the test than did those who had less information on the screen. These results were not found in the mathematics portion of the test. It appeared that scrolling between the passage text and the test items caused the difference. In terms of web-based testing, this finding implies that the definition of a standardized proctored test needs to be expanded to include appropriate hardware and software settings.

However, it should be noted that the verbal test used was composed of reading comprehension items in which the participants read a passage and then answered

questions regarding the text. Therefore, these results may only apply to this particular type of question and not to all items that assess verbal ability. Also, the test used in that particular study was a power test with what the authors called a generous time limit. It is unknown at this point what effect a more stringent time limit might have had on the results.

Measuring usability. The Internet has posed an interesting problem in assessing usability. Traditionally, the interface that usability measures is the whole system. Lewis (1995) stated that proper usability evaluations of a computer system would include not just what a person sees on the screen, but also the keyboard, mouse, and other hardware that is being used. Usability evaluations of web-based systems seem to evaluate only the software interface. Limiting the scope of the interface to what is seen on the screen has both strengths and weaknesses.

The strength is that this procedure simplifies the usability evaluation since all combinations and permutations of hardware and software settings are not tested, thereby saving resources. One weakness of this approach is that it can introduce additional error into the measurement of usability. Since this possibility for additional error exists, then the recommendation by Lundby and Mack (2003) of taking different types of measurements, such as an efficiency measurement and a satisfaction measurement, when conducting a usability evaluation seems to be even more important. An additional problem with this limited definition of interface is that traditional user satisfaction questionnaires were designed to measure the usability of the complete system.

For web-based testing, the definitions offered by Nielsen (2003) and Lundby and Mack (2003) are inadequate. For example, how are errors measured on a test in which

the items are scored either right or wrong? If the test has detailed instructions, then what is measured by Nielsen's (2003) learnability (how easy it is for users to accomplish tasks the first time they visit a website)? Also, learnability is not very useful in testing since most tests seem to share a very similar layout. Effectiveness as part of the usability construct seems very important in purchasing a product over the Internet, but seems to have to do more with the overall quality of the test than with the usability of a web-based cognitive ability test. Finally, memorability is not a property that is usually associated with psychological tests. Memorability (Nielsen, 2003) is defined as a property of a web page that assesses how easy it is for a person to reestablish proficiency upon return to the website after a period of time has passed. Once again, if there are well-written instructions and a standard test format, then memorability is not a property that applies to a web-based test. If we adapt the learnability and memorability definitions to web-based tests, it appears that there is a new component of usability that would have to do with the instructions used with the web-based test.

Efficiency as defined by Nielsen (2003) is how quickly users can perform tasks. Lundby and Mack (2003) defined efficiency as the level of resources expended by the users in completing tasks. When measuring efficiency, both definitions suggest using measures such as time spent on task. Nielsen defined satisfaction as how pleasant the design is to use, whereas Lundby and Mack defined satisfaction as how satisfied users are with the overall experience. These two components of website usability seem to tie into web-based test usability. For example, efficiency could apply to what the test taker has to do to input an answer.

With the use of the three hypothesized usability components, it appeared possible to use them to generate new measures of usability as they relate to web-based tests. By integrating the various definitions, it appeared that a usability measure for a web-based test should be composed of a rating of the instructions that accompany the form, an assessment of ease of completing the test, and an overall estimate of satisfaction with the test. To meet these needs, a survey was developed (see Appendix 2) for use in measuring the usability of web-based tests. This survey appeared to be similar to the survey used by Sinar and Reynolds (2004) that defined user friendliness as system speed, efficiency, ease of navigation, and instructions for completion.

Besides assessing usability for informational and development purposes, there exists a theoretical reason to measure usability as well. If hardware and software settings affect test performance as they did in the Bridgemen, Lennon, and Jackenthal (2003) study, then it is possible that the adverse effects of the hardware and software settings could be detected by a usability measure. Therefore, when comparing a test that is administered in a proctored environment to the same test administered in an unproctored environment, then a comparison of the usability measures could reveal whether the hardware and software settings had an effect on the test.

After reviewing the research on this issue, it is apparent that very little is understood about the usability of web-based tests. In addition, very little information is available regarding how a person would measure usability for a web-based timed test.

Research Questions

After reviewing the available research on web-based cognitive ability testing, it is obvious that more research needs to be done in the area. To date, the few studies that

have tried to examine whether these types of tests are equivalent to their paper-and-pencil versions or whether they are equivalent in proctored or unproctored settings have found conflicting results. An additional problem exists since the majority of these studies used limited statistical comparisons by relying on either correlations or statistics based on linear equations. There has not been a study that used the Differential Functioning of Items and Tests (DFIT) framework, which uses a nonlinear analysis. Finally, it is also apparent that the effects of technology on web-based cognitive ability testing have been overlooked in all of the studies specifically investigating test equivalence.

This study sought to investigate these neglected areas. Specifically, in this study the following research questions were investigated:

1. To what extent are proctored paper-and-pencil and proctored web-based versions of a timed cognitive ability test equivalent?
2. To what extent are web-based proctored and web-based unproctored versions of a timed cognitive ability test equivalent?
3. To what extent did the relationship between a timed cognitive ability test and a criterion variable remained the same?
4. To what extent do the different hardware and software settings in an unproctored administration, as measured by a usability questionnaire, have an effect on scores?
5. Under what conditions are unproctored tests completed?

Methods

Participants

Participants in this study were 220 students in introductory psychology classes at a large southeastern public university who were randomly assigned into either the proctored web-based group (112 participants) or the unproctored web-based group (108 participants). In exchange for their participation, the subjects received one research credit that was used to partially satisfy their course requirements. Additionally, archival data from 650 adults, obtained from Wonderlic Inc, were analyzed as a part of this study. These 650 participants were part of one of two groups, 325 participants that completed the paper-and-pencil version of the *Wonderlic Personnel Test* (WPT) and 325 participants that completed the web-based WPT in a proctored setting.

Measures

The WPT is a timed test of cognitive ability for use in personnel selection. Participants in this study completed form I of the WPT. Each test taker had 12 minutes to complete the 50-item test (Wonderlic, 2006).

As part of this study, the undergraduate participants completed the web-based *Wonderlic Personnel Test-Quicktest* (WPT-Q). The WPT-Q is a 30-item timed test of cognitive ability for use in personnel selection that is available for administration over the Internet. Each test taker had 8 minutes to complete the WPT-Q (Wonderlic, 2005).

In addition to the WPT-Q, the undergraduate participants completed three questionnaires during this study. The first questionnaire, Environmental Questionnaire (see Appendix A), was designed to assess the conditions in which the test is completed. The second questionnaire, Usability Questionnaire (see Appendix B), was designed to

assess usability issues associated with the Internet administrations. These issues included connection speed, ease of use, and visual layout. The final questionnaire, Demographic Questionnaire (see Appendix C), was designed to gather data on various demographic variables (age, gender, ethnicity, primary language, college GPA, year in college, and SAT scores) that maybe associated with the participants' scores on the WPT-Q.

Procedures

Participants registered for an administration time at the campus Experimetrix website and were randomly assigned to one of two testing conditions; proctored Internet or unproctored Internet. Testing occurred in computer laboratories for the proctored Internet-based administration in a standardized setting. Administration of the instruments occurred in groups no larger then 30. The participants in the unproctored Internet-based administration were able to complete the test wherever they chose.

Proctored administration. Upon entering the computer laboratory, the undergraduate participants were instructed to sit in front of a computer terminal containing instructions that guided them through the rest of the study. The computers that were used in this study where either a Dell Precision 650 with an 18" LCD Flat Panel monitor (with resolution set at 1280x1024) or a Dell Dimension 4700C with a 15" LCD flat panel monitor (with resolution set at 1024x768). First, participants were required to read an Internet-based informed consent form. Then, they received instructions on taking the WPT-Q. Administration of the WPT-Q followed the standardized instructions exactly. After completing the WPT-Q, the participants were required to fill out the Environmental Questionnaire, Usability Questionnaire, and the Demographic Questionnaire.

Unproctored administration. Participants in this group first reported to a computer laboratory similar to that in the proctored administration. Once there, they were directed to sit at a computer and read over an electronic informed consent form, directions for participating in the study, and complete a web-based form on which they provided their contact information. After the participants left they room, they were then sent three e-mails. The first e-mail was a set of instructions explaining how to participate in the study. The second e-mail was an invitation to complete the WPT-Q. The third and final e-mail contained a link to the questionnaire. Administration of the WPT-Q followed the standardized instructions. After completing the WPT-Q, the participants completed the Usability Questionnaire, the Environment Questionnaire, and the Demographic Questionnaire.

Results

The results of this research were divided into two main sections. The first section was concerned with the analysis of the WPT (Research Question 1) while the second section was devoted to the WPT-Q (Research Questions 2 - 5). In each of these sections, the results of an exploratory factor analysis of the items for the test and a Differential Functioning of Items and Tests analysis are reported. These analyses are followed by additional analyses that are specific to each test.

Wonderlic Personnel Test (WPT)

Research Question 1

Since there was little or no variability in items 46-50, these items were dropped from the data sets and all subsequent analyses. Since this was a timed test, very few test takers were able to reach these five items on either version of the WPT. Of those that reached these five items, few were able to answer the questions correctly. In some cases, no one was able to answer the items correctly. In fact, only item 48 had correct responses on both the web-based version and paper-and-pencil version.

Exploratory Factor Analysis (EFA). As a first step in the analysis, an EFA was conducted on items 1-45 from the WPT paper and pencil version to check for unidimensionality using *Mplus* software version 3.13 (Muthen & Muthen, 2004). Since the data were dichotomous, tetrachoric correlations were used with promax rotation. Based on the analysis of the scree plot (see Figure 1), a three-factor solution seemed the best fit for the data (see Appendix D for factor loadings). The first factor was named verbal ability, the second factor was named logical reasoning, and the third factor was

named mathematical ability. Using this three-factor framework, the 45 items were split into three scales. Items 1, 4, 8, 13, 22, and 24 did not appear to load on any factor.

Differential Functioning of Items and Test (DFIT). For the DFIT analysis, the data were coded as correct, incorrect, or not reached for the participant's responses to each item per Ludlow and O'Leary's (1999) recommendation. The three factors revealed in the previous analysis, as well as items 1-45, were then subjected to a DFIT analysis using the 2PL model. The analyses were conducted using the *Bi-log-MG v3.0* (Zimowski, Muraki, Mislevy, & Bock, 2002), *Equate v2.1* (Baker, 1995), and *DFITD5* (Raju, 1999) programs. The DFIT analysis generated 88 differential functioning indexes (39 Non-Compensitory Differential Item Function (NCDIF) indexes plus 3 Differential Test Functioning (DTF) indexes, one for the each of the three scales, 45 NCDIF indexes and 1 DTF index for items 1-45). The NCDIF indexes, along with the χ^2 statistic, are shown in Tables 1-4 for verbal ability, logical reasoning, mathematical ability, and items 1-45, respectively. The results of the DTF analyses are shown in Table 5. In DFIT analyses, a difference in true scores for members of the various groups is considered significant when the associated χ^2 is significant and the NCDIF for an item exceeds an *a priori* specified critical value. As show in Tables 1-4, several items demonstrated Differential Item Functioning (DIF). While no items displayed DIF on the Verbal Ability, both Logical Reasoning and Mathematical Ability had items that displayed DIF. These items are 17, 34, 43 on Logical Reasoning and items 27, 33, 39, 42, and 44 on Mathematical Ability. However, none of the three scales displayed DTF. The results are similar for items 1-45 (Table 4).

Comparisons of the different versions. Table 6 contains information on the number of correct responses, number of participants who attempted each item, and the percent of participants who got each item correct in relation to the number who attempted the item. The most interesting information that this table reveals was that for about the first half of the test, the web-based version had more participants attempt items than did the paper-and-pencil version. This pattern changed for item 22 and the remaining items. The paper-and-pencil version had more participants attempt items than did the web-based version. However, for almost every item, the ratio of correct versus attempted was higher for the web-based group then it was for the paper-and-pencil group. It appears then that even though the web-based WPT had a higher correct to attempted ratio, it took longer for people to complete.

For technical reasons, several of the response options differ for the two versions of the WPT, and some of these differences warrant consideration. In general, participants inputted their answers through the Internet in one of three ways, either through a checkbox, a radio button, or a textbox. Of these three, checkboxes and radio buttons are very similar. The checkboxes allowed participants to make multiple selections from a list, whereas a radio button allowed only one selection to be made from a list. Textboxes allowed a participant to type in data using the keyboard. Table 7 contains information on the three response options and the items that use them. When this table was examined, two things became clear. First, the checkbox was only used once for items 1-45 (item 22). Second, more textboxes appear on the second half of the test than on the first half.

Wonderlic Personnel Test – Quicktest (WPT-Q)

Only 212 participants out of the original 220 where used in the analysis. Of the 8 participants whose data were not included, one individual had missing data and was therefore removed from the analysis. In addition, technical problems with the WPT-Q necessitated removing data from seven participants in the unproctored group. The technical problems that were reported came from a variety of sources. Three of these participants were disconnected from the Internet while completing the test, three participants had problems with their computers, and the final participant was unable to complete the test as a result of problems with Wonderlic's web server.

Research Question 2

Demographics. As a preliminary step, various demographic factors were analyzed to compare the groups' equivalence. All *t* tests were two-tailed. The two groups were equivalent on English as a first language $\chi^2(1, N = 205) = 0.0427$, $p > .05$ (Table 8), ethnicity $\chi^2(6, N = 204) = 2.3802$, $p > .05$ (see Table 9), age $t(188) = 0.83$, $p = .83$, $d = .12$ (see Table 10), credit hours completed $t(195) = -1.14$, $p = .2516$, $d = -.16$ (see Table 10), GPA $t(141) = 0.804$, $p = .2334$, $d = .14$ (see Table 10 for means), verbal SAT score $t(174) = .78$, $p = .4368$, $d = .12$ (see Table 10), and Quantitative SAT score $t(174) = -0.42$, $p = .6715$, $d = -.06$ (see Table 10). However, the two groups were different in terms of their gender makeup in that the proctored group contained significantly more males than females $\chi^2(1, N = 207) = 4.8791$, $p \leq .05$ (see table 10).

To further test the gender factor, a Fisher's *z* transformation was performed. The data were grouped by gender to compare Pearson correlations coefficients between total SAT scores and WPT-Q scores, ignoring the effects of proctoring. The correlation for

males was $r = .47$, $df = 109$, $p < .0001$ and for females was $r = .62$, $df = 63$, $p < .0001$. These correlations were then analyzed using the Fisher's z transformation. This analysis resulted in an insignificant difference between the two correlations ($z = 1.353$, $p = .1761$, two-tailed).

Exploratory Factor Analysis (EFA). As a first step in the analysis, an EFA was performed on the proctored WPT-Q to check for unidimensionality using *Mplus* software version 3.13 (Muthen & Muthen, 2004). Since the data were dichotomous, tetrachoric correlations were used with promax rotation. Based on the analysis of the scree plot (Figure 2), two-factor and three-factor solutions were considered. Using the combined principles of parsimony and interpretability, a 2-factor solution was retained (see Appendix E for factor loadings). Analysis of the items indicated that the first factor was Verbal Ability, and the second factor was Quantitative Ability.

Differential Functioning of Items and Test (DFIT). For the DFIT analysis, the data were coded as correct, incorrect, or not reached for the participants responses to each item per Ludlow and O'Leary's (1999) recommendation. The two factors revealed in the previous analysis, as well as items 3-29, were then subjected to a DFIT analysis using the 2PL model. The analyses were conducted using the *Bi-log-MG v3.0* (Zimowski, Muraki, Mislevy, & Bock, 2002), *Equate v2.1* (Baker, 1995), and *DFITD5* (Raju, 1999) programs. The DFIT analysis generated 56 differential functioning indexes (26 NCDIF indexes plus 2 DTF indexes, one for each of the two scales, 27 NCDIF indexes and 1 DTF index for items 3-29). The NCDIF indexes, along with the χ^2 statistic, are shown in Tables 14-16 for verbal ability, quantitative ability, and items 3-29, respectively, and the results of the DTF analyses are reported in Table 17. In DFIT analyses, a difference in

true scores for members of the various groups is considered significant when the associated χ^2 is significant and the NCDIF for an item exceeds an *a priori* specified critical value. As shown in the tables 12-13, several items demonstrated DIF. These items were item 8 on Verbal ability and items 22, 28, and 29 on Quantitative Ability. However, neither factor displayed DTF. The results were similar for items 3-29 (see Table 16).

Research Question 3

Comparison of correlation coefficients. The correlation coefficients for the proctored and unproctored group were compared as an additional check for measurement equivalence. As a first step, the Pearson correlation coefficients were computed between the WPT-Q scores and the combined Verbal and Quantitative SAT scores for the proctored ($r = .41$, $df = 90$, $p < .0001$) and unproctored group ($r = .63$, $df = 82$, $p < .0001$). These correlations were then analyzed using Fisher's z transformation. This analysis resulted in a significant difference, $z = 2.551$, $p = .0107$.

Research Question 4

Environmental Questionnaire. While the proctored group participated in the research in one of several computer laboratories, the unproctored group completed the study in a variety of locations. The locations are summarized in Table 12. The environment in which the WPT-Q was completed in for the proctored and unproctored groups were compared using the results of the environmental questionnaire. The results are shown in Table 13.

Research Question 5

Usability Questionnaire. To compare the usability of the proctored and unproctored environment, responses to the usability questionnaire were scored 1 for "Unsatisfactory" to 5 for "Excellent". An EFA was conducted on the data using SAS software version 9.1 to verify that the questionnaire was unidimensional. The Kaiser Criterion and Scree Plot analysis both indicated a single factor solution. A Cronbach's Alpha reliability analysis yielded an $\alpha = .88$. Therefore, it was concluded that the usability questionnaire measured a strong single factor.

Since the previous analysis concluded a single factor for the usability questionnaire, each participant's responses were summed. The results were then analyzed using an independent-samples *t* test, two-tailed. The sample means for the proctored and unproctored groups were, respectfully, 30.76 and 29.74 and were not significantly different, $t(206) = 1.90$, $p = .0584$, $d = .26$.

Discussion

Following the format of the results section, this section has been divided into two parts. The results of each test are first discussed independently. Then the overall conclusions from this study are discussed.

Wonderlic Personnel Test (WPT)

Data from 650 participants who completed either the paper-and-pencil version of the WPT or a web-based version were analyzed to investigate research question 1 of whether or not paper-and-pencil cognitive ability tests are equivalent to web-based cognitive ability tests. The analysis of the WPT revealed several items that displayed differential item functioning (DIF) between the paper-and-pencil version and the web-based version. Although only a minority of the items demonstrated differential functioning, it can be reasonably concluded that the two versions of the test are not completely equivalent.

Since only test scores were available for the analysis, causes for this differential functioning are only speculative in nature, and the impact of the differential functioning on test interpretation is unknown. However, in spite of the limited evidence, there existed several possible explanations regarding the reasons why these two versions of the test showed differential functioning. More insight regarding why the test behaved differently in the two groups was gained by looking at the specific items.

From the analysis of the items used in the test it was indicated that participants who completed the web-based WPT took longer to complete the test than did those who completed the paper-and-pencil WPT and more textbox questions were used in the second half of the test. Taken separately, these two pieces of information are interesting.

However, if these findings are considered together with the fact that the WPT is a timed test, then a possible psychological explanation surfaces regarding why the 8 items demonstrated differential functioning -- namely self-efficacy.

Computer self-efficacy is a specific form of self-efficacy and is defined as a judgment of one's capability to use a computer (Compeau & Higgens 1995). Potosky and Bobko (2004) hypothesized that individuals with high Internet self-efficacy would find a web-based test easier and less stressful than would those with low Internet self-efficacy, thereby allowing them to outperform individuals with low Internet self-efficacy.

It is possible that participants with low computer self-efficacy took longer to complete the test than did participants with high computer self-efficacy. The textbox questions could have been particularly problematic for the low self-efficacy participants. If this phenomenon exists, then low computer self-efficacy could have a larger impact on timed web-based tests than on tests that are not timed.

Wonderlic Personnel Test – Quicktest (WPT-Q)

Data from 212 participants were analyzed to investigate research question 2 -- namely whether or not a timed cognitive ability test functions the same in a proctored web-based environment as it does in an unproctored web-based environment. Results of the DFIT analyses indicated that several items on the WPT-Q did function differently for the two conditions. The source of this differential function could have resulted from a number of factors. The relatively small sample size used in this analysis may have contributed to this result. However, the results of the Fisher's *z* test, and the answer to the research question 3, indicated that the correlations between the WPT-Q scores and the

total SAT scores were significantly different for the proctored and unproctored groups, thus strengthening the research findings for the second research question.

Demographic variables were ruled out as a factor in the differential functioning. All but one of the analyses conducted on the demographic variables in this study showed no significant difference between the proctored and the unproctored groups. The only variable that did have a significant difference between the groups was gender. However, since the two groups were equivalent in terms of SAT scores and GPA and the results of the Fisher's *z*, it is believed that the gender factor can probably be ruled out as the cause of DIF.

These results were consistent with what appears to be the state of the field in gender differences in cognitive ability. In this research area, the existence of differences in cognitive ability between males and females seems to depend on the specific ability being measured and on the sample on which the study was conducted (Feingold, 1996).

Research question 4 sought to see if the technology used was a source of the differential functioning. In this study, the participants in the unproctored group had a lower usability score, however the difference was not significant at the .05 level (but would have been at the $p < .1$ level). This reported lower average usability score is another possible reason for the differential functioning between the proctored and unproctored test participants. These results corresponded to the results of Bridgemen, Lennon, and Jackenthal (2003). These results agreed with their findings that screen settings and resolution matter; however, the results go beyond their findings since the current study seems to indicate that usability can affect non-reading comprehension items as well.

The most likely source of the differential functioning, and the answer to research question 5 of this study, was the environment in which the test was completed. Although the proctored environment was not always ideal, the unproctored environment suffered a much larger departure from ideal conditions. This finding was consistent with the research conducted by Sinar and Reynolds (2004). However, what still remains unclear is what in the unproctored environment might have caused the DIF. Additionally, these results do not directly support the results of Boggs and Simon (1968) since it appeared that performance did not suffer.

Unfortunately, the effects of computer or Internet self-efficacy cannot be ruled out as the source of the differential functioning. However, it was believed that since these specific self-efficacies are attributes of the person, any impact that they might have had was controlled by the random assignment to the different conditions. Since random assignment to the groups was used, it was assumed that varying levels of these self-efficacies were equivalent for the two groups.

A final untested possible explanation to the differential functioning is that participants in the unproctored group were subjected to a certain effects of group testing. It has long been known that tests that are administered in a group setting have certain disadvantages or risks over tests individually administered. One disadvantage is that a person's actual score will be different from their true score because of motivational problems or difficulty following the directions. Additionally, these scores will not be recognized as invalid (Gregory, 2000). It is likely that unproctored tests, even if administered individually, share these problems with group-administered tests. It is very possible that participants used resources that were expressly forbidden in the instructions

because they were not aware of the instructions. Even though it is likely that such was the case, it is unfortunately impossible to determine how extensive this problem was or what affect it had on participant's scores.

Overall Conclusions

In the present research, provocative results were found. The most important findings related to the equivalence of proctored and unproctored web-based cognitive ability tests and the equivalence of web-based tests and paper-and-pencil tests. These results directly contradicted some of the research that has been conducted in this area (Sinar and Reynolds, 2004; Oswald, Carr, & Schmidt, 2001); however, the findings supported other research that has been conducted (Potosky and Bobko, 2004; Bridgemen, Lennon, and Jackenthal, 2003). Several possible explanations exist as to why these differences were found. It could imply that measurement equivalence between modes of administrations could rest on the test specific test or the specific abilities that are being measured. However, it is also possible that the type of statistical analysis made a difference.

In general, the results showed that measurement equivalence was less than perfect. The amount of measurement variance that was detected was small and limited to only a few items. The fact that in all of the DFIT analyses that were conducted as part of this study, no instance of DTF was discovered demonstrates that timed cognitive ability measures still measure cognitive ability regardless of the mode of administration. Therefore, the construct validity of web-based cognitive ability tests in general, and the Wonderlic tests in specific, remain in tact.

However, even though the presence of DIF was limited, there was enough DIF to change the relationships between predictors and criterions. This result was most apparent in a significantly different Fisher's *z* test between the WPT-Q scores and the combined SAT scores. The practical implications of this finding are most apparent for an organization that wishes to change the mode of the their selection tests, such as from paper-and-pencil to web-based proctored test or web-based proctored to web-based unproctored. These organizations will need to conduct new criterion-related validity studies because the criterion-related validity of the measures cannot be assumed to remain the same when mode of test administration is changed.

Although unrelated to the question of equivalence of the administration modes, this research produced provocative results in another area, i.e., the factor structure of the WPT and WPT-Q. Both of these tests are supposed to measure *g* (Wonderlic, 2005; Wonderlic, 2006). If these tests did measure *g*, then the factors should have been highly correlated. However, in this study, the factors were moderately correlated at best and uncorrelated at worst. Therefore, the results of this study suggest that the WPT and WPT-Q should be better thought of as tests that measure several specific abilities and not necessarily as a test that measures *g*.

Limitations

The major limitation of this study was the small sample size used in the analysis of the WPT-Q data. In future research, larger sample sizes should be used, not only for the analysis of proctored versus unproctored tests, but also with the analysis of paper-and-pencil versus web-based tests. An additional limitation of the present research was the lack of measures for self-efficacy, cheating, or group testing effects. Inclusion of

such measures might give more insight into the possible causes of the differential functioning in research comparing the equivalence of web-based tests.

Another potential limitation in this research was the use of data from low-stakes testing situations. The participants in this research did not have any external motivation, such as getting a job, to perform at their best. Therefore, the results found in this study might not replicate results from a high stakes testing situation.

Directions for Future Research

Research on these same questions needs to be continued. One issue that does not seem to be resolved yet is whether the results of research investigating web-based test equivalence are limited to specific tests, or can these results be generalized to the method of administration of any test? Until more research is conducted, this question cannot be answered.

Also, research incorporating both usability measures and self-efficacy measures should become standard procedures when conducting this type of research. Very little seems to be known about how either of these constructs is related to web-based tests. The current research made significant contributions since not only was a method for measuring usability in timed web-based tests used successfully, but also a new measure of usability for web-based tests was introduced.

Research should also be conducted on how different response options, e.g., checkboxes, textboxes, and radio buttons, affect the outcome of web-based tests. It is possible that usability and self-efficacy measures could assist in understanding what factors can have an impact on web-based tests.

In conclusion, provocative results were found during this research. Some of these results supported past research, and some of them did not. In all, results suggested that mode of administration matters. Only additional research will help to settle the question of the equivalence of web-based cognitive ability testing. Future research should continue to focus not only on what affects the equivalence of web-based tests, but also on why researchers have found conflicting results.

References

Baker, F. B. (1995). Equate 2.1: Computer program for equating two metrics in item response theory [Computer Program]. Madison: University of Wisconsin, Laboratory of Experimental Design.

Beaty, J. C., Fallon, J. D., & Barrett, C., 2002. Proctored versus unproctored web-based administration of a cognitive ability test. Paper presented at the 17th Annual Conference for the Society for Industrial and Organizational Psychology (SIOP), Toronto, Ontario.

Boggs, D. H. & Simon, J. R. (1968). Differential effect of noise on tasks of varying complexity. *Journal of Applied Psychology, 52*, 148 – 153.

Bridgeman, B., Lennon, M.L., & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education, 16(3)*, 191-205.

Buchanan, T. (2000). Potential of the Internet for personality research. In M. H. Birnbaum (Ed.) *Psychological experiments on the Internet* (121 – 140). San Diego, CA: Academic Press.

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

Cascio, W. F. (1998). *Applied psychology in human resource management (5th ed.)*. Upper Saddle River, NJ: Prentice Hall.

Compeau, D.R. & Higgins, C.A. (1995). Computer self-efficacy: development of a measure and initial test. *MIS Quarterly*, 19, 189-211.

Dembowski, J.M & Callans, M.C. (2000). Comparing computer and paper forms of the Wonderlic Personnel Test. Paper presented at the Society for Industrial and Organizational Psychology, New Orleans, LA.

Epstein, J., Klinkenberg, W.D., Wiley, D., and McKinley, L. (2001). Insuring sample equivalence across Internet and paper-and-pencil assessments. *Computers in Human Behavior, 17*, 339-346.

Feingold, A. (1996). Cognitive gender differences: where are they and why are they there? *Learning and Individual Differences, 8*(1), 25 – 32.

Gregory, R. J. (2000). *Psychological testing* (3rd ed.). Needham Heights, MA: Allyn and Bacon.

Lewis, J. R. (1995). IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use. *International Journal of Human-Computer Interaction, 7(1)*, 57 – 78.

Lievens, F. & Harris, M. M. (2003). Research on Internet recruiting and testing: current status and future directions. In C. L. Cooper and I. T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology* (131 - 165). Chichester, UK: Wiley.

Ludlow, L.H. & O'Leary, M. (1999). Scoring omitted and not-reached items: practice data analysis implications. *Educational and Psychological Measurement, 59*(4), 615-630.

Lundby, K & Mack, M. (2003). Usability research: An introduction, general overview, and practical applications for I/O psychology. Paper presented at Society for Industrial/Organizational Psychology, Orlando, FL.

McBride, J. R. (1998). Innovations in computer-based ability testing: promise problems and perils. In M.D. Hakel (Ed.) *Beyond multiple choice: Evaluating alternatives to traditional testing for selection* (23 – 40). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Mead, A.D. & Drasgrow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114(3),* 449-458

Muthen, L. & Muthen, B. (2004). Mplus 3.13 [Computer Program]. Los Angeles, Muthen & Muthen.

Naglieri, J.A., Drasgrow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., & Velasquez, R. (2004). Psychological testing on the Internet, new problems, old issues. *American Psychologist, 59*(3), 150-162.

Nielsen, J (2003). Usability 101: Introduction to usability. Retrieved August 1, 2004 from http://www.useit.com/alertbox/20030825.html.

Oswald, F. L. Carr, J.Z., & Schmidt, A.M. (2001). The medium and the message: Dual effects of supervision and web-based testing on measurement equivalence for ability and personality measures. Paper presented at the Society for Industrial and Organizational Psychology, San Diego, CA.

Polyhart, R.E., Weekly, 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.

Potosky, D. & Bobko, P. (2004). Selection testing via the Internet: practical considerations and exploratory empirical findings. *Personnel Psychology*, 57, 1003 –1034.

Raju, N.S. (1999). DFITD5: A Fortran program for calculating DIF/DTF [Computer Program]. Chicago, Illinois Institute of Technology.

Sinar, E. & Reynolds, D. (2004). Exploring the impact of unstandardized Internet testing. Paper presented at the Society for Industrial Organizational Psychology, Chicago, IL.

Wonderlic (2005). Wonderlic Personnel Test-Quicktest. Retrieved January 25, 2005 from http://www.wonderlic.com/products/product.asp?prod_id=35

Wonderlic (2006). Wonderlic Personnel Test. Retrieved January 31, 2006 from http://www.wonderlic.com/products/product.asp?prod_id=4

Zimowski, M., Muraki, E., Mislevy, R. & Bock, D. (2002). BILOG-MG 3 [Computer Program]. Chicago, Scientific Software, Inc.

Table 1 – DIF Statistics for Factor 1 of the WPT - Verbal Ability. NC-DIF Index and χ^2 Statistic for each item in the scale. NC-DIF Index is marked with an * if greater then .006. NC-DIF = non-compensatory differential item functioning.

Item	NC-DIF	χ^2	Prob
2	.004	2363.87	.0000
3	.000	895.26	.0000
5	.005	325.70	.4631
6	.000	387.95	.0084
7	.001	510.53	.0000
9	.002	1474.54	.0000
14	.001	377.41	.0218
15	.002	348.32	.1689
16	.002	555.92	.0000
20	.001	792.39	.0000
21	.001	1031.33	.0000
25	.000	1212.08	.0000
28	.002	410.16	.0008

Table 2 - DIF Statistics for Factor 2 of the WPT - Logical Reasoning. NC-DIF Index and χ^2 Statistic for each item in the scale. NC-DIF Index is marked with an * if greater then .006. NC-DIF = non-compensatory differential item functioning.

Item	NC-DIF	χ^2	Prob
11	.003	345.17	.2004
17	.012*	3095.90	.0000
19	.001	70216.83	.0000
32	.001	35984.71	.0000
34	.008*	16059.35	.0000
35	.000	2255.30	.0000
36	.001	552.25	.0000
38	.002	1651.11	.0000
40	.002	340.96	.2480
41	.002	379.53	.0181
43	.019*	114404.10	.0000
45	.004	740.82	.0000

Table 3 - DIF Statistics for Factor3 of the WPT - Mathematical Ability. NC-DIF Index and χ^2 Statistic for each item in the scale. NC-DIF Index is marked with an * if greater then .006. NC-DIF = non-compensatory differential item functioning.

Item	NC-DIF	χ^2	Prob
10	.002	37433.93	.0000
12	.003	1592.48	.0000
18	.002	413.66	.0005
23	.003	6941.10	.0000
26	.003	331.71	.3720
27	.009*	9264.13	.0000
29	.001	443.12	.0000
30	.002	494.12	.0000
31	.011*	341.51	.2413
33	.011*	2625.79	.0000
37	.003	586.72	.0000
39	.008*	1619.99	.0000
42	.012*	501.57	.0000
44	.009*	572.45	.0000

Table 4 - DIF Statistics for Items 1-45 of the WPT. NC-DIF Index and χ^2 Statistic for each item in the scale. NC-DIF Index is marked with an * if greater then .006. NC-DIF = non-compensatory differential item functioning.

Item	**NC-DIF**	χ^2	**Prob**	Item	NC-DIF	χ^2	**Prob**
1	.001	682.84	.0000	**24**	.001	2844.08	.0000
2	.006	1560.34	.0000	**25**	.000	361.99	.0716
3	.000	414.86	.0005	**26**	.000	330.43	.3909
4	.000	646.34	.0000	**27**	.005	7018.51	.0000
5	.004	336.31	.3072	**28**	.003	369.67	.0408
6	.000	640.06	.0000	**29**	.001	367.10	.0496
7	.001	498.32	.0000	**30**	.005	423.36	.0002
8	.000	392.27	.0056	**31**	.008*	343.32	.2205
9	.000	3164.28	.0000	**32**	.005	1415.62	.0000
10	.004	717.18	.0000	**33**	.006	2990.67	.0000
11	.001	2470.80	.0000	**34**	.008*	14578.33	.0000
12	.007*	36078.92	.0000	**35**	.000	2476.19	.0000
13	.001	451.99	.0000	**36**	.002	401.37	.0022
14	.001	334.78	.3282	**37**	.003	360.91	.0772
15	.000	374.27	.0283	**38**	.001	120035.90	.0000
16	.001	477.49	.0000	**39**	.007*	1598.60	.0000
17	.003	23810.98	.0000	**40**	.000	411.28	.0007
18	.001	388.22	.0082	**41**	.001	326.14	.4562
19	.000	559.18	.0000	**42**	.006	718.85	.0000
20	.001	1570.79	.0000	**43**	.026*	5133.34	.0000
21	.001	527.79	.0000	**44**	.009*	533.30	.0000
22	.003	874.61	.0000	**45**	.004	4768.75	.0000
23	.005	17272.32	.0000				

Table 5 - Differential Test Functioning (DTF) Results for the WPT.

	DTF	**DTF Cutoff**	χ^2	**Prob**
Factor 1	.00016	.078	489.65	.0000
Factor 2	.00073	.072	709.47	.0000
Factor 3	.00116	.084	558.36	.0000
Items 1-45	.01907	.270	716.17	.0000

Table 6 – Correct and Attempted information for items 1-45 on the Wonderlic Personnel Test. P = Paper and Pencil WPT, W = Web-based WPT

	Correct		Attempted		Percent Correct	
Item	**P**	**W**	**P**	**W**	**P**	**W**
1	320	322	325	325	98.46	99.08
2	274	285	325	325	84.31	87.69
3	298	311	325	325	91.69	95.69
4	265	280	325	325	81.54	86.15
5	299	300	325	325	92.00	92.31
6	257	287	325	325	79.08	88.31
7	295	306	325	325	90.77	94.15
8	189	210	325	325	58.15	64.62
9	282	298	325	325	86.77	91.69
10	113	175	325	325	34.77	53.85
11	241	270	325	325	74.15	83.08
12	169	239	325	325	52.00	73.54
13	250	276	325	325	76.92	84.92
14	210	244	325	325	64.62	75.08
15	242	272	325	325	74.46	83.69
16	215	243	323	325	66.56	74.77
17	213	258	323	325	65.94	79.38
18	131	156	318	320	41.19	48.75
19	244	264	315	320	77.46	82.50
20	230	233	313	316	73.48	73.73
21	192	224	310	312	61.94	71.79
22	106	151	304	300	34.87	50.33
23	208	248	296	289	70.27	85.81
24	125	119	293	279	42.66	42.65
25	234	237	285	271	82.11	87.45
26	134	153	276	248	48.55	61.69
27	117	114	253	229	46.25	49.78
28	197	184	250	218	78.80	84.40
29	54	59	218	179	24.77	32.96
30	73	91	207	173	35.27	52.60
31	29	37	185	141	15.68	26.24
32	109	84	180	131	60.56	64.12
33	28	24	162	103	17.28	23.30
34	53	28	147	99	36.05	28.28
35	64	52	136	94	47.06	55.32
36	49	39	123	86	39.84	45.35
37	6	5	103	66	5.83	7.58
38	32	21	100	62	32.00	33.87
39	5	10	91	52	5.49	19.23
40	22	14	89	47	24.72	29.79
41	38	23	83	44	45.78	52.27
42	5	7	71	34	7.04	20.59
43	13	15	65	32	20.00	46.88
44	5	4	50	25	10.00	16.00
45	8	5	46	24	17.39	20.83

Table 7 – Response Types for the Web-Based WPT items 1-45.

Response Type	Items in which it occurs
Text Box (Open response)	10, 12, 15, 16, 18, 23, 26, 27, 29, 31, 33, 37, 39, 40, 42, 44, 45
Radio Button (Fixed number of response options, only one can be selected)	1-9, 11, 13, 14, 17, 19, 20, 21, 24, 25, 28, 30, 32, 34 –36, 38, 41, 43
Check Box (Fixed number of response options, multiple options can be selected)	22

Table 8 – English as Primary Language.

	Proctored	Unproctored
English as a first language	97	88
English as a foreign language	10	10

Table 9 – Ethnicity.

Ethnicity	Proctored	Unproctored
African American	15	12
Asian	9	10
Caucasian	80	69
East Indian	0	0
Hispanic/Latino	1	2
Middle Eastern	1	2
Native American	0	0
Pacific Islander	0	1
Other	1	1

Table 10 - Means for the Proctored and Unproctored Group.

	Proctored		Unproctored	
	Mean	SD	Mean	SD
Age	20.4340	3.2397	20.1111	2.2673
Credit Hours Completed	30.0481	31.9339	35.5161	35.4455
GPA	2.9048	0.6039	3.0314	0.6604
Verbal SAT	586.2934	113.1933	573.1071	110.9092
Quant SAT	612.0435	100.9720	618.6786	106.2104
WPT-Q Scores	21.2500	3.9105	20.9200	3.6004
Usability	30.7593	3.6091	29.7400	4.1111

Table 11 – Gender.

Gender	Proctored	Unproctored
Male	76	55
Female	32	44

Table 12 – Location of the Unproctored WPT-Q Administration.

Location	Frequency	Percent
Dorm Room	37	37
Computer Lab	13	13
Library	0	0
Apartment	20	20
House	25	25
Store	1	1
Office	1	1
Other	3	3

Table 13 – Possible Distracters at the Testing Site for both the Proctored and Unproctored Administrations.

Events that occurred during testing	Proctored Environment			Unproctored Environment		
	Not Present	Present but not distracting	Present and Distracting	Not Present	Present but not distracting	Present and Distracting
Pop-up advertisements	108	0	0	93	5	2
TV on	107	0	0	75	21	4
Music Playing	108	0	0	78	19	3
Someone Interrupted	105	1	1	66	21	12
Telephone Rang	100	5	2	89	7	4
Notified of receipt of e-mail	107	1	0	97	2	1
Received Instant Message	108	0	0	85	10	5
Someone Talking	89	16	2	48	39	12
Other	97	10	1	91	7	2

Table 14 – DIF Statistics for Factor 1 of the WPT-Q - Verbal Ability. NC-DIF Index and χ^2 Statistic for each item in the scale. NC-DIF Index is marked with an * if greater then .006. NC-DIF = non-compensatory differential item functioning.

Item	NC-DIF	χ^2	Prob
1	.003	228.63	.0000
4	.001	1010.31	.0000
5	.001	128.48	.0452
6	.004	105.89	.4030
8	.018*	781.70	.0000
9	.000	189.16	.0000
11	.002	1214.47	.0000
13	.000	600.31	.0000
14	.005	536.36	.0000
15	.002	168.51	.0000
16	.003	394.82	.0000
18	.000	110.17	.2964

Table 15 - DIF Statistics for Factor 2 of the WPT-Q - Quantitative Ability. NC-DIF Index and χ^2 Statistic for each item in the scale. NC-DIF Index is marked with an * if greater then .006. NC-DIF = non-compensatory differential item functioning.

Item	NC-DIF	χ^2	Prob
10	.002	58159.25	.0000
12	.001	163.03	.0002
17	.002	250.17	.0000
19	.002	182.81	.0000
20	.004	723.64	.0000
21	.003	202.38	.0000
22	.013*	4589.44	.0000
23	.000	104.21	.4482
24	.001	3124.81	.0000
25	.002	104.77	.4330
26	.002	1828.81	.0000
27	.002	2605.88	.0000
28	.010*	6512.99	.0000
29	.007*	4496.30	.0000

Table 16 - DIF Statistics for Items 3-29 of the WPT-Q. NC-DIF Index and χ^2 Statistic for each item in the scale. NC-DIF Index is marked with an * if greater then .006. NC-DIF = non-compensatory differential item functioning.

Item	NC-DIF	χ^2	Prob
3	.000	268.65	.0000
4	.001	444.51	.0000
5	.001	104.03	.4530
6	.004	127.56	.0508
7	.001	2343.12	.0000
8	.003	309.61	.0000
9	.000	1148.88	.0000
10	.002	22434.01	.0000
11	.000	432.43	.0000
12	.000	387.03	.0000
13	.005	261.80	.0000
14	.005	263.77	.0000
15	.010*	878.57	.0000
16	.005	733.19	.0000
17	.004	158.25	.0004
18	.002	602.77	.0000
19	.002	262.26	.0000
20	.005	827.76	.0000
21	.005	182.13	.0000
22	.015*	1469.63	.0000
23	.000	112.01	.2558
24	.004	236.97	.0000
25	.004	108.17	.3444
26	.002	1350.50	.0000
27	.002	590.18	.0000
28	.010*	26830.11	.0000
29	.008*	1891.84	.0000

Table 17 - WPT-Q Differential Test Functioning (DTF) Results.

	DTF	DTF Cutoff	χ^2	Prob
Verbal Ability	.00999	.072	289.68	.0000
Quantitative Ability	.00272	.084	116.96	.1640
3-29	.01655	.162	110.84	.2813

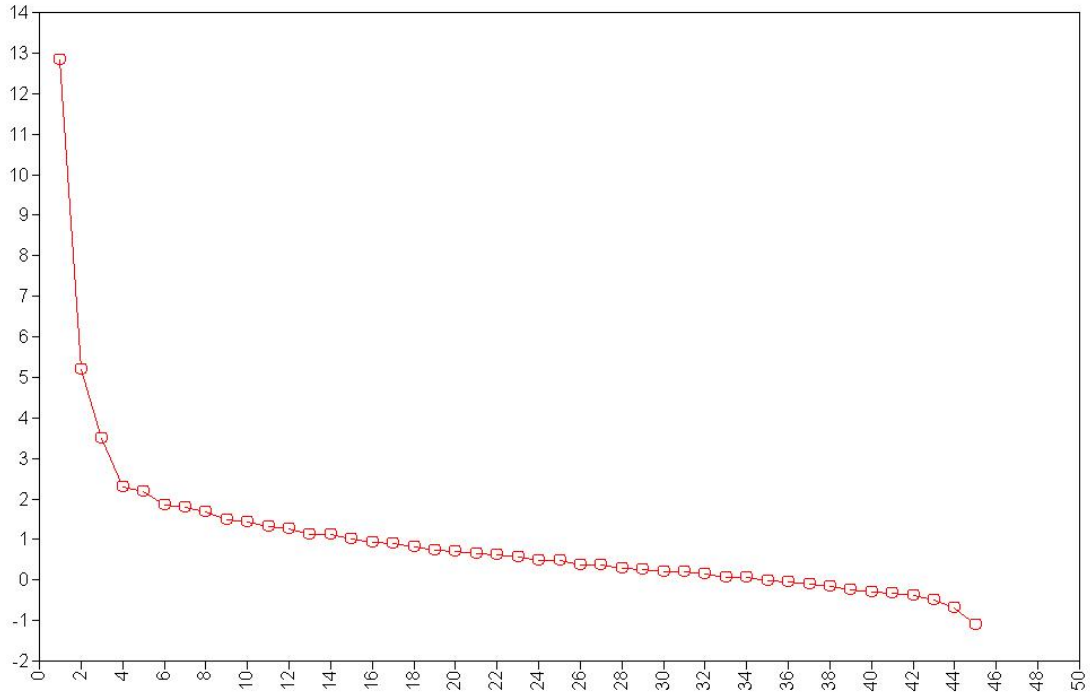

Figure 1. Scree plot for EFA analysis on items 1-45 from the WPT

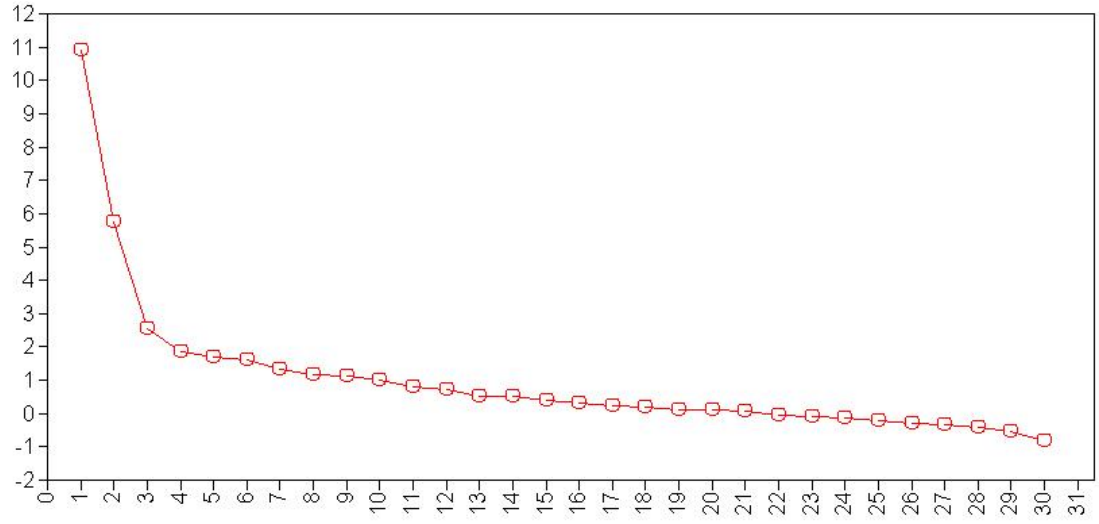

Figure 2. Scree plot for EFA analysis on items 1-30 from the WPT-Q

Appendices

Appendix A

Environment Questionnaire

Now that you have completed the cognitive ability test, I would like to get a better understanding of the location where you are in this study. Please be as specific as possible in answering these questions.

1. Where are you participating in this study (ex dorm room, coffee shop, computer lab, etc.)?
   - Dorm Room
   - Computer Lab
   - Library
   - Apartment
   - House
   - Store (such as a coffee house or book store)
   - Office
   - Other

2. While you were completing the test, did you experience any pop up advertisements?
   - No
   - Yes, but it wasn't distracting
   - Yes, and it was distracting

3. While you were completing the test, was there a TV on?
   - No
   - Yes, but it wasn't distracting
   - Yes, and it was distracting

4. While you were completing the test, was there music playing?
   - No
   - Yes, but it wasn't distracting
   - Yes, and it was distracting

5. While you were completing the test, did someone interrupt you?
   - No
   - Yes, but it wasn't distracting
   - Yes, and it was distracting

6. While you were completing the test, did the telephone ring?
   - No
   - Yes, but it wasn't distracting
   - Yes, and it was distracting

7. While you were completing the test, were you notified of receiving an e-mail?
   No
   Yes, but it wasn't distracting
   Yes, and it was distracting

8. While you were completing the test, did you receive an instant message?
   No
   Yes, but it wasn't distracting
   Yes, and it was distracting

9. While you were completing the test, did you hear anyone talking?
   No
   Yes, but it wasn't distracting
   Yes, and it was distracting

10. While you were completing the test, did anything else happen?
   No
   Yes, but it wasn't distracting
   Yes, and it was distracting

If yes, what happened?

12. Did you experience any technical problems while completing this test? If so, what were they?

13. Did you experience any other problems while completing the test? If so, what were they?

Appendix B

Usability Questionnaire

I would now like you to answer a few questions about the cognitive-ability test that you completed. Please be as specific as possible in answering these questions.

Please rate your satisfaction with the Wonderlic Personnel Test on the following dimensions:

Unsatisfactory Poor Adequate Good Excellent

14. Readability of text
15. Quality of graphics
16. Instructions
17. Test appearance
18. Ease of marking answers
19. Speed of site
20. Overall ease of taking the test

21. What type of Internet connection are you using to complete this test?
    - Dial-up
    - Cable Modem or DSL
    - Network Connection
    - Wireless Network
    - Other

Appendix C

Demographic Questionnaire

I would now like you to answer the following questions about yourself. Please be as specific as possible in answering these questions.

22. What is your Date of Birth (month/day/year):

23. How many hours of course credit have you completed?

24. What is your current GPA (If this is your first semester in college, please mark N/A)?

25. What year did you take the SAT?

26. What was your Verbal score on the SAT?

27. What was your Quantitative (Math) score on the SAT?

28. Is English the first language you learned?
    Yes
    No

29. What is your gender?
    Male
    Female

30. What is your ethnicity?
    African American (Black)
    Asian
    Caucasian (White)
    East Indian
    Hispanic/Latino
    Middle Eastern
    Native American
    Pacific Islander
    Other:

Appendix D

Table D1 - Promax Rotated Loadings for the WPT

	1	2	3
Y1	0.216	0.353	-0.122
Y2	0.767	0.131	0.104
Y3	0.528	0.040	-0.051
Y4	0.252	0.003	0.116
Y5	0.646	0.165	0.039
Y6	0.653	0.010	0.037
Y7	0.586	0.177	0.115
Y8	0.061	0.060	0.315
Y9	0.785	0.074	-0.007
Y10	0.357	0.084	0.482
Y11	0.114	-0.149	0.208
Y12	0.142	0.099	0.514
Y13	0.313	0.127	0.164
Y14	0.387	0.000	0.299
Y15	0.683	0.007	0.151
Y16	0.414	0.020	0.066
Y17	0.359	-0.013	0.166
Y18	0.082	-0.070	0.601
Y19	0.386	-0.018	0.080
Y20	0.140	-0.068	0.145
Y21	0.291	0.059	0.275
Y22	0.338	-0.157	0.280
Y23	0.358	0.042	0.428
Y24	0.032	-0.074	0.273
Y25	0.532	-0.042	0.249
Y26	0.369	0.075	0.614
Y27	0.234	-0.193	0.604
Y28	0.632	-0.150	-0.060
Y29	0.300	-0.056	0.405
Y30	0.357	-0.334	0.402
Y31	0.235	-0.332	0.802
Y32	0.347	-0.437	0.235
Y33	0.172	-0.380	0.619
Y34	-0.111	-0.465	0.133
Y35	0.221	-0.713	0.092
Y36	0.034	-0.795	0.002
Y37	-0.187	-0.835	0.684
Y38	0.131	-0.568	-0.030
Y39	-0.225	-0.871	0.531
Y40	0.301	-0.602	-0.309
Y41	0.154	-0.837	-0.021
Y42	-0.168	-0.871	0.518
Y43	0.376	-0.509	-0.548
Y44	-0.287	-0.882	0.603
Y45	0.341	-0.505	-0.400

Table D2 - Promax Factor Correlations for the WPT

	1	2	3
1	1.000		
2	-0.312	1.000	
3	0.279	0.014	1.000

Appendix E

Table E1 - Promax Rotated Loadings for the WPT-Q

	1	2
Y1	0.757	0.081
Y2	0.687	0.215
Y3	0.160	0.117
Y4	0.923	0.118
Y5	0.969	-0.122
Y6	0.635	0.523
Y7	0.226	-0.130
Y8	0.446	-0.195
Y9	0.693	0.258
Y10	0.102	0.127
Y11	0.520	0.099
Y12	0.076	0.393
Y13	0.492	0.026
Y14	0.935	-0.226
Y15	0.935	-0.226
Y16	0.793	0.208
Y17	0.384	0.354
Y18	0.392	0.512
Y19	0.478	0.545
Y20	0.385	0.539
Y21	-0.167	0.567
Y22	0.190	0.696
Y23	0.180	0.862
Y24	0.169	0.697
Y25	-0.078	0.883
Y26	-0.089	0.915
Y27	-0.211	0.837
Y28	-0.331	0.722
Y29	-0.258	0.739
Y30	-0.582	0.407

Table E2 - Promax Factor Correlations for the WPT - Q

	1	2
1	1.000	
2	0.254	1.000

Appendix F

Item Parameter Estimates and Standard Error Estimates for the WPT

Table F1 - Factor 1 Item Parameters for the WPT

Item	Paper		Internet		Paper		Internet	
	a	**SE**	**a**	**SE**	**b**	**SE**	**b**	**SE**
2	1.040	0.206	0.714	0.154	-1.900	0.172	-2.024	0.354
3	.565	0.146	0.505	0.144	-3.369	0.529	-3.997	1.056
5	.708	0.172	1.449	0.414	-3.033	0.397	-1.770	0.226
6	.954	0.195	0.849	0.185	-1.650	0.162	-1.873	0.314
7	.744	0.179	0.537	0.163	-2.731	0.345	-3.438	0.917
9	1.311	0.323	0.973	0.246	-1.908	0.161	-2.034	0.343
14	.654	0.129	0.527	0.110	-1.111	0.152	-1.434	0.290
15	.896	0.181	0.625	0.139	-1.430	0.150	-1.860	0.353
16	.435	0.098	0.628	0.129	-1.478	0.248	-1.248	0.234
20	.259	0.072	0.343	0.084	-2.838	0.628	-1.891	0.490
21	.414	0.094	0.318	0.080	-1.114	0.218	-1.827	0.514
25	.505	0.124	0.602	0.150	-2.360	0.389	-2.186	0.479
28	.510	0.128	0.780	0.202	-2.035	0.352	-1.531	0.325

Note – Item parameter estimates (but not standard error estimates) for the paper-and-pencil version of the WPT (Paper) have been transformed to the metric of the Internet version of the WPT (Internet) using Equate v2.1 (Baker, 1995).

Table F2 - Factor 2 Item parameters for the WPT

Item	Paper		Internet		Paper		Internet	
	a	**SE**	**a**	**SE**	**b**	**SE**	**b**	**SE**
11	.849	0.184	0.431	0.114	-1.191	0.212	-2.405	0.581
17	.424	0.094	0.786	0.217	-1.232	0.334	-1.336	0.255
19	.641	0.137	0.682	0.179	-1.566	0.324	-1.672	0.329
32	.677	0.143	0.645	0.191	-.706	0.219	-0.522	0.254
34	.303	0.075	0.428	0.132	.902	0.564	1.538	0.537
35	.708	0.159	0.680	0.213	-.153	0.208	-0.100	0.241
36	.542	0.130	0.680	0.213	.238	0.302	0.290	0.241
38	.401	0.101	0.568	0.185	.900	0.555	0.955	0.406
40	1.004	0.297	0.658	0.223	.638	0.315	0.991	0.426
41	.849	0.219	1.188	0.498	-.098	0.233	-0.005	0.235
43	.712	0.201	0.622	0.214	1.050	0.521	0.222	0.410
45	1.307	0.514	0.994	0.444	.738	0.411	1.174	0.460

Note – Item parameter estimates (but not standard error estimates) for the paper-and-pencil version of the WPT (Paper) have been transformed to the metric of the Internet version of the WPT (Internet) using Equate v2.1 (Baker, 1995).

Table F3 - Factor 3 Item Parameters for the WPT

Item	Paper		Internet		Paper		Internet	
	a	**SE**	**a**	**SE**	**b**	**SE**	**b**	**SE**
10	.717	0.146	0.716	0.127	.132	0.134	-0.165	0.122
12	.812	0.157	0.703	0.134	-.660	0.104	-1.093	0.189
18	.533	0.105	0.748	0.130	-.088	0.145	0.053	0.120
23	.645	0.130	0.846	0.176	-1.577	0.160	-1.654	0.256
26	.932	0.173	1.277	0.277	-.460	0.104	-0.345	0.110
27	1.024	0.217	0.924	0.173	-.397	0.105	0.035	0.126
29	.551	0.123	0.713	0.152	.891	0.257	0.814	0.200
30	.736	0.158	0.541	0.113	.088	0.155	-0.083	0.202
31	1.806	0.920	0.854	0.220	.768	0.163	1.116	0.228
33	1.085	0.307	1.553	0.523	.816	0.198	1.062	0.185
37	1.501	1.024	1.321	0.648	1.534	0.310	1.950	0.415
39	1.078	0.632	1.099	0.348	1.772	0.439	1.215	0.328
42	1.124	0.585	1.730	1.181	1.503	0.410	0.856	0.415
44	1.564	1.327	1.127	0.573	.895	0.447	1.476	0.499

Note – Item parameter estimates (but not standard error estimates) for the paper-and-pencil version of the WPT (Paper) have been transformed to the metric of the Internet version of the WPT (Internet) using Equate v2.1 (Baker, 1995).

Table F4 – Item Parameter Estimates for Items 1-45 of the WPT

Item	Paper		Internet		Paper		Internet	
	a	SE	a	SE	b	SE	b	SE
1	.423	0.148	.931	0.430	-6.697	1.991	-3.727	1.03134
2	1.013	0.155	.555	0.111	-1.991	0.147	-2.418	0.42403
3	.519	0.118	.374	0.096	-3.634	0.543	-5.171	1.36181
4	.343	0.077	.435	0.103	-3.218	0.537	-2.734	0.58696
5	.686	0.149	1.243	0.304	-3.162	0.384	-1.882	0.21873
6	.692	0.118	.623	0.118	-1.961	0.198	-2.288	0.36687
7	.672	0.138	.423	0.113	-2.960	0.3411	-4.189	1.04958
8	.297	0.063	.275	0.060	-1.229	0.251	-1.358	0.37844
9	.905	0.152	.799	0.179	-2.228	0.195	-2.279	0.34954
10	.846	0.138	.607	0.097	.106	0.106	-.186	0.12245
11	.362	0.079	.318	0.079	-2.353	0.355	-3.120	0.75233
12	.589	0.087	.638	0.108	-.616	0.119	-1.168	0.18463
13	.407	0.082	.272	0.068	-2.444	0.340	-3.905	0.98977
14	.566	0.098	.419	0.078	-1.264	0.150	-1.719	0.32954
15	.871	0.130	.862	0.147	-1.519	0.133	-1.529	0.18784
16	.426	0.080	.544	0.093	-1.585	0.235	-1.385	0.23255
17	.422	0.075	.425	0.087	-1.505	0.210	-2.071	0.41094
18	.551	0.094	.718	0.109	-.032	0.134	.059	0.10834
19	.397	0.079	.442	0.087	-2.459	0.362	-2.293	0.43087
20	.224	0.058	.274	0.063	-3.275	0.705	-2.319	0.57231
21	.447	0.079	.338	0.070	-1.151	0.178	-1.734	0.39876
22	.570	0.099	.732	0.115	.320	0.151	.021	0.10954
23	.660	0.109	.768	0.146	-1.393	0.159	-1.717	0.24596
24	.225	0.056	.276	0.063	.369	0.338	.736	0.29781
25	.610	0.119	.636	0.134	-2.166	0.275	-2.100	0.37868
26	.972	0.141	1.096	0.195	-.334	0.089	-.314	0.10067
27	.922	0.150	.857	0.142	-.251	0.097	.094	0.11152
28	.426	0.094	.769	0.159	-2.399	0.406	-1.556	0.26022
29	.586	0.11518	.781	0.152	.946	0.234	.810	0.15877
30	.901	0.169	.544	0.108	.153	0.121	-.035	0.18197
31	1.837	0.596	.962	0.246	.796	0.139	1.078	0.16625
32	.629	0.127	.433	0.100	-.947	0.168	-.732	0.31943
33	.994	0.232	1.283	0.369	.954	0.195	1.137	0.15452
34	.217	0.060	.302	0.082	1.117	0.599	2.057	0.66502
35	.544	0.122	.501	0.112	-.320	0.195	-.171	0.26795
36	.328	0.089	.505	0.121	.314	0.372	.355	0.27809
37	1.502	0.782	1.014	0.425	1.561	0.252	2.210	0.49946
38	.340	0.091	.342	0.095	.895	0.464	1.319	0.57501
39	1.126	0.521	1.132	0.339	1.788	0.388	1.275	0.30237
40	.407	0.116	.371	0.105	1.246	0.503	1.402	0.65744
41	.643	0.178	.798	0.231	-.352	0.218	-.282	0.27071
42	1.258	0.611	1.509	0.956	1.504	0.358	1.050	0.32904
43	.370	0.123	.528	0.162	1.791	0.803	.013	0.41350
44	1.633	1.302	1.146	0.611	.959	0.329	1.547	0.42108
45	.568	0.206	.531	0.177	1.171	0.580	1.797	0.66521

Note – Item parameter estimates (but not standard error estimates) for the paper-and-pencil version of the WPT (Paper) have been transformed to the metric of the Internet version of the WPT (Internet) using Equate v2.1 (Baker, 1995).

Appendix G

Item Parameter Estimates and Standard Error Estimates for the WPT-Q

Table G1 – Item Parameters for Factor 1 of the WPT-Q

	Proctored		Unproctored		Proctored		Unproctored	
Item	**a**	**SE**	**a**	**SE**	**a**	**SE**	**a**	**SE**
1	1.132	0.477	1.024	0.371	-2.515	0.803	-3.621	1.730
4	1.443	0.903	.979	0.296	-2.753	1.142	-3.033	1.135
5	1.116	0.591	1.940	1.037	-3.063	1.223	-2.270	0.543
6	1.516	0.722	.738	0.227	-1.761	0.440	-3.022	1.334
8	.563	0.194	1.056	0.304	-1.057	0.431	-1.241	0.333
9	1.034	0.484	.931	0.293	-2.632	1.029	-3.119	1.205
11	.887	0.311	.908	0.234	-1.602	0.436	-1.924	0.510
13	.672	0.243	.775	0.215	-1.987	0.680	-1.899	0.625
14	1.434	0.894	.866	0.279	-3.294	****	-3.049	1.311
15	1.434	0.597	1.466	0.611	-3.294	****	-2.492	0.758
16	1.381	0.794	1.233	0.338	-2.305	0.660	-1.975	0.476
18	.858	0.308	.963	0.268	-1.973	0.579	-1.764	0.520

Note – Item parameter estimates (but not standard error estimates) for the unproctored WPT-Q administration (Unproctored) have been transformed to the metric of the proctored WPT-Q administration (Proctored) using Equate v2.1 (Baker, 1995).

Table G2 – Item Parameters for the Factor 2 of the WPT-Q

	Proctored		Unproctored		Proctored		Unproctored	
Item	**a**	**SE**	**a**	**SE**	**b**	**SE**	**b**	**SE**
10	.462	0.131	.485	0.130	-.696	0.356	-1.091	0.396
12	.763	0.216	.593	0.156	-.610	0.264	-.618	0.293
17	.761	0.232	.566	0.155	-1.035	0.329	-1.091	0.382
19	.905	0.449	1.202	0.652	-3.763	1.397	-2.449	0.621
20	.526	0.177	1.017	0.451	-3.268	1.119	-2.349	0.610
21	.617	0.190	.893	0.257	.494	0.289	.548	0.256
22	.583	0.171	.444	0.124	-1.129	0.435	-.353	0.353
23	.986	0.502	.773	0.341	-2.921	0.954	-3.617	1.401
24	1.084	0.372	1.074	0.318	-.662	0.255	-.508	0.229
25	1.031	0.372	.745	0.216	-.441	0.252	-.558	0.272
26	.573	0.173	.428	0.131	-2.340	0.802	-2.463	0.867
27	.802	0.265	.729	0.230	.316	0.284	.058	0.299
28	.470	0.141	.528	0.158	.598	0.460	-.256	0.377
29	.631	0.212	.572	0.171	-.030	0.375	-.646	0.387

Note – Item parameter estimates (but not standard error estimates) for the unproctored WPT-Q administration (Unproctored) have been transformed to the metric of the proctored WPT-Q administration (Proctored) using Equate v2.1 (Baker, 1995).

Table G3 – Item parameters for items 3-29 of the WPT-Q

Item	Proctored		Unproctored		Proctored		Unproctored	
	a	**SE**	**a**	**SE**	**b**	**SE**	**b**	**SE**
3	.462	0.119	.431	0.115	-.272	0.261	-.360	0.278
4	1.093	0.624	.957	0.429	-3.327	0.984	-3.062	0.750
5	.720	0.330	1.327	0.781	-4.396	1.958	-2.647	0.518
6	1.190	0.496	.438	0.141	-2.032	0.409	-4.270	1.326
7	.351	0.101	.378	0.106	-1.947	0.625	-1.416	0.449
8	.327	0.091	.553	0.151	-1.719	0.590	-1.451	0.349
9	.882	0.363	.979	0.525	-3.020	0.874	-3.026	0.747
10	.397	0.104	.394	0.105	-.795	0.342	-1.305	0.393
11	.590	0.170	.565	0.177	-2.177	0.559	-2.368	0.560
12	.581	0.144	.530	0.135	-.741	0.256	-.697	0.253
13	.830	0.267	.409	0.126	-1.813	0.378	-2.662	0.751
14	.050	0.000	.690	0.279	-34.627	0.000	-3.461	1.002
15	.050	0.000	.834	0.426	-34.627	0.000	-3.620	1.128
16	1.009	0.495	.871	0.340	-2.791	0.769	-2.328	0.509
17	.820	0.228	.467	0.128	-1.005	0.251	-1.291	0.362
18	.728	0.228	.475	0.147	-2.271	0.553	-2.732	0.706
19	.842	0.414	.954	0.439	-3.891	1.522	-2.817	0.672
20	.595	0.213	1.156	0.513	-2.932	0.942	-2.264	0.441
21	.483	0.133	.852	0.253	.575	0.306	.595	0.220
22	.581	0.161	.363	0.101	-1.131	0.369	-.383	0.336
23	.942	0.471	.672	0.307	-2.920	0.973	-4.043	1.463
24	1.088	0.324	.771	0.191	-.657	0.200	-.589	0.218
25	1.040	0.343	.634	0.168	-.452.	0.206	-.597	0.239
26	.582	0.171	.443	0.133	-2.298	0.712	-2.427	0.706
27	.688	0.207	.576	0.175	.320	0.249	.103	0.275
28	.444	0.128	.403	0.118	.647	0.417	-.294	0.387
29	.631	0.201	.521	0.161	-.005	0.308	-.720	0.340*

Note – Item parameter estimates (but not standard error estimates) for the unproctored WPT-Q administration (Unproctored) have been transformed to the metric of the proctored WPT-Q administration (Proctored) using Equate v2.1 (Baker, 1995).