

# Comparison of a Paper-and-Pencil to a Web-Enabled Version of the Thurstone Test of Mental Alertness (TMA™)

*Michael R. Cunningham, Ph.D., University of Louisville, Kelly D. Dages, Ph.D., Vangent, Inc., and John W. Jones, Ph.D., Vangent, Inc.*

*October 27, 2008*

## EXECUTIVE SUMMARY

The purpose of this research paper was to document the comparability level of the Thurstone Test of Mental Alertness (TMA™) when administered in both a paper-and-pencil format and an online, web-enabled format. Two studies are reported in this paper. First, a review of meta-analytic research that explored the impact of computer versus paper-and-pencil test administration on test scores was conducted to ensure that our expectation related to the impact of different test administration mediums (i.e., online versus paper-and-pencil) on TMA scores was grounded in scientific research. Sure enough, a review and extension of all accessible research on this topic documented a very high degree of similarity between computerized and paper-and-pencil administration methods with cognitive ability tests. That is, based on 51 available correlations ( $N = 15,866$ ), the current analysis found an estimated  $\rho = .78$  between computerized and paper-and-pencil administered cognitive assessments. The second study focused exclusively on the TMA and studied a sample of diverse college students. The obtained results were consistent with the Study I results and supported the notion that students would obtain similar scores on the Internet version of the TMA when compared to the scores they produced when they took the TMA in the traditional paper-and-pencil format. Hence, a fairly reasonable conclusion was reached that the reliability and validity data produced to support the paper-and-pencil version of the TMA can also be used to support the reliability and the validity of the Internet version of the TMA. Future studies are planned to further support this conclusion.

Paper accepted at *Midwestern Psychological Associates*, Annual Conference, Chicago, IL, May 2009.

## Comparison of a Paper-and-Pencil to a Web-Enabled Version of the Thurstone Test of Mental Alertness (TMA™)

Computerized presentation of pre-employment cognitive ability tests offers increases in both efficiency and effectiveness. Computerized testing is more efficient, because respondents enter their responses in a format that can be instantly scored and analyzed, eliminating the need for the test administrator to use a scoring template, or fax the responses to the test publisher for analysis. Online presentation of a computerized test can increase effectiveness by allowing employment recruitment and selection to occur anywhere, at anytime. That can expand the pool of available prospects to organizations that are eager to find talented employees to fill their positions.

Whenever a technological solution is implemented, it is wise to question whether the new approach produces the same accuracy and reliability as the tried-and-true predecessor. The degree of similarity of scores produced on paper-and-pencil versus computer tests may vary depending on whether the test is speeded or allows an unlimited amount of time for assessment (called power tests), and on other aspects of test procedures and administration, such as the nature of the hardware, the test taking environment, and whether or not the test is proctored or not. According to the American Psychological Association (1985), scores of conventional and computerized versions of a test may be considered equivalent if the rank-order of individuals tested with these two tests are close.

This paper presents the outcomes of two studies that support the comparability of paper-and-pencil and computerized administration methods for cognitive ability tests in general, and for the Thurstone Test of Mental Alertness (TMA) in particular. The first study, a meta-analytic investigation, was conducted to examine the equivalence of computerized tests compared with paper-and-pencil tests. The purpose of conducting the first study was to ensure that the expectation that there would be little difference in test scores regardless of whether the TMA was administered via the Internet or in the more traditional paper-and-pencil format was grounded in research. The second study was conducted to specifically compare the performance on internet and paper-and-pencil administrations of the Thurstone Test of Mental Alertness (TMA) using a sample of college students.

### STUDY I: META-ANALYSIS OF ONLINE COMPUTERIZED AND PAPER-AND-PENCIL COGNITIVE ABILITY TESTS

Meta-analysis is a statistical technique for combining and reviewing previous quantitative research. The current meta-analysis is an update to a meta-analysis completed by Mead and Drasgow (1993). The Mead and Drasgow meta-analysis investigated the impact of test media on performance using 29 published papers and unpublished papers from conferences and individual researchers. The estimated correlation between paper-and-pencil and computer versions of an untimed test was  $\rho = .95$ , based on 123 correlations.<sup>1</sup> This indicates

---

<sup>1</sup> The effect size measurement ( $\rho$ ) was the mean of the correlation between the two presentation media, weighted by the sample size, and corrected for unreliability (attenuation).

a high degree of similarity in performance between the two administration methods. The results are a bit more complex for speeded tests, because it may take a respondent a differing amount of time to darken a bubble with a pencil versus click a button to indicate a response, and respondents may have differing amounts of experience responding quickly with a computer keyboard or mouse. In addition, fewer studies were located that used speeded tests compared to power tests. The cross-media correlation for speeded tests was estimated to be  $\rho = .67^2$ , based on 36 correlations. This correlation is moderately strong, but it does indicate that speeded tests may be slightly affected by the mode of administration, both in terms of the way that the item stimuli are recorded on screen, and how the responses are captured. For example, it may take less time to move a pencil one inch to mark a response on a paper-and-pencil test than to first click a button to indicate the answer choice, then click “next” to see the next item on a computerized test. Mead and Drasgow (1993) cautioned that “Empirically established validity of inferences made from a paper-and-pencil speeded test should not be assumed to automatically generalize to a corresponding computerized test.”

To further examine the equivalence of internet computerized tests compared with paper-and-pencil tests, a meta-analysis of the literature was conducted. The literature review focused on studies that were conducted after the Mead and Drasgow (1993) review was published. A comprehensive search of the literature, dissertation abstracts, and relevant conference papers was completed.<sup>3</sup> The selection criteria for papers included the following: (1) the study must include data from both a paper-and-pencil and a computerized test; (2) the sample must be normal adults, rather than school children, mentally challenged individuals or clinical patients; (3) the test must measure general cognitive ability, rather than mastery of a specific subject matter, such as computer programming; (4) participants in the paper-and-pencil and computerized test conditions should be drawn from the same population. There were studies that did not meet these criteria, and they were not included in the meta-analysis.

A handful of useable studies were located following the Mead and Drasgow meta-analysis. Some studies used a within-subjects design while others used a between-subjects design. Additionally some researchers considered order effects of the mode of administration, while others did not. For example, one study tested the equivalence of paper-and-pencil and computerized tests in proctored versus unproctored environments (Oswald, Carr, and Schmidt, 2001). They investigated the Air Force Qualifying test using a 2 x 2 between-subjects factorial research design (proctored vs. unproctored and paper-and-pencil vs. web). The authors reported that no differences were evident for performance in the paper-and-pencil versus the computerized test, or the proctored versus unproctored conditions, suggesting that the tests were equivalent across the media of presentation. A narrative summary of each study used in the meta-analysis is presented in Appendix A. Additionally, Appendix A includes a table with the means, correlations and other data used to complete the meta-analysis. The results of the meta-analysis are described below.

---

<sup>2</sup> This rho is corrected for attenuation, but with no r that is corrected above unity (1.0). Allowing correlations above 1.0 produces  $\rho = .72$  (a figure that is cited in other papers).

<sup>3</sup> Authors of dissertations were contacted for copies of their work, as were authors of papers presented at the Society of Industrial and Organizational Psychology (SIOP).

## Results

### Meta-analysis of correlations

The present meta-analysis combined all 36 of the correlations for speeded tests ( $n=10,339$ ) available from Mead and Drasgow, and merged them with correlations produced since that report. The meta-analysis of correlations was conducted using the procedures of Hedges & Olkin (1961). The results of the meta-analysis for the Mead and Drasgow correlations, three additional studies using within-subjects design, and the combined correlations are presented in Table 1. As mentioned above, the Mead and Drasgow data estimated the relation between scores on a computerized test compared to a paper-and-pencil test to be  $\rho = .67$  ( $p < .05$ ). Three studies using a within-subjects design that became available since Mead and Drasgow reported a total of 15 correlations between paper-and-pencil and computerized test scores (Neuman and Baydoun, 1998; deBeer and Visser, 1998; and Potosky and Bobko, 2004). Because Cronbach's alphas were not available for all studies, none of the parameters are corrected for attenuation. For the newer studies,  $\rho = .91$  ( $p < .05$ ). Taking all 51 available correlations, the  $\rho = .78$  ( $p < .05$ ). Such results indicate high similarity between the results obtained in computerized versus paper-and-pencil assessments of cognitive abilities.

**Table 1. Correlations between Computer and Paper-and-pencil Administration**

Source	Rho	Number of correlations	N	Confidence Limits
Mead and Drasgow (1993)	.67	36	10,339	.66 -.68
Neuman and Baydoun (1998), deBeer and Visser (1998) and Potosky and Bobko (2004)	.91	15	4,907	.90 -.92
<b>Total</b>	<b>.78</b>	<b>51</b>	<b>15,866</b>	<b>.77 -.79</b>

### Meta-analysis of mean differences

A strong correlation between computer administered and paper-and-pencil versions of a test indicated that the test takers maintained comparable ranks from one form of the test to another. The possibility remains, however, that the scores could differ between the two forms of the test, requiring the use of separate norms tables for interpreting the results produced by the two media. Mead and Drasgow (1993) did not present data on the difference between computer and paper-and-pencil administered tests, but 26 pairs of means were available from studies produced since that time. Ten of the pairs of means were based on between-subjects designs (Van de Vijver & Harsveld; Oswald, Carr, & Schmidt; and Huff) and 16 were based on within-subjects designs (Neuman & Baydoun; de Beer & Visser; Dembowski & Callans, Potosky & Bobko). Neuman & Baydoun did not report repeated measures ANOVA or paired-sample t-tests, and paired sample t-tests cannot be calculated on the basis of descriptive statistics. Consequently, independent sample t-tests were calculated as a proxy (halving the number of subjects per group to produce appropriate degrees of freedom). A meta-analysis of the between-group difference ( $d'$ ) scores was conducted using the procedures described by Hunter, Schmidt, and Jackson (1982). Table 2 presents the effect sizes for the group differences for the within-subjects and between-subjects studies. Ignoring the direction of the difference, the effect size for the between-group studies and the 16 within-group difference scores were small effects. Combining all 26 contrasts produced a small difference. If the direction of the difference is retained, such that outcomes

favoring paper-and-pencil tests are scored positively and outcomes favoring computerized tests are scored negatively, even smaller differences are noted. The between-group difference shows a negligible advantage for computerized tests, while within-group studies show a small advantage for paper-and-pencil tests. When all 26 contrasts are combined, the effect shows a negligible advantage for paper-and-pencil. These results indicate test takers do not produce meaningfully higher scores on either paper-and-pencil or computerized tests.

**Table 2. Mean Differences between Computer and Paper-and-pencil Administration**

Contrast Type	Number of Contrasts	N	d' <sup>4</sup>	References
<b>Direction of Contrast Ignored</b>				
Between group	10	3,518	.36	Van de Vijver & Harsveld (1994); Oswald, Carr, & Schmidt (2001); and Huff (2007)
Within-group	16	5,157	.28	Neuman & Baydoun (1998); de Beer & Visser (1998); Dembowski & Callans (2000); Potosky & Bobko (2004)
<b>Total</b>	<b>26</b>	<b>8,675</b>	<b>.31</b>	
<b>Direction of Contrast Retained</b>				
Between group	10	3,518	-.07	Van de Vijver & Harsveld (1994); Oswald, Carr, & Schmidt (2001); and Huff (2007)
Within-group	16	5,157	.23	Neuman & Baydoun (1998); de Beer & Visser (1998); Dembowski & Callans (2000); Potosky & Bobko (2004)
<b>Total</b>	<b>26</b>	<b>8,675</b>	<b>.10</b>	

## Conclusions

The meta-analysis demonstrates a high degree of similarity between computerized and paper-and-pencil administration methods for cognitive ability tests. This was an update to the Mead and Drasgow (1993) meta-analysis, which found an estimated correlation for speeded tests to be  $\rho = .67$ . Based on 51 available correlations ( $n = 15,866$ ), the current analyses found an estimated correlation of  $\rho = .78$  between computerized and paper-and-pencil administered cognitive assessments. Additionally, a comparison of the mean difference between assessment scores suggest small to negligible differences in assessment performance. Such results indicate high similarity between the results obtained in computerized versus paper-and-pencil assessments of cognitive abilities.

<sup>4</sup> Cohen (1988) defined an effect size of  $d' \leq .20$  as negligible,  $d' > .20$  and  $< .50$  as small,  $d' \geq .50$  and  $< .80$  as medium and  $d' > .80$  as large.

## STUDY II: ONLINE COMPUTERIZED AND PAPER-AND-PENCIL COMPARISON OF TMA

In this second study, an experiment was conducted to compare the scores of individuals who completed both the paper-and-pencil and online computerized versions of the Thurstone Test of Mental Alertness (TMA). The TMA is designed to assess general mental ability and knowledge of general information. The TMA includes both verbal and quantitative items. According to Thurstone and Thurstone's (1996) TMA Information Guide:

*It measures an individual's capacity to acquire new knowledge and skills and apply them to problem solving. It also measures individual differences in ability to learn and perform mental tasks of varying types and complexity. Four job-related tasks are assessed by the TMA test: adjusting to new situations, learning new skills quickly, understanding complex or subtle relationships, and thinking flexibly. (p. 1)*

Previous research has demonstrated that the TMA is a reliable and valid assessment. The test-retest reliability of the TMA is  $r=.95$  over a one-month period (Thurstone & Thurstone, 1997, p.8). A series of studies was conducted to examine the concurrent validity of the TMA (Thurstone & Thurstone, 1997, p.11-13). The concurrent validity studies are summarized in Table 3. As expected, the studies demonstrate that the TMA has strong relationships with other cognitive ability assessments.

**Table 3. TMA Concurrent Validity Studies**

Assessment	Sample	Correlation with TMA
Iowa Test of Educational Development	9th grade students (n=977)	.83
Iowa Test of Educational Development	12th grade students (n=545)	.85
SRA <sup>®</sup> High School Placement Test (HSPT)	High School freshmen (n=102)	.73
Wechsler Adult Intelligence Scale (WAIS)	Supervisors, managers, staff, and salespeople aged 22 to 70 years (n=200)	.61a
Wechsler Adult Intelligence Scale Revised (WAIS-R)	Undergraduate students (23 females and 9 males) aged 19 to 49 years (n=32)	.74

<sup>a</sup> This coefficient is lower than would probably result if the coefficients were based on a full range of scores from the general population (only four cases in this group had WAIS scores lower than 100).

A variety of additional studies demonstrated that the TMA was positively correlated with on-the-job performance. Some of the previously completed TMA studies are summarized in Table 4. Such results indicated that “The TMA Total score appears to be an adequate measure of intelligence for the vocational and industrial-organizational settings for which it was originally developed and for which it continues to be used and recommended” (Thurstone & Thurstone, 1997, p.15).

**Table 4. TMA Job-related Validation**

Job Performance Criteria	Sample	Relationship with TMA
General effectiveness ratings	Sales Supervisors (n=202)	Top 15% scored significantly higher than bottom 15%
General effectiveness ratings	Retail sales employees (n =1274)	Top 15% scored significantly higher than bottom 15%
General effectiveness ratings	Retail store managers (n=201)	r = .54
Sales Figures	Retail store managers (n=61)	r =.49
Supervisor ratings	Bank tellers (n=77)	r =.55
Performance under pressure rating	Clerical workers (n=27)	r =.56
Supervisor ratings	Salaried office workers (n=192)	r =.31

The initial studies on the reliability and validity of the test were conducted using the paper-and-pencil version of the TMA. It is reasonable to ask if individuals obtain similar scores on the computerized, on-line version of the test compared to what they obtain with the paper-and-pencil version. Therefore, an experiment was conducted to compare the scores of individuals who complete both versions of the tests.

## Method

*Research Participants.* One hundred and sixteen potential participants volunteered from lower level undergraduate courses at a large university in the southeast United States. Of the 116 volunteers, seventy-two reported (62%) for the study at the laboratory of the first author. Participants received course extra credit for participating. Most research participants were in the workforce, forty-six participants (63.9%) reported being employed, and twenty-six (36.1%) reported being full-time students. The participants included thirty males (41.7%) and forty-two females (58.3%). The sample was diverse; it included fifty-two Whites (72.2%), seventeen African-Americans (23.6%), three Asians (2.8%) and one Latino (1.4%). Forty participants were 18-20 (55.6%), twenty-two were 20-25 (30.6%), eight were 25-30 (11.1%), and two were older than 30 (2.8%).

*Instruments.* Form A of the Thurstone Test of Mental Alertness was used for both the paper-and-pencil and computerized internet conditions. The internet and paper-and-pencil versions were administered in a proctored, laboratory environment.

The TMA consists of 126 items, and respondents are required to correctly answer as many questions as possible within 20-minutes. The TMA Verbal items include Same-Opposite questions, which ask the respondent to select a word that means the same or the opposite of the word presented, and Definitions, which asks the respondent to read a definition and mark the word that matches the definition. The Quantitative items include Arithmetic, which asks the respondent to provide a numerical solution to a word problem, and Number Series, which asks the respondent to determine what number comes next in a series of numbers that have a definite numerical order. The verbal and quantitative items are scored separately and then combined for a total score.

In the paper-and-pencil version of the TMA, respondents are required to make a pencil mark to designate which of four or five possible answers is correct. The answers are presented to the right of the questions within the five pages of the test booklet. Respondents are able to see many questions at the same time, to jump ahead or back, and to change their answers. Time is kept by a test proctor. In the computerized version of the test, the same TMA items are presented to respondents via the internet. The items are presented one at a time, and the respondent clicks a button to indicate the chosen answer, and a second button to move to the next question. Time is kept by a computerized clock which counts backwards from 20 minutes. Respondents click a button to submit their answers when they have completed all 126 items, or the time has expired.

*Procedure.* Experimental sessions were offered at various times throughout the day and involved one to five participants at a time. There was always at least one experimenter present and most of the time there were two<sup>5</sup>. When research participants arrived at the laboratory, they were reminded that the study was designed to compare a paper-and-pencil test with a computerized online version of the same test, and were asked to read and sign an informed consent statement. Research participants were seated at individual carrels with side walls that prevented research participants from seeing each other's work. Each carrel contained a computer that was already turned on and linked to the test site. Despite the computer, the carrels had ample space for participants to work on the paper-and-pencil version of the test. The carrel also was supplied with the paper-and-pencil test, two pieces of scratch paper, and a sharpened #2 pencil.

Research participants were randomly assigned to the paper-and-pencil version first or internet version first experimental conditions by flipping a coin before the start of the experimental session. Thirty-four (47.2%) participants completed the paper-and-pencil test first, and thirty-eight subjects (52.8%) completed the computerized version first.

All individuals began one version of the test at the same time. After 20 minutes had elapsed, research participants were asked to discontinue their work on the test. The first scratch sheet was collected, and the paper-and-pencil test if applicable, so participants could not see their prior work or answers. As soon as all research participants were ready, they completed the other version of the test within the same 20 minute time period. Contrary to instructions, two research participants started on the paper-and-pencil test immediately after completing the computerized test. Because their start time was not recorded, their data were discarded. Research participants were not informed of their performance on either version of the test.

---

<sup>5</sup> The authors would like to express their gratitude for the conscientious assistance of Justin Combs from the University of Louisville as an experimenter for this research.

## Results

Preliminary analysis was conducted to compare the performance of the research participants who took the computerized online version of the TMA first with a moderately large sample ( $n=596$ ) of individuals who completed the computerized online version of the TMA as part of the job application process for a variety of industries. The mean scores for the student research participants and the job applicants are presented in Table 5. The scores of the two groups did not differ significantly on the total TMA score, verbal scores, or the quantitative scores. For more descriptive information on the sample of job applicants that completed the online version of the TMA, see Appendix B. A comparable sample of employment applicants who took the paper-and-pencil version of the TMA was not available, but the fact that half of a randomly divided experimental sample was comparable to an job applicant sample seems adequate to demonstrate that the experimental sample was representative of the job applicant population, and appropriate to test the impact of mode of administration on test performance.

**Table 5. Online Administered TMA Scores for Research Sample and Job Applicants**

	Student Research Sample	Online Job Applicants	<i>t</i> -test
Verbal	30.97	33.53	$t(631) = -1.71, p < .09$
Quantitative	24.76	25.83	$t(631) = -.90, p < .37$
<b>Total TMA</b>	<b>55.73</b>	<b>59.35</b>	<b><math>t(631) = -1.46, p &lt; .15</math></b>

The initial set of primary analyses contrasted the research participant's total TMA scores on the paper-and-pencil tests with the same individual's score on the computerized test, regardless of which test was presented first. The scores obtained on the computerized and paper-and-pencil tests were remarkably similar (see Table 6). Comparable results were obtained for the verbal scores and the quantitative scores. The similarity of the scores supports the expectation that the online computerized version and the paper-and-pencil versions are comparable.

**Table 6. TMA Scores for Paper and Internet Administration**

	Paper Administration	Internet Administration	<i>t</i> -test
Verbal	33.96	34.13	$t(67) = -.14, p < .89$
Quantitative	28.37	27.10	$t(67) = 1.39, p < .17$
<b>Total TMA</b>	<b>62.32</b>	<b>61.24</b>	<b><math>t(67) = .54, p &lt; .59</math></b>

Next, a series of repeated measures ANOVAs were conducted on the TMA total, verbal, and quantitative scores. There were no significant effects of the medium of presentation on test performance nor was there any impact of whether the paper-and-pencil or computer version was presented first. The resulting  $F$  statistics for each effect are presented in Table 7. There was, however, a significant interaction effect. The interaction indicated that if research participants took the paper-and-pencil test first, they received a slightly higher score on the computerized test, whereas if they took the computerized test first they received a slightly higher score on the paper-and-pencil test. Thus, regardless of medium of presentation, research participants received a higher score the second time that they took the TMA (total/ verbal/ quantitative scores for paper-and-pencil = 54.94/ 30.24/ 24.70; computer = 55.77/30.71/25.06) than the first time that they took the TMA (total/ verbal/ quantitative scores for paper-and-pencil = 67.03/37.46/29.27; computer = 69.29/37.76/31.83). Research participants probably became more accustomed to the questions, and faster at answering them as a result of taking the first test, despite the absence of performance feedback.

**Table 7. Comparison of TMA Scores by Administration Method and Condition**

	Test Media $F$ -test	Administration Condition $F$ -test	Interaction $F$ -test
Verbal	$F(1,66) = .19, p < .67$	$F(1,66) = .01, p < .97$	$F(1,66) = 64.47, p < .0001$
Quantitative	$F(1,66) = 3.39, p < .07$	$F(1,66) = .64, p < .43$	$F(1,66) = 90.65, p < .0001$
<b>Total TMA</b>	<b><math>F(1,66) = .31, p &lt; .58</math></b>	<b><math>F(1,66) = .16, p &lt; .69</math></b>	<b><math>F(1,66) = 99.54, p &lt; .0001</math></b>

Finally, correlations were calculated between the individuals' scores on the two administrations. For research participants who completed the paper-and-pencil test first, the test-retest correlations were all highly significant. The correlations are presented in Table 8. For those who completed the computerized test first, the test-retest correlations were even higher. It is not clear why scores were particularly stable when research participants took the computerized test first, but it appears that the computerized assessment produced an estimate of the individual's ability that was quite reliable, as confirmed by the paper-and-pencil assessment.

**Table 8. Correlations between TMA Scores by Administration Condition**

	Paper-and-pencil Version first	Internet Version first
Verbal	.71**	.94**
Quantitative	.86**	.79**
<b>Total TMA</b>	<b>.81**</b>	<b>.93**</b>

\*\*correlations significant ( $p < .01$ )

---

## Conclusions

The similarity of scores obtained on the computerized version of the TMA delivered via the internet with the scores produced by the same individuals who took the TMA in the traditional paper-and-pencil format supports the conclusion that the two tests are comparable. As a consequence, the reliability and validity data produced to support the paper-and-pencil version of the TMA also can be used to support the reliability and validity of the computerized version of the TMA.

Similar studies have also found equivalence between paper-and-pencil and computerized versions of cognitive tests. One such study, included in the Study I meta-analysis, demonstrated the equivalence of a timed online computerized version of a cognitive ability test with a paper-and-pencil version (Dembowski & Callans, 2000). The study used a similar within-subjects design with order of administration method counter-balanced. Overall, the researchers concluded that they demonstrated equivalence between the administration modes. Additionally, the meta-analysis results indicate high similarity between the results obtained in computerized versus paper-and-pencil assessments of cognitive abilities.

Future research will continue to focus on ways to further develop and refine the online version of the TMA. For example, both the meta-analytic and the TMA-specific set of analyses in Study 2 confirmed the high degree of comparability between the two test administration mediums (i.e. online versus paper-and-pencil). Still, some TMA users might prefer to compare the online TMA scores to a normative sample of online job applicants. In fact, Appendix B actually presents an early version of such norms for a sample of N=596 online job applicants. This online norm group will continue to be increased and refined. In addition, future studies with the online version of the TMA can further explore the value of administering the online TMA in an untimed format, and subsequently developing untimed online norms. For now, however, both Studies I and II together indicate that with proper administration, online computerized versions of cognitive ability tests, including the web-enabled TMA, can be considered equivalent to their corresponding paper-and-pencil versions.

---

## REFERENCES

- American Psychological Association, Committee on Professional Standards and Committee on Psychological Tests and Assessment (1985). Guidelines for computer-based tests and interpretations. Washington, DC: American Psychological Association.
- Booth-Kewley S., Edwards J.E., & Rosenfeld P. (1992). Impression management, social desirability, and computer administration of attitude questionnaires: Does the computer make a difference? *Journal of Applied Psychology*, 77, 563-566.
- 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.
- Cohen, J. (1988). Statistical power analyses for the behavioral sciences (2<sup>nd</sup> Ed.). Hillsdale, N.J.: Lawrence Erlbaum.
- de Beer, M., & Visser, D. (1998). Comparability of the paper-and-pencil and computerized adaptive versions of the General Scholastic Aptitude Test. *South African Journal of Psychology*, 28, p21-27.
- 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.
- Gibson, W., & Weiner, J. (1997). Equivalence of computer-based and paper-based cognitive ability tests. Paper presented at the twelfth annual conference of the Society for Industrial and Organizational Psychology, St. Louis, MO.
- Hedges, L., & Olkin, I. (1985). Statistical Methods for Meta-Analysis. New York: Academic Press.
- Huff, K. (2006). The Effects of Mode of Administration on Timed Cognitive Ability Tests. Unpublished dissertation, North Carolina State University, Greensboro.
- Huff, K (2007). DFIT analysis of Web-based and Paper-based Versions of the WPT. Poster presented at the Society for Industrial and Organizational Psychology.
- Huff, K., & Michael J. (2007). The Effects of Proctoring on Web-based Timed Cognitive Ability Scores. Paper presented at the Society for Industrial and Organizational Psychology.
- Hunter, J.E., Schmidt, F.L., & Jackson, G.B. (1982). Meta-Analysis: Cumulating Research Findings Across Studies. Beverly Hills: Sage.

- King W.C. & Miles E.W. (1995). A quasi-experimental assessment of the effect of computerizing non- cognitive pencil-and-paper measurements: A test of measurement equivalence. *Journal of Applied Psychology*, 80, 643-651.
- Mead A.D., & Drasgow F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, 114(3), 449-458.
- Neuman, G., & Baydoun, R. (1998). Computerization of paper-and-pencil tests: When are they equivalent? *Applied Psychological Measurement*. 22, 71-83.
- 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.
- Potosky, D., & Bobko, P. (2004). Selection testing via the internet: Practical considerations and exploratory empirical findings. *Personnel Psychology*, 57, 1003-1034.
- Preckel, F., & Thiemann, H. (2003). Online- versus paper-and-pencil version of a high potential intelligence test. *Swiss Journal of Psychology*, 62, 131-138.
- Sinar, E., & Reynolds, D. (2004). Exploring the impact of unstandardized internet testing. Paper presented at the Society for Industrial Organizational Psychology, Chicago, IL.
- Thurstone, L.L., & Thurstone, T.G. (1997). Thurstone Test of Mental Alertness (TMA™) Examiner's Manual. Arlington, VA: Vangent, Inc.
- Thurstone, L.L., & Thurstone, T.G. (1996). Thurstone Test of Mental Alertness (TMA™) Information Guide. Arlington, VA: Vangent, Inc.
- Van de Vijver F.J.R., & Harsveld M. (1994). The incomplete equivalence of the paper-and-pencil and computerized versions of the General Aptitude Test Battery. *Journal of Applied Psychology*, 79(6), 852-859.

## APPENDIX A: META-ANALYSIS RESEARCH STUDY SUMMARIES

Van de Vijver and Harsveld (1994) compared performance on the computerized version of the General Aptitude Test Battery (GATB) with the paper-and-pencil version using 326 applicants to the Royal Military Academy in the Netherlands. More items were answered, and more items were correctly solved, using the computerized version of the Name Comparison, Three-Dimensional Space, and Tool Matching subtests, whereas respondents performed better on the paper-and-pencil version of Vocabulary and Form Matching, likely due to differences in the format of the presentation. Examination of the proportion of items correctly solved to the items attempted suggested that the two media may have induced different levels of speed and accuracy, with faster responding in the computerized version, and more accurate responding in the paper-and-pencil version. All differences were reported to be  $p < .05$ , but statistics were not provided.

de Beer and Visser (1998) conducted a within-subjects design. They compared achievement in one form of the General Scholastic Aptitude Test (GSAT) paper-and-pencil test with the GSAT computerized adaptive test. In some cases the paper-and-pencil test was administered first followed by the standard computerized version, while in other cases the order of administration was reversed. In both of these versions, the items and the sequence in which they were administered were the same. According to the authors: "The results indicate that achievement in the paper-and-pencil GSAT and the standard computerized version of the GSAT were not equivalent because the examinees performed better in the paper-and-pencil version of the GSAT than in the standard computerized version of the GSAT. The scores on the six subtests of the GSAT standard test were calculated separately in order to determine whether any of the subtests were influenced more than the others in the computerized version. It appears that the overall trend of the difference in total score is reflected in each of the separate subtests." The authors' conclusion might be unduly negative. At the same time that differences in the absolute scores were obtained, there were strong correlations between performances on the two versions of the test.

Neuman and Baydoun (1998) used a within-subjects design to compare the paper-and-pencil and computer-based version of the Office Skills Test. The Office Skills Test is a battery of 10 tests designed to assess the clerical ability of job applicants for entry level positions, and includes checking, coding, filing, grammar, numerical skills, oral directions, punctuation, reading comprehension, spelling, and vocabulary. Each test is completed in a set time period, monitored via a stopwatch in the paper-and-pencil format. The subjects were 411 undergraduate students who were randomly assigned to receive first either the paper-and-pencil test, completed in groups of 15-20, or the computerized test, completed at one of 3 computer terminals. The time between administrations of the two tests was four weeks. Strong correlations were found between performance on the paper-and-pencil tests and computer-based tests (CBTs). Unfortunately, paired t-test for the differences between groups in mean performance were not presented.

One of the few studies to demonstrate the equivalence of the web-based computerized version of a cognitive ability test with a paper-and-pencil version was conducted with the *Wonderlic Personnel Test* (Dembowski & Callans, 2000). The subjects were 250 student applicants to a national computer applications training center. The materials used in the study included two forms of the WPT, Form IV and Form V, which were administered to each applicant in both paper-and-pencil and computerized formats. The WPT is a short-form

test of general cognitive ability. It contains 50 questions and is timed at 12 minutes. As part of this study, applicants were given the opportunity to complete both tests, computer and paper, untimed (after the timed section). Subjects were randomly assigned to either the computer or paper group, and half received Form IV, while the other half received Form V. At time 2, the subjects received the opposite mode and form of the test. The entire design was counter-balanced. The time between testing was a minimum of one-half hour. Using the alternate forms conversion table, Form V scores were adjusted downward by 1 point. Using the age adjustment table, the appropriate number of points was added to the 12 minute raw score. They also adjusted untimed scores by subtracting six points to estimate the applicant's true timed score. The authors reported that there was a mean difference of -.06 between the computer and paper tests such that the computer test was slightly easier. There was a mean difference of .42 between Form IV and Form V, such that Form V was slightly easier. There also was a median difference of 1.0 between the computer and paper tests (the paper test is slightly higher), and a median difference of -1.0 between Form IV and Form V (Form V is slightly higher). Overall, the researchers concluded that they demonstrated equivalence, not only between the two alternate forms, but also between the administration modes.

In a study to test the equivalence of paper-and-pencil and computerized tests in proctored versus unproctored environments, Oswald, Carr, and Schmidt (2001) conducted research on the math and verbal subsections of the Air Force Qualifying test with 410 psychology undergraduates. Using a 2 x 2 between subjects factorial research design (proctored vs. unproctored and paper-and-pencil vs. web), the researchers used both descriptive statistics and confirmatory factor analysis to compare the equivalence of the tests in the four conditions. The authors reported that no differences were evident for performance in the paper-and-pencil versus the computerized test or the proctored versus unproctored conditions, suggesting that the tests were equivalent across the media of presentation.

Potosky & Bobko (2004) conducted a repeated measures study of the Test of Learning Ability (TLA), in which respondents completed both the computerized and paper-and-pencil versions of the test. The Test of Learning Ability is a 54-item, multiple-choice, timed test of cognitive ability that assesses spatial reasoning (block counting), mathematical ability (arithmetic), and verbal ability (vocabulary). *T*-test results indicated that mean scores on the Web version of the TLA ( $M = 39.19$ ,  $SD = 8.51$ ) were significantly lower than mean scores on the paper-and-pencil version of the TLA ( $M = 40.55$ ,  $SD = 6.71$ ;  $t = 2.27$ ,  $p < .05$ ,  $df = 53$ ). Regardless of mode of administration, scores were higher the second time the TLA was administered, when averaged across modes ( $t = 5.16$ ,  $p < .01$ ,  $df = 53$ ). The cross-mode correlations for each subscale varied depending on the test; math  $r = .74$ , vocabulary  $r = .58$ , and block counting  $r = .44$ . Overall, the correlation between the Internet version and the paper-and-pencil version of the TLA was  $r = .60$  ( $p < .001$ ). Thus, there was evidence for equivalence in respondent ranking, but not in equivalence of total score.

Huff (2006) conducted a two-study dissertation research project, which was presented at SIOP. The first study compared two archival data sets involving the *Wonderlic Personnel Test, Form I* (WPT-I), which is a 50 item test of cognitive ability for use in personnel selection, which has a 12 minute time limit, and is available in both paper-and-pencil and web-based versions. Wonderlic provided that data, involving 325 participants who completed the paper-and-pencil version of the WPT-I and 325 participants who completed the web-based WPT-I in a proctored setting. In the analysis of the two versions of the WPT-I, the means for the

paper-and-pencil and web-based of the test were 20.52 ( $std = 6.267$ ) and 22.10 ( $std = 5.669$ ), respectively. The results of a  $t$ -test to compare the means yielded  $t(648) = 3.380, p = .001, d = .27$ . Inspection of the scores on individual items indicated that in 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. This would suggest that the web-based version took longer to complete. However, for almost every item, the ratio of correct versus attempted was higher for the web-based group than it was for the paper-and-pencil group. Taken together, it appears that the participants in the web-based group tended to work at slower pace because they were working more carefully to complete the test. The investigator also reported effects due to the type of response format (a checkbox, a radio button, or a textbox), and the size of the computer monitor. Thus, this study showed a small amount of non-equivalence between the two test media.

Huff's second study (Huff & Michael, 2007) did not meet the selection criteria for the meta-analysis, but the results are noteworthy. This study compared proctored and unproctored administration of the *Wonderlic Personnel Test-Q*. The *Wonderlic Personnel Test-Quicktest* (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. 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). There was no significant difference (proctored  $M=21.25, sd=3.91$ , unproctored  $M=20.92, sd=3.60$ ). To determine if distractions in the environment had an effect on test performance, the investigator placed subjects in one of two groups, regardless of whether the participants were originally in the proctored or in the unproctored group. The first group ( $n = 103$ ) consisted of participants who reported no distracting events occurring in the environment when they completed the WPT-Q and had a mean 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 they had a mean score of 20.88 ( $sd = 3.613$ ). Respondents did not have to find these events distracting in order to be placed in the second group. The difference between the two groups was not significant ( $t(206) = .833, p = .406, d = .161$ ), suggesting that test performance is not strongly influenced by events in the testing environment.

Table 9. Meta-analysis Study Summaries

Author	Date	Test	Subscale	Design	P & P Means	Computerized Means	df	t	p	d'	k	r
Mead & Drasgow	1993	ASVAP									22	0.79
Mead & Drasgow	1993	DAT									6	0.34
Mead & Drasgow	1993	Misc									8	0.6
Neuman & Baydoun	1998	OST	Checking	Within	20.88	21.30	409	-0.861	0.39	-0.085	1	0.97
Neuman & Baydoun	1998	OST	Coding	Within	45.78	44.50	409	3.294	0.001	0.325	1	0.98
Neuman & Baydoun	1998	OST	Filing	Within	34.92	35.76	409	-0.607	0.54	-0.061	1	0.90
Neuman & Baydoun	1998	OST	Grammar	Within	22.84	23.22	409	-0.741	0.44	-0.073	1	0.83
Neuman & Baydoun	1998	OST	Numerical Skills	Within	15.88	14.80	409	2.714	0.001	0.267	1	0.93
Neuman & Baydoun	1998	OST	Punctuation	Within	11.82	11.86	409	-0.087	0.84	-0.009	1	0.92
Neuman & Baydoun	1998	OST	Oral Directions	Within	10.65	10.15	409	1.388	0.18	0.137	1	0.91
Neuman & Baydoun	1998	OST	Reading	Within	10.66	10.84	409	-0.647	0.52	-0.064	1	0.84
Neuman & Baydoun	1998	OST	Spelling	Within	43.46	43.56	409	-0.089	0.84	-0.009	1	0.84
Neuman & Baydoun	1998	OST	Vocabulary	Within	25.78	24.22	409	2.347	0.02	0.232	1	0.80
de Beer & Visser	1998	GSAT	Verbal P&P 1st	Within	50.42	50.30	147	0.4	0.69	0.07	1	0.96
de Beer & Visser	1998	GSAT	Verbal Comp 1st	Within	51.63	46.76	220	15.67	0.0001	2.1	1	0.91
de Beer & Visser	1998	GSAT	Nonverbal P&P 1st	Within	45.78	45.00	147	1.64	0.1	0.27	1	0.88
de Beer & Visser	1998	GSAT	Nonverbal Comp 1st	Within	49.51	44.57	222	11.89	0.0001	1.6	1	0.85
Potosky & Bobko	2004	TLA		Within	40.55	39.19	53	2.27	0.05	0.43	1	0.60
Van de Vijver & Harsveld	1994	GATB	Name Comparison	Between	75.50	87.00	348	-6.92	0.0001	-0.74		
Van de Vijver & Harsveld	1994	GATB	Computation	Between	24.30	24.60	348	-0.748	0.455	-0.08		
Van de Vijver & Harsveld	1994	GATB	3D Space	Between	26.20	24.40	348	3.393	0.0001	0.36		
Van de Vijver & Harsveld	1994	GATB	Vocabulary	Between	32.10	30.00	348	3.89	0.0001	0.42		
Van de Vijver & Harsveld	1994	GATB	Tool Matching	Between	36.80	40.20	348	-5.83	0.0001	-0.62		
Van de Vijver & Harsveld	1994	GATB	Arithmetic Reasoning	Between	16.60	16.60	348	0	1	0		
Van de Vijver & Harsveld	1994	GATB	Form Matching	Between	35.40	31.40	348	5.944	0.0001	0.64		
Dembowski & Callans	2000	WPT		Between	19.75	19.81	248	-0.18	0.857	-0.02		
Oswald	2001	AFQT	Supervised	Between	12.26	12.45	256	-0.389	0.697	-0.05		
Oswald	2001	AFQT	Unsupervised	Between	12.34	12.14	158	0.33	0.742	0.12		
Huff	2006	WPT-I		Between	20.52	22.10	648	-3.38	0.001	-0.38		

## APPENDIX B: ONLINE JOB CANDIDATE SAMPLE

The sample of job candidates (n=596) that completed the TMA online consists of respondents from a variety of industries including retail, education, service, manufacturing, and technology. The majority of the sample consists of job applicants (approximately 90%). The sample includes approximately 12% managerial candidates, 35% professional level candidates, and 53% entry/non-exempt level candidates. In Table 10, the descriptive statistics for this sample of job candidates are presented. Table 11 presents the percentile distribution for the TMA Verbal, Quantitative, and Total scores.

**Table 10. Online Administered TMA Scores for Job Candidates**

	Mean	Standard Deviation	Range
Verbal	33.53	8.75	10 - 60
Quantitative	25.83	6.98	6 - 47
<b>Total TMA</b>	<b>59.35</b>	<b>14.59</b>	<b>22 - 101</b>

**Table 11. TMA Percentiles for Online Administered TMA Scores**

Percentile	Verbal Score	Quantitative Score	Total TMA Score	Percentile
99	≥52	≥42	≥93	99
95	48	38	85	95
90	45	35	79	90
80	41	31	72	80
70	38	29	66	70
60	35	27	62	60
50	32	25	58	50
40	30	24	55	40
30	28	22	51	30
20	26	20	46	20
10	22	17	41	10
5	19	14	36	5
1	≤14	≤10	≤27	1