Graduate Management Admission Council®

Assessing the Equivalence of Internet-Based vs. Paper-and-Pencil Psychometric Tests

Michal Baumer, Keren Roded, and Naomi Gafni

National Institute for Testing and Evaluation (NITE), Jerusalem, Israel

Presented at the CAT Research and Applications Around the World Poster Session, June 2, 2009

Abstract

The equivalence of paper-and-pencil (P&P) and computer-based tests (CBTs) has become an important focus of research in the past 20 years. However, few studies have specifically addressed the equivalence of Internet-based tests (IBTs) and P&P administrations of high-stakes admissions tests (Potosky & Bobko, 2004). Despite the fact that there is a shortage of evidence with regard to the equivalence of scores obtained in the IBT and P&P modalities, the number of tests administered via the Internet is constantly rising. The goal of the present study was to compare the scores of examinees who took the P&P version of a scholastic ability test with the scores of those who took it via the Internet. The study was conducted using the Psychometric Entrance Test used for admission to institutions of higher education in Israel. 370 examinees participated in the study. Half were given a Web-based format in a computer lab and the other half were given the same test in P&P format. The study confirmed the equivalence between IBT and traditional P&P versions of the test for the sample.

Acknowledgment

Presentation of this paper at the 2009 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Citation

Baumer, M., Roded, K., & Gafni, N. (2009). Assessing the equivalence of Internet-based vs. paper-and-pencil psychometric tests. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

Keren Roded, National Institute for Testing and Evaluation,
P.O.B 26015, Jerusalem, 91260, Israel.
Email: keren@nite.org.il

Assessing the Equivalence of Internet-Based vs. Paper-and-Pencil Psychometric Tests

Background

The equivalence of paper-and-pencil (P&P) and computer-based tests (CBTs) has become an important focus of research in the past 20 years. However, few studies have specifically addressed the equivalence of Internet-based tests (IBTs) and P&P administrations of high-stakes admissions tests (Potosky & Bobko, 2004). Also, most of the research occurs in the personality arena with very little testing occurring for Web-based ability tests (Huff, 2006). Despite the fact that there is a shortage of evidence with regard to the equivalence of scores obtained in the IBT and P&P modalities, the number of tests administered via the Internet is constantly rising.

The issue of equivalence arises because certain differences exist between CBTs and P&P tests, and in addition there are differences between IBTs and other CBTs. In the former case, the differences lie in the presentation of the items (particularly reading comprehension passages and questions that have graphic components), in the method of answering, and in time allotment method. In the case of IBTs, additional factors come into play, for example: interruptions to the power supply, non-standardized computers in different laboratories, Internet server problems (particularly the impact of heavy traffic on the server), greater risk of items becoming known to other examinees, and the challenge of handling problems during the administration itself.

In general, studies dealing with the question of equivalence have yielded mixed results. K-12 research on the comparability of CBT and P&P tests generally show that computer and paper versions of multiple-choice tests are comparable across grades and academic subjects (Paek, 2005; Wang, Jiao, Young, Brooks, & Olson, 2008). Bicanich, Slivinski, Hardwicke & Kapes (1997) compared students' performance on IBT and P&P versions of a multiple-choice test and found the two modalities to be equivalent. Dembowski & Callans (2000) demonstrated equivalence between IBT and P&P versions of a cognitive ability test using the Wonderlic Personnel Test. Other researchers, however, obtained different results. For example, Potosky & Bobko (2004) conducted a study comparing P&P versions of a timed cognitive ability test to an IBT, and found moderate cross-mode correlation and significantly different means between the two modalities of administration.

One possible reason for non-equivalence of computer- and paper-based tests might be the specific state of the technology at the time of the research. For example, Paek (2005) stated in her research review that comparability studies tend to show that, in cases of tests that include extensive reading passages, there is lower performance on computer-based tests than on paper tests. However, she also stated that as computer interfaces include more tools that enhance a student's reading comprehension, this gap might disappear. Another example relates specifically to IBTs: in some studies, the researchers encountered Internet transmission problems (Potosky & Bobko, 2004; Bicanich, Slivinski, Hardwicke, & Kapes, 1997). However, IBTs administered over a high-speed Internet connection without network failures would appear very similar to CBTs (Huff, 2006). It seems that as long as the technology continues to develop, there will be an ongoing need for comparability studies to ensure the equivalence of test scores obtained in the two modalities.

Objective

The goal of the present study was to compare achievement scores on a linear IBT with scores on a P&P version of the test. The study was conducted using the Psychometric Entrance Test (PET). The PET, developed and administered by NITE, is used for admission to institutions of higher education in Israel. It is a high-stakes examination designed to assess ability in three domains, each consists of several types of items:

1. **Verbal Reasoning:** Analogies, Logic, Reading Comprehension, Letter Switching, Sentence Completions, Words and Expressions
   (two sections, 30 items in each)
2. **Quantitative Reasoning:** Questions and Problems in Algebra and Geometry, Diagrams and Tables, Quantitative Comparison
   (two sections, 25 items in each)
3. **English as a Foreign Language:** Reading Comprehension, Restatements, Sentence Completions
   (two sections, 27 items in each)

All items are in a multiple-choice format, and the time allotted for each section is 25 minutes. The number-correct score in each domain is scaled to range from 50 to 150 with a mean of 100 and a SD of 20 in the base population. An overall score is computed as the weighted sum of the domain scores and scaled back to a mean of 500 and a SD of 100 (range: 200-800). The relative weights of the three domains are 2, 2, and 1 for Verbal Reasoning, Quantitative Reasoning, and English respectively.

At the present time, most examinees take the P&P version of the PET. It is anticipated that administration of IBTs will be expanded. Given that this process will be gradual, with the test being administered in two parallel modalities for a period of time, establishing the equivalence of scores is of paramount importance.

Method

381 subjects were tested in an experimental administration of PET. A sample of 1,844 applicants who registered for the October 2008 Hebrew administration of the PET were invited to participate in the experiment (on a voluntary basis). Those who agreed to participate were randomly divided into two groups, one of which was given an IBT and the other a P&P version. They had no prior knowledge as to which version they would be given. After the test, the participants were asked to complete a feedback questionnaire. They were then given their scores.

Of the 381 examinees tested in the experimental administration, 370 (185 per group) took the operational PET as scheduled (one month after the experiment) and were therefore included in the analyses.

Instruments of Measurement

1. *One PET form in two versions*: Internet-based and P&P.
2. *Feedback questionnaires for the two groups* (IBT and P&P). The questionnaires elicited background information about the examinees, as well as their attitudes toward the idea of an Internet-based test.

Procedure

An identical PET form was administered to two groups (IBT and P&P). This experimental administration was conducted in classrooms and computer laboratories, in conditions approximating those of a real test. After the scheduled operational administration, four scores (one for each domain and an overall score) were computed for each examinee, both in the experimental and operational tests. The experimental test scores obtained in the two modalities were then compared, after correcting for differences in ability between groups (as reflected in the operational test).

Results

Examinees from both experimental groups achieved slightly higher results in the operational test than in the experimental administration (Table 1).

Table 1. Means and SDs of Experimental and Operational Test Scores for the IBT and P&P Groups (N=185 per group)

	Experimental PET			
	IBT		P&P	
Domain	**Mean**	**SD**	**Mean**	**SD**
Verbal	115	19	112	17
Quantitative	113	16	112	17
English	113	22	113	22
Overall score	***578***	***90***	***569***	***91***
	Operational PET			
	IBT		P&P	
Domain	**Mean**	**SD**	**Mean**	**SD**
Verbal	117	18	115	18
Quantitative	118	17	118	17
English	114	24	117	23
Overall score	***597***	***96***	***595***	***91***

Following the operational administration, minor differences in ability between the two groups were controlled for using analysis of covariance (with the score on the operational test serving as the covariate) and comparable scores in the experiment were extracted (Table 2). No

significant differences between the two groups were found in the overall score, in the verbal reasoning score, or in the quantitative reasoning score. In the English domain, the mean score on the IBT was significantly higher than that on the P&P, though the effect size was small (0.13 SD).

Table 2. Means of Experimental Test Scores Controlled for Ability for the IBT and P&P Groups (N=185 per group)

Domain	IBT Mean	P&P Mean	Pr > F
Verbal	114	113	0.2766
Quantitative	112	112	0.9294
English	115	111	0.0005
Overall score	***577***	***571***	***0.0751***

Pearson correlations between scores on the experimental and operational test were computed for each group (Table 3). The correlations between the overall scores, as well as the scores in each domain, were similar for the two groups, and corresponded to the established PET test-retest correlations.

Table 3. Correlations Between Experimental and Operational Test Scores for the IBT and P&P Groups

Domain	IBT	P&P
Verbal	0.84	0.89
Quantitative	0.83	0.85
English	0.93	0.91
Overall score	***0.93***	***0.94***

Performance on Different Types of Items

Performance on items by type (on the experimental test) was analyzed for each group (Table 4). For most item types performance on the computerized test was higher, and in all item types the effect size was small. These results can rule out concerns that certain item types might become more difficult when viewed on the computer screen.

Table 4. Number of Correct Responses in the Experimental Test (Controlled for Ability), by Group and Item Type

		IBT	P&P		
Domain	**Item Type**	**Mean**	**Mean**	**Number of Items**	**Effect size**
Verbal	Analogies *	8.1	7.7	12	0.17
	Logic	7.4	7.6	12	-0.08
	Reading Comprehension (~450 words)	6.4	6.2	10	0.08
	Letter Switching	5.7	5.6	8	0.06
	Sentence Completions *	7.6	7.9	10	-0.15
	Words and Expressions *	5.4	5.1	8	0.15
Quantitative	Questions and Problems	16.7	16.7	29	-0.01
	Diagrams and Tables	5.6	5.7	8	-0.06
	Quantitative Comparison	9.2	8.9	12	0.15
English	Reading Comprehension ** (~350 words)	14	13	20	0.21
	Restatements *	7.8	7.4	12	0.13
	Sentence Completions	14.2	13.7	22	0.09

* Significant difference at level of 0.05.
** Significant difference at level of 0.01.

Computer Familiarity

The relationship between level of computer familiarity and performance on the IBT was examined. Examinees were surveyed with regard to how frequently they use a computer, and two distinct categories emerged: "low" (0-5 times per week) and "high" (6+ times per week). The scores of those who use computers more often were significantly higher in both groups (IBT and P&P) – see Table 5 and Figures 1 and 2. After controlling for ability, no meaningful relationship between level of computer familiarity and performance on the IBT was found (p <= .11).

Table 5. Overall Score on the Experimental and Operational Tests, by Computer Familiarity

	Experimental PET					
	IBT			P&P		
Computer Familiarity	**Mean**	**SD**	**N**	**Mean**	**SD**	**N**
Low	558	86	101	543	86	88
High	606	86	82	594	89	96
	Operational PET					
	IBT			P&P		
Computer Familiarity	**Mean**	**SD**	**N**	**Mean**	**SD**	**N**
Low	576	95	101	573	90	88
High	626	91	82	615	88	96

Figure 1. Overall Score on the Experimental Test, by Group and Computer Familiarity

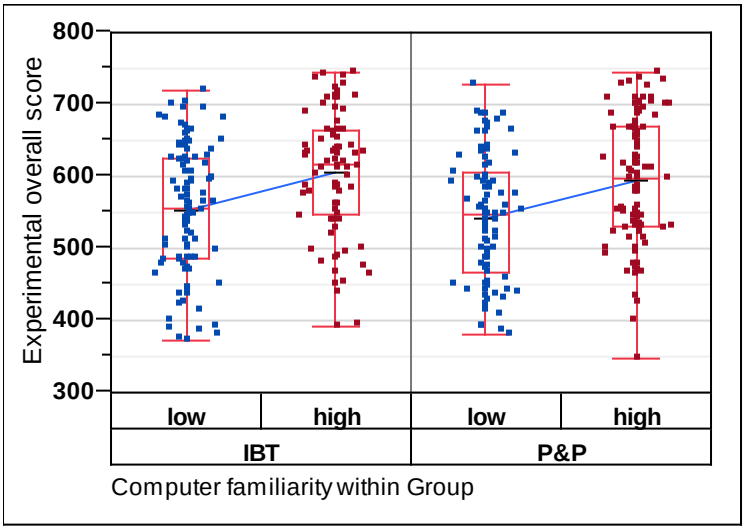
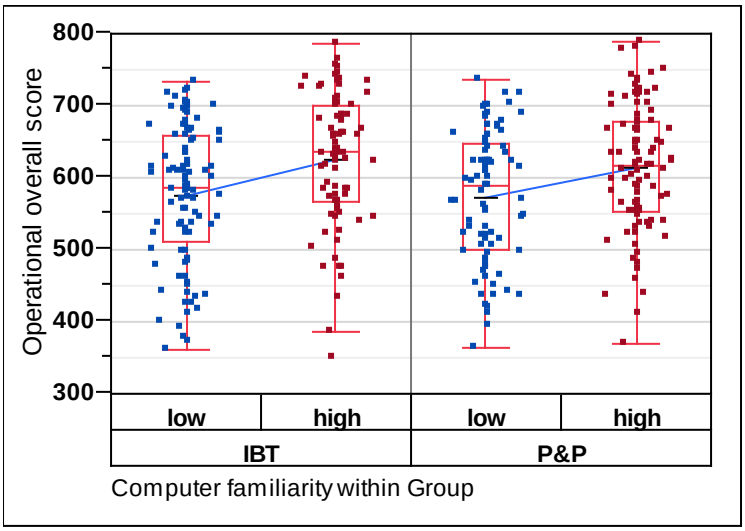


Figure 2. Overall Score on the Operational Test, by Group and Computer Familiarity



Gender Differences

Performance on the IBT was analyzed for gender differences. Males obtained significantly higher scores than females in both groups (IBT and P&P) – see Table 6 and Figures 3and 4. No significant interaction between gender and group was found (p <= .96). It appears that the performance gap between males and females that is already in evidence on the P&P tests did not grow larger on the IBT.

Table 6. Overall Score on the Experimental and Operational Tests, by Gender

	Experimental PET					
	IBT			P&P		
Gender	**Mean**	**SD**	**N**	**Mean**	**SD**	**N**
Male	619	83	69	597	87	86
Female	552	85	110	542	85	93
	Operational PET					
	IBT			P&P		
Gender	**Mean**	**SD**	**N**	**Mean**	**SD**	**N**
Male	632	90	69	617	85	86
Female	573	95	110	572	90	93

Figure 3. Overall Score on the Experimental Test, by Group and Gender

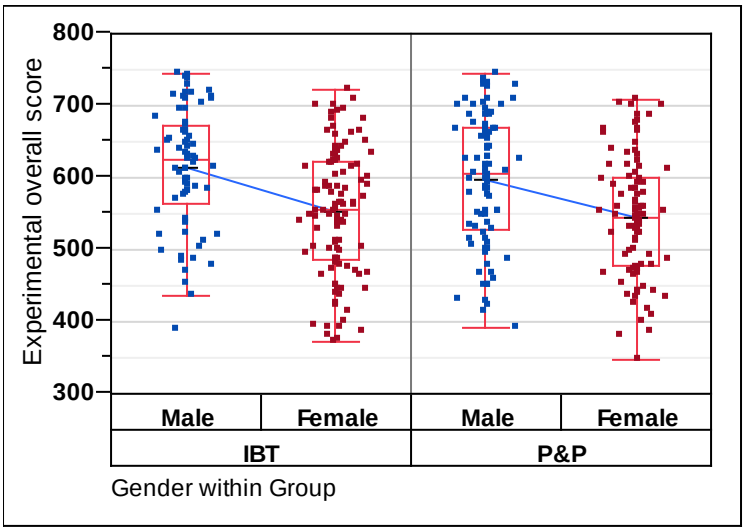
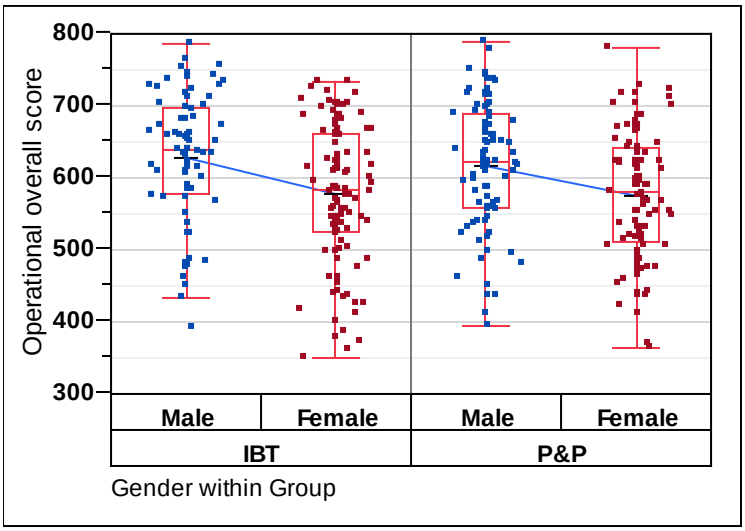


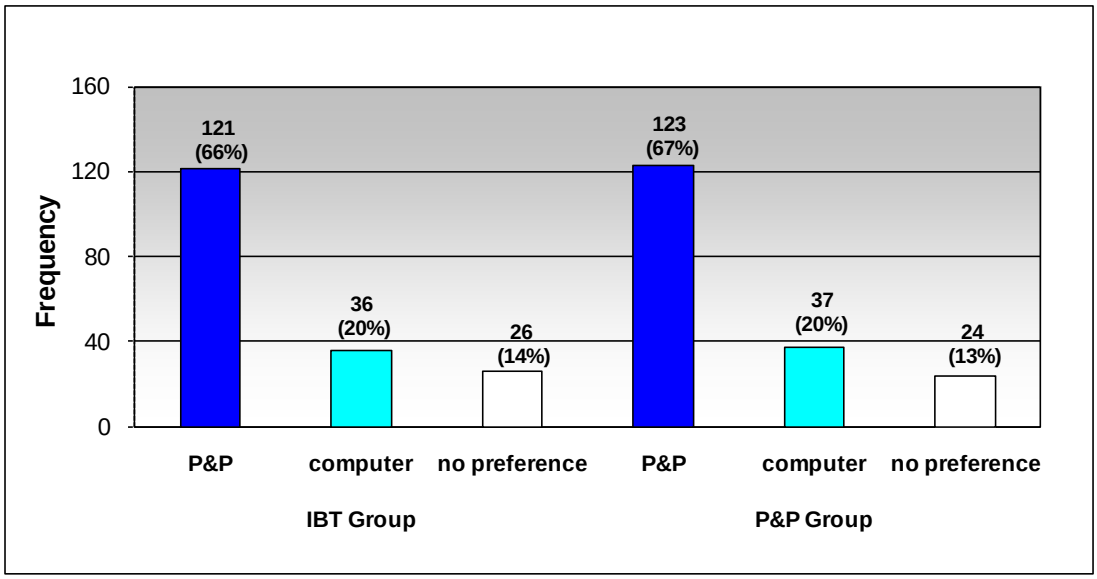
Figure 4. Overall Score on the Operational Test, by Group and Gender



Examinee Preferences

The examinees were asked about their preferred test modality ("Would you prefer to take a test: on computer; on paper; no preference"). The results were almost identical in both groups (Figure 5), in favor of the P&P mode. This outcome might be attributed to the fact that the PET is currently administered as P&P: the examinees prefer the modality with which they are already familiar and which they perceive as the standard. This hypothesis was confirmed in another study, where the examinees preferred the IBT when it was the operational test (Gafni, Blum, & Baumer, 2009).

Figure 5. Examinees' Preferred Test Modality, by Group



Summary and Conclusions

The study was designed to compare achievement scores on an Internet-based test with those on a paper-and-pencil version of the same test. An identical PET form was administered to two groups—one was given an IBT and the other a P&P version. The administration of the IBT went fairly smoothly with no major issues. A few tests were interrupted due to a network problem, but were immediately continued using a recovery procedure. Scores in the two modalities (IBT and P&P) were compared, with the operational test score serving as covariate. The results suggested that the modality of administration does not affect test performance, and that concerns that the IBT might be more difficult than the P&P test are unfounded. Analysis of scores also revealed that computer familiarity does not affect performance on the IBT, and that the IBT does not pose a disadvantage to examinees of any gender. The results support administration of the test in both modalities simultaneously.

Limitations

The experimental cohort was comparable to the overall PET population in age and male-female ratio, but there was a difference in test performance. The examinees who participated in the experiment achieved higher operational scores than the general PET population (an average overall score of 596 vs. 556). It is, therefore, not clear to what degree the results obtained in this study can be generalized to the overall PET population, and in particular to lower-ability examinees. It is also important to emphasize that the results hold only for the item types currently used in PET. They might not generalize, for example, to tests with longer reading passages.

References

Bicanich, E., Hardwicke, S. B., Slivinski, T., & Kapes, J. T. (1997). Internet-based training: A vision or reality? *Technological Horizons in Education, 24*, 61-65

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*, 191-205

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

Gafni, N., Blum N., & Baumer, M. (2009). The use of internet based admissions test (MEMAD) to preparatory schools (In Hebrew). A paper presented at the Annual Meeting of the Israeli Psychometric Association, February 9, Jerusalem, Israel

Huff, K.C. (2006). *The effects of mode of administration on timed cognitive ability tests*. 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. Available from http://io.psych.uiuc.edu/SIOP2007/Huff%20SIOP%202007.pdf

Paek, P. (2005). *Recent trends in comparability studies* (PEM Research Report 05-05). Available from http://www.pearsonedmeasurement.com/downloads/research/RR_05_05.pdf

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

Wang, S., Jiao, H., Young, M. J., & Brooks, T. (2008). Comparability of computer-based and paper-and-pencil testing in K–12 reading assessments. *Educational and Psychological Measurement, 68(1)*, 5-24