

Computer-Based & Paper-Pencil Test Comparability Studies

In some testing applications, Computer-Based Test (CBT) delivery is gaining popularity over the traditional Paper-Pencil-Test (PPT) delivery due to the several potential advantages that it offers, such as immediate scoring and reporting of results, more flexible test scheduling, the opportunity to include innovative item formats that are made possible by the use of technology, and reduced costs of test production, administration, and scoring. In most applications, test items on PPTs and CBTs are identical in terms of content and delivery sequence. The primary difference is the test delivery mode. During the last two decades, numerous comparability studies have been conducted on a variety of educational and psychological tests to examine whether test presentation mode affects examinee performance.

The several articles that have summarized or reviewed previous research on this topic are not conclusive about the comparability of PPTs and CBTs. Mazzeo and Harvey (1988) provided one of the earliest reviews, which included some 30 comparability studies about a range of tests such as ones focusing on intelligence, aptitude, personality, and achievement. Their review revealed mixed evidence concerning the comparability of CBTs and PPTs. They found test mode seemed to have no effect on power tests, but a considerable effect on speeded tests. Mead and Drasgow (1993), who used meta-analysis to examine the mode effect on timed power tests and speeded tests, arrived at a similar conclusion. However, in a meta-analysis of ability measure tests performed by Kim (1999), CBTs and PPTs were found to have comparable average scores.

More recently, with the growing interest in CBTs in K-12 education, a number of comparability studies have been conducted focusing on these applications. The Texas Education Agency (TEA) issued a technical report (2008) that reviewed comparability studies across different content areas (Mathematics, English Language Arts including Reading and Writing, Science and Social Studies) in K-12 tests. In each content area, they found discrepancies between the conclusions of some empirical studies—some studies indicate that a CBT is more difficult than a PPT or vice versa (e.g. Choi & Tinkler, 2002) while some studies show that CBTs and PPTs are comparable (e.g., Kim & Hyunh, 2007). A similar pattern was observed by Paek (2005), although she concluded that “in general, computer and paper versions of traditional multiple-choice tests are comparable across grades and academic content” (p.17).

“The results indicate that administration has no statistically significant effect on student math or reading achievement scores.”

This trend toward CBT and PPT comparability has also been echoed in studies that use meta-analysis to examine the mode effect for K-12 populations. Wang, Jiao, Young, Brooks, and Olson (2007, 2008) conducted two meta-analysis studies on the K-12 student math and reading achievements, respectively. The results indicate that administration mode has no statistically significant effect on student math or reading achievement

scores. Likewise, Kingston (2009) synthesized 81 comparability studies in K-12 multiple-choice tests performed between 1997 and 2007 and found that the estimated effect size was small across all the studies.

Although the majority of recent comparability studies have indicated that CBT and PPT are comparable across delivery medium, the results are not unanimous. The inconsistency in the findings is not surprising, given that these comparability studies involve a wide range of variations in content areas, participants, data collection designs, and item format.

Many researchers have attempted to provide a rationale for observed medium differences. Leeson (2006) cited a few potential factors related to mode effect from two perspectives: participant and technology. The participant factors include demographic characteristics of examinees such as gender and ethnicity, cognitive processing, examinees' ability, familiarity with computers, and anxiety when interacting with computers. The technological issues involve computer interface legibility (such as screen size and resolution, font characteristic, line length) and user interface (such as scrolling to locate reading, permission for item review and how to present the items—one a time or several a time).

"How to measure computer familiarity with adequate validity is an important issue."

Among the factors related to participant characteristics, computer familiarity was examined by several researchers.

Nevertheless, the results of the effect of computer familiarity on examinees' performance are inconclusive. Some studies indicate that computer familiarity is a significant predictor of examinees' performance, suggesting students who have less experience with computers will score lower on computer-administered tests (e.g. Goldenburg & Pedulla, 2002; Pomplun & Custer, 2005, Pomplun, Ritchie & Custer, 2006, Bennett, Braswell, Oranje, Sandene, Kaplan, & Yan, 2008), while others do not replicate these results (e.g. Mazzeo, Druesne, Raffeld, Checketts, & Muhlstein, 1991; Clariana & Wallace, 2002; Taylor, Kirsch, & Eignor, 1999). One possible reason for this inconsistency is that computer familiarity is defined and measured differently in different studies. For example, Pomplun et al. (2006) used SES (measured by free lunch) as an indicator of computer familiarity while in Taylor et al.'s (1999) study, a self-developed survey instrument was employed. Hence, how to measure computer familiarity with adequate validity is an important issue that needs further research in comparability studies.

Subgroup difference related to demographic attributes, such as gender and ethnicity, is another important factor that may have an impact on mode effect. Studies investigating this issue also result in inconsistent results. Many recent studies suggest no difference in the administration mode for gender and/or ethnicity subgroups (e.g. Bennett et al., 2008; Clariana & Wallace, 2002). However, in an early study of CBT and PPT versions of Graduate Record Examination (GRE), Parshall and Kromrey (1993) found examinees' gender, race, and age were associated with the test mode although the results varied across the three subscales (verbal, quantitative and analytic) of the GRE. Some consistent patterns for ethical/gender subgroup differences have

also been found by Gallagher, Bridgeman, and Cahalan (2000) in the investigation of performance difference on PPT and CBT, although the differences are small. Differences in the examinee groups may be such a great threat to the comparability that the fairness of the test will be questioned, so additional research on the subgroup differences as they relate to mode effect is needed.

A few studies have examined the relationship between administration mode and computer characteristics such as screen size, resolution, and font size. For example, McKee and Levinson (1990) indicate these factors may change the nature of a task so dramatically that items administered in CBTs and PPTs no longer measure the same construct. Additional related studies can be found in Leeson (2006). In addition, the flexibility with which test takers can interact with the computer is another possible explanation for observed performance differences across administration media. Examples of this flexibility include item review, revision, skipping, etc., which are natural features in PPTs but often unavailable in computerized tests. Several studies have explored these features (e.g., Lunz, 1995; Vispoel, 2000) but the findings seem not clear-cut.

“The flexibility with which test takers can interact with the computer is another possible explanation for observed performance differences.”

In addition to the participants and computer interface issues, the

comparability of CBTs and PPTs may also be related to the test characteristics such as speededness. The Mead and Drasgow (1993) meta-analysis revealed that the construct being measured by CBTs and PPTs was not equivalent for speeded test. In an investigation of a speeded reading comprehension placement test (Pomplun, Frey & Becker, 2002), the computerized version of the test yielded higher mean scores than its paper-based counterpart.

The authors hypothesized that this score difference was due to the response procedure—that is, clicking a mouse might advantage an examinee to move quicker through the items than recording answers with pencil and answer sheet. However, Neuman and Baydoun (1998) found no differences between CBT and PPT on a series of speeded clerical tests. Even though small distributional differences across mode were found in their study, they concluded that the comparability of the two versions could be achieved when speeded CBT followed the same administration and response procedures as the corresponding PPT.

In addition, content area might constitute another factor that has impact on the comparability of alternative test versions. In Kingston’s (2009) meta-analysis of K-12 multiple-choice tests, he found that computer administration seemed to provide a small advantage for the English language arts and social studies tests while a math test favored paper test. With the open-ended items, Russell (1999) examined the mode effect on students’ performance in three subject areas: science, math, and language arts. The results indicate that computer groups performed better than the paper group in science, but no significant mode effects were found for language arts and math tests. In fact, content area is often an important factor in determining item format (close- or open-ended

questions), and examinees' response procedures may be intertwined with the administration mode to impact the comparability.

Although the findings from various comparative studies are not consistent, there seems to be a trend that the CBTs are comparable to their PPT counterparts. **With the ever-changing characteristics of computer technology and broader accessibility to computers, the incomparability due to computer unfamiliarity is likely to decrease over time.** In addition, recent surveys indicate that students tested online feel comfortable with taking tests on the computer and tend to prefer it to traditional paper testing (Way, Davis & Fitzpatrick. 2006). These positive trends bode well for the CBT development. However, the comparability between the alternative test versions cannot be taken for granted, and related investigations must be conducted to ensure that the examinees are not treated unfairly due to the testing medium.

-- Hong Wang, University of Pittsburgh
Chingwei David Shin, Pearson

REFERENCES

- Bennett, R.E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment*, 6(9). Retrieved from <http://www.jtla.org>.
- Clariana R., & Wallace P. (2002). Paper-based versus computer-based assessment: key factors associated with the test mode effect. *British Journal of Educational Technology*, 33 (5), 593-602.
- Choi, S. W., & Tinkler, T. (2002). *Evaluating comparability of paper and computer based assessment in a K-12 setting*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.
- Gallagher, A., Bridgeman, B., & Cahalan, C. (2000). *The effect of computer-based tests on racial/ethnic, gender, and language groups* (GRE Board Professional Report No. 96-21P). Princeton, NJ: Education Testing Service.
- Goldberg, A., & Pedulla, J.J. (2002). Performance differences according to test mode and computer familiarity on a practice GRE. *Educational and Psychological Measurement*, 62(6), 1053-1067.
- Kim, D.-H., & Huynh, H. (2007). Comparability of computer and paper-and-pencil versions of Algebra and Biology assessments. *Journal of Technology, Learning, and Assessment*, 6(4). Available from <http://www.jtla.org>.
- Kim, J. P. (1999). *Meta-analysis of equivalence of computerized and P&P tests on ability measures*. Paper presented at the Annual Meeting of the Mid-Western Educational Research Association. Chicago, IL.
- Kingston N. M. (2009). Comparability of computer- and paper-administered multiple-choice tests for K-12 populations: A synthesis. *Applied Measurement in Education*, 22(1), 22-37.
- Leeson H. V. (2006). The mode effect: A literature review of human and

- technological issues in computerized testing. *International Journal of Testing*, 6 (1), 1-24.
- Lunz, M. E., & Bergstrom, B. A. (1994). An empirical study of computerized adaptive test administration conditions. *Journal of Educational Measurement*, 31(3), 251-263.
- Mazzeo, J., Druesne, B., Raffeld, P., Checketts, K., & Muhlstein, A. (1991). *Comparability of computer and paper-and-pencil scores for two CLEP general examinations* (College Board Report No. 91-5). New York: College Entrance Examination Board.
- Mazzeo, J., & Harvey, A. L. (1988). *The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature*. (College Board Report 88-8). New York: College Entrance Examination Board.
- McKee, L. M., & Levinson, E. M. (1990). A review of the computerized version of the Self-Directed Search. *Career Development Quarterly*, 38(4), 325-333.
- Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, 114, 449-458.
- Neuman, G., & Baydoun, R. (1998). Computerization of paper-and-pencil tests: When are they equivalent. *Applied Psychological Measurement*, 22, 71-83.
- 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.
- Parshall, C., & Kromrey, J. D. (1993). *Computer testing versus paper-and-pencil testing: An analysis of examinee characteristics associated with mode effect*. Paper presented at the Annual Meeting of the American Educational Research Association. Atlanta, GA.
- Pomplun, M., Frey, S., & Becker, D.F. (2002). The score equivalence of paper and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement*, 62(2), 337-354.
- Pomplun, M., & Custer, M. (2005). The score comparability of computerized and paper-and-pencil formats for K-3 reading tests. *Journal of Educational Computing Research*, 32(2), 153-166.
- Pomplun M., Ritchie, T., & Custer M. (2006). Factors in paper-and-pencil and computer reading score differences at the primary grades. *Educational Assessment*, 11(2), 127-143.
- Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives*, 7(20). Available at <http://epaa.asu.edu/epaa/v7n20>.
- Taylor, C., Kirsch, I., Eignor, D., & Jamieson, J. (1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning*, 49, 219-274.
- Texas Education Agency. (2008). *A review of literature on the comparability of scores obtained from examinees on computer-based and paper-based tests*. Available from http://ritter.tea.state.tx.us/student.assessment/resources/techdigest/Technical_Reports/2008_literatur

- [e_review_of_comparability_report.pdf](#).
- Vispoel, W. P. (2000). Reviewing and changing answers on computerized fixed-item vocabulary tests. *Educational and Psychological Measurement, 60*, 371-384.
- Wang, S., Jiao, H., Young, M. J., Brooks, T. E., & Olson, J. (2007). A meta-analysis of testing mode effects in Grade K-12 mathematics tests. *Educational and Psychological Measurement, 67*, 219-238.
- Wang, S., Jiao, H., Young, M. J., Brooks, T. E., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K-12 assessment: A meta-analysis of testing mode effects. *Educational and Psychological Measurement, 68*, 5-24.
- Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006). *Score comparability of online and paper administrations of Texas assessment of knowledge and skills*. Paper presented at the Annual Meeting of the National Council on Measurement in Education. San Francisco, CA.