Web-Based Testing: Exploring The Relationship Between Hardware Usability And Test Performance

Kyle Huff, Georgia Gwinnett College, USA
Melinda Cline, Georgia Gwinnett College, USA
Carl S. Guynes, University of North Texas, USA

ABSTRACT

Web-based testing has recently become common in both academic and professional settings. A web-based test is administered through a web browser. Individuals may complete a web-based test at nearly any time and at any place. In addition, almost any computer lab can become a testing center. It is important to understand the environmental issues that may influence test performance. This study furthers our understanding of web-based testing. The research is conducted using an experimental method with 220 undergraduate student participants in an academic environment. Test performance effects are examined based on administration environment, computer hardware configuration, and distractions. Results indicate that minor differences in hardware configurations may have a significant effect on test results.

Keywords: Web-Based Testing; Internet Tests; Cognitive Ability; Usability

INTRODUCTION

The browser determines how information is displayed and the plug-ins (additional software that is used to enhance the web experience) that can be used. Web-based tests have a number of advantages over traditional testing methods. One advantage is that it is possible for an individual to complete a web-based test at nearly any time and at any place. In some cases, the required computer hardware is supplied by the tester, such as in a computer laboratory, and in other cases, the required internet access hardware is supplied by the test taker.

This study furthers our understanding of web-based testing. The research is conducted using an experimental method with 220 undergraduate student participants in an academic environment. Test performance effects are examined based on administration environment, computer hardware configuration, and distractions. Results indicate that minor differences in hardware configurations may have a significant effect on test results.

PRIOR RESEARCH

Prior research indicates that web-based tests may have increased psychometric properties over their paper-and-pencil counterparts (Buchanan, 2000). Advantages include researchers being able to collect data 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 do those of other administration modes. Furthermore, the direct input of answers by test-takers provides for more accurate scoring than 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; and Oswald, Carr, & Schmidt, 2001).

One of the major issues associated with web based testing is the administration mode - proctored vs. self-administered. Proctored web-based testing occurs when the test-taker completes the test form in the presence of a test administrator. Unproctored testing (self-administration) 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 web-based testing is increasingly being administered in unproctored settings. Oswald, Carr, and Schmidt (2001) conducted research on web-based personality and cognitive ability tests versus their paper-and-pencil counterparts. 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. They also found equivalence in cognitive ability tests. However, the researchers did not address practice effects, gather information about the environment in which the participants took the test in the unproctored setting, or measure the usability of the web-based tests.

Nielsen (2003) defines usability as a quality attribute of interfaces that assesses how easy it is to use a particular user interface. With a broad range of hardware and software settings available, differing configurations produce differences in the amount of information that is displayed on the screen and the legibility of that information. Nielson found that the usability of a website has been known to affect people's perceptions of a website as well as their willingness to use the website. Usability is important in test taking because the greater the usability of an interface, the quicker and easier a person will be able to complete the task. Poor usability may result in a person having difficulty figuring out how to take the test, how to answer the questions, how to read figures, or how to read the text of the test. It may cause people to complete the form incorrectly and/or take longer to complete, thereby introducing extraneous error into the assessment process.

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 from their paper-and-pencil version when administered on a computer (Buchanan, 2000; Gregory, 2007; and McBride, 1998). When a speed test is administered via a computer, people complete more of the items than they would otherwise because participants are able to select an answer quicker using the computer. 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. 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; however, these results were not found in the mathematics portion of the test. Scrolling between the passage text and the test items is posited to have caused the difference.

It is important to consider usability during web-based testing in both a proctored testing environment using uniform hardware and software configurations and in an unproctored environment in which computers can have a much broader range of hardware and software settings. 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.

Additionally, factors in the test taker's specific environment are important and may, under certain circumstances, cause a contamination of test scores. Environmental factors include temperature, humidity, illumination, and noise. Noise, in particular, is a factor that must be controlled in testing (Gregory, 2007). Noise has been shown to cause a decrease in performance on tasks, especially when the noise is unpredictable, intermittent, and loud (Boggs & Simon, 1968).

RESEARCH QUESTION AND MEASUREMENTS

The specific research question asked in this study is: "Does the computer configuration used by a participant, a proctored environment, and/or distractions in the environment impact a student's test score?" An experimental approach was chosen to investigate this question as it is suited to research focused on a small number of isolated variables (proctored/unproctored, usability, and distractions) and is often used when the investigator benefits from being able to manipulate the research setting (Yin, 1984).

The test used in this study is the *Wonderlic Personnel Test-Quicktest 2005* (WPT-Q). The WPT-Q is a 30-item, 8-minute timed test of cognitive ability widely used in personnel selection that is available for administration over the Internet. In addition to the WPT-Q, the participants were also asked to complete three questionnaires. The first assessed the conditions in which the test was completed to assess distractions in the environment, the second questionnaire assessed usability issues associated with the Internet administrations (connection speed, ease of use, and visual layout), and the third gathered data on various demographic variables (age, gender, ethnicity, primary language, college GPA, year in college, and SAT scores).

PARTICIPANTS

Participants in this study included 220 students enrolled in introductory psychology classes at a large southeastern public university during 2005. The students were randomly divided into three groups. Group 1 participated in the test using a 15 inch monitor in a standard computer laboratory environment with an assigned proctor. Group 2's environment was the same as Group 1's with the exception of the monitor size, which was 18 inches. Group 3 used their own computer equipment in a non-proctored environment. In exchange for their participation, the students each received one research credit to be used as partial fulfillment of course requirements.

PROCEDURES

Participants registered for an administration time at the campus Experimetrix website and were randomly assigned to the computer laboratory or self-administration groups. The procedures for each are described as follows:

Computer Laboratory

Upon entering the computer laboratory, participants were instructed to sit in front of a computer terminal containing instructions that guided them through the study. Computers used in this study were 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). Participants were required to first read an Internet-based informed consent form and then they received instructions on taking the WPT-Q. Administration of the WPT-Q followed the standardized instructions. After completing the WPT-Q, participants completed the three additional questionnaires.

Self-administration

Participants in this group first reported to a computer laboratory similar to the one described above. They were then asked to sit at a computer to read an electronic informed consent form and the directions for participating in the study. They also completed a web-based form to provide their contact information. They were then excused from the computer laboratory and were sent three subsequent e-mails. The first e-mail was a set of instructions explaining how to participate in the study. The second was an invitation to complete the WPT-Q and the third and final e-mail contained a link to the questionnaires.

RESULTS

Only 212 participants out of the original 220, were used in the analysis. Of the eight 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 self-administration group. The technical problems that were reported arose 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.

Demographics

As a preliminary step, various demographic factors were analyzed to compare the computer laboratory and self-administration groups' equivalence. All *t* tests were two-tailed. The groups were equivalent on English as a

first language $\chi^2(1, N = 205) = 0.0427$, $p > .05$, ethnicity $\chi^2(6, N = 204) = 2.3802$, $p > .05$, age $t(188) = 0.83$, $p = .83$, $d = .12$, credit hours completed $t(195) = -1.14$, $p = .2516$, $d = -.16$, GPA $t(141) = 0.804$, $p = .2334$, $d = .14$, verbal SAT score $t(174) = .78$, $p = .4368$, $d = .12$ and Quantitative SAT score $t(174) = -0.42$, $p = .6715$, $d = -.06$. However, the two groups were different in terms of their gender makeup in that the computer laboratory group contained significantly more males than females $\chi^2(1, N = 207) = 4.8791$, $p \leq .05$.

To further test the gender factor, the investigators performed a Fisher's *z* transformation. The data were grouped by gender to compare Pearson correlation coefficients between total SAT scores and WPT-Q scores, ignoring the effects of test administration. 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 a statistically non-significant difference between the two correlations ($z = 1.353$, $p = .1761$, two-tailed).

RESEARCH QUESTION ASSESSMENT

To see if the tests were equivalent across administration methods, first a t-test was conducted to compare both proctored groups to the self-administration group. The result ($t(206) = .632$, $p = .528$) showed no significant differences. However, when results of the WPT-Q scores were compared between the three groups, a different picture emerged. Results of a one-way ANOVA revealed a significant difference ($f(2,207) = 3.068$, $p = .049$, *eta square* = .029). Results of a LSD post-hoc analysis revealed that the participants who completed the test in the computer lab with the 15" monitors scored significantly higher than those who completed the test in the lab with the 18" monitor (mean difference = 1.725, $p = .018$) and the self-administration group (mean difference = 1.288, $p = .05$). Table 1 includes the means and standard deviations for the three test administration methods.

Table 1: Means and Standard Deviations for Self-Administered, 15-inch monitor, and 18-inch monitor groups

	Self-Administered		15-inch Monitor		18-inch Monitor	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Score	20.92	3.600	22.21	3.439	20.48	4.119
Number attempted	26.86	3.830	26.96	3.242	25.35	4.250
Usability Score	29.74	4.111	31.75	3.199	29.97	3.746

The correlation coefficients between the proctored and unproctored groups' WPT-Q scores and criterion measure 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 computer laboratory ($r = .41$, $df = 90$, $p < .0001$) and self-administration groups ($r = .63$, $df = 82$, $p < .0001$). These correlations were then analyzed using Fisher's *z* transformation. This analysis resulted in a significant difference - $z = 1.991$, $p = .0465$.

To compare the usability of the computer laboratory and self-administration 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 $\alpha = .88$. 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 a two-tailed independent-samples *t* test. The sample means for the computer laboratory and self-administered groups were 30.76 and 29.74, respectively, and were not significantly different - $t(206) = 1.90$, $p = .0584$, $d = .26$.

When the effects of the hardware and software settings of the two different computer laboratories were compared for Groups 1 and 2, however, a different picture emerged. Additional ANOVAs were conducted on the number of questions attempted by the participants (an estimate of efficiency) and the usability scores as independent variables to try to assess the overall usability of the systems and WPT-Q scores as the dependent variable. Results

revealed significant differences for both independent variables, $f(2,207) = 3.465$, $p = .033$, partial $\eta^2 = .033$, for the last question answered and $f(2,207) = 4.769$, $p = .009$, partial $\eta^2 = .044$.

To determine if distractions in the environment had an effect on test performance, the scores were regrouped based on the respective responses on the environmental questionnaire, regardless of whether the participants were originally in the proctored or unproctored group. Participants were placed in one of two groups. The first group ($n = 103$) consisted of participants who reported no events occurring in the environment when they completed the WPT-Q and had a mean WPT-Q score of 21.31 ($sd = 3.908$). The second group ($n = 105$) consisted of participants who reported that some event occurred while they were completing the WPT-Q and had a mean score of 20.88 ($sd = 3.613$). They did not have to find these events distracting in order to be placed in this group. The two groups were found to be not significantly different on test scores - $t(206) = .833$, $p = .406$, $d = .1161$.

DISCUSSION

The results indicate that the correlations between the WPT-Q scores and the total SAT scores were significantly different for the computer laboratory and self-administered groups. Demographic variables were ruled out as a factor in the differences. 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* analysis, it was believed that the gender factor could be ruled out as the cause of the differences. This is consistent with other research on gender differences in cognitive ability (Feingold, 1996).

Results indicate that differences in scores are attributable to differences in environmental factors. Groups 1 and 2, which both completed the test in a proctored computer laboratory environment, had significant differences not only in usability scores, but also in number of items attempted. Statistical results indicate that participants who took the test on the smaller monitor (15 inch) out-performed participants who took the test on the larger monitor (18 inch) and they rated the test as easier to use. These results contradicted the results of Bridgemen, Lennon, and Jackenthal (2003) who found that participants who completed the test on a larger monitor had higher scores on reading comprehension items. However, these results agreed with their general findings that screen settings and resolution matter. In addition, these results went beyond their findings as this study indicates that usability can affect non-reading comprehension items. Based on these results, it is concluded that hardware and software settings have an impact on timed test performance. Environmental distractions were not found to be a significant issue in any of the testing environments observed.

Two plausible explanations for the performance differences due to screen size have to do with mouse movements and the display area size. Mouse movements may have been a factor during test taking because even though the layout on the two different monitors was very similar in the amount of information displayed on the screen, it was believed that fine movements were needed to move the mouse pointer. Perhaps it took longer to complete these movements on the larger monitors. If true, it could mean that participants would have had less time to spend on test items, thus leading to the lower tests scores.

An alternate explanation has to do with the display area size of the monitors. The larger monitor has a larger display area than the smaller monitor. Since the display area was larger, participants had to pay attention to more space on the larger monitor than on the smaller monitor. Perhaps the increased cognitive load led to participants taking longer to complete each item. Since it took longer to answer each item, participants were unable to attempt as many items on the larger monitor. Once again, this delay would lead to lower test scores.

These findings indicate that usability of the specific computer equipment influences test performance and should be considered as a significant component in assessment preparation, particularly when test administration with be timed. One of the often listed advantages of computer and internet testing is that it allows the testing of new constructs or existing constructs in ways that were impossible with paper-based tests (McBride, 1998; and Gregory, 2007). These new tests could include having test takers listen to audio files or watch video clips. Regardless of what these types of test would use, it is important to keep in mind the simple lesson of usability; otherwise, error will be introduced into the assessment.

LIMITATIONS

A potential limitation of 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. Other limitations include sample size, the inability to control for cheating in the self-administration group, and a lack of detail data about each computer configuration among the self-administration participants. Further studies are required to examine the possibility of the importance of these factors.

DIRECTIONS FOR FUTURE RESEARCH

Continued research is necessary to determine optimal environments for assessments, how varying time allotments should be made for a test taker based on the specific computer configuration used, and the best possible format for a test to reduce the impact of the technology on the outcome score. This line of research is more important now than ever as we see an integration of a number of devices for Web access. In addition to the traditional desktop and laptop computers, students and professionals now commonly use smart phones, net books, and tablet computers to access the Web to complete a plethora of information-based tasks. It is important to continue to examine how the information is presented and to examine the effects of differences in performance based on the method of access. By doing so, we can learn how best to reduce technology bias in assessment results. In conclusion, interesting results were found. Some of these results supported past research and some did not. In all, this research suggests that mode of administration matters.

AUTHOR INFORMATION

Dr. Kyle Huff is an Assistant Professor of Management in the School of Business at Georgia Gwinnett College in Lawrenceville, Georgia. He teaches general management, human resource management, management of technology, and quantitative analysis. His education includes a B.S. in Management from the Georgia Institute of Technology, an M.S. in General Psychology from Georgia College and State University, and a Ph.D. in Industrial/Organizational and Vocational Psychology at North Carolina State. His project work has focused on numerous clients including the US Military, O*NET, North Carolina State Bureau of Investigations, and NASA. In addition, he has worked as a Test Engineer for VeriTest and as a Researcher for Predvari. E-mail: khuff@ggc.edu.

Dr. Melinda Cline is an Associate Professor of Management Information Systems in the School of Business at Georgia Gwinnett College in Lawrenceville, Georgia.. She received her Ph.D. in Management and Information Science from Florida State University in 1999, after having spent 15 years in the railroad industry. Her professional experience includes numerous leadership and project management positions while designing and implementing information systems for CSX Transportation, New Zealand Railways, Queensland Railways, and the Southern Pacific Transportation Company. Her research interests include information systems evaluation, information security, and project management. Her research is published in numerous journals including the *Journal of Computer Information Systems, Information Systems Management, Decision Support Systems,* and the *Managerial Auditing Journal*. E-mail: mcline@ggc.edu.

Carl S. Guynes is a Regents Professor of Business Computer Information Systems at the University of North Texas. He received a doctorate in quantitative analysis from Texas Tech University. Dr. Guynes' areas of specialization are client/server computing, data administration, and information resource management. His most recent research efforts have been directed in the areas of client/server computing and data administration. Some of the journals in which Dr. Guynes has published include and *Communications of the ACM, Information & Management, The Journal of Information Systems Management, Journal of Accountancy, Journal of Systems Management, The Journal of Database Management, The CPA Journal, The Journal of Computer Information Systems, Information Strategy, Computers and Security, and Computers and Society*. E-mail: Guynes@unt.edu. Corresponding author.

REFERENCES

1. Boggs, D. H. & Simon, J. R. (1968). Differential effect of noise on tasks of varying complexity. *Journal of Applied Psychology, 52*, 148 – 153.
2. 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.
3. 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.
4. 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.
5. Feingold, A. (1996). Cognitive gender differences: where are they and why are they there? *Learning and Individual Differences, 8*(1), 25 – 32.
6. Gregory, R. J. (2007). *Psychological testing: history, principles, and applications* (5th ed.). Boston: Pearson Education.
7. 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.
8. 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.
9. 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.
10. Nielsen, J (2003). Usability 101: Introduction to usability. Retrieved August 1, 2004 from http://www.useit.com/alertbox/20030825.html.
11. 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.
12. 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.
13. Sinar, E. & Reynolds, D. (2004). Exploring the impact of unstandardized Internet testing. Paper presented at the Society for Industrial Organizational Psychology, Chicago, IL.
14. Wonderlic (2005). Wonderlic Personnel Test-Quicktest. Retrieved January 25, 2005 from http://www.wonderlic.com/products/product.asp?prod_id=35
15. Yin, R. (1984). *Case study research: Design and Methods* (1st ed.). Beverly Hills, CA: Sage Publishing.
16. Zimowski, M., Muraki, E., Mislevy, R. & Bock, D. (2002). BILOG-MG 3 [Computer Program]. Chicago, Scientific Software, Inc.

NOTES

