

Comparability of Computerized Adaptive and Paper-Pencil Tests

Hong Wang, University of Pittsburgh
Chingwei David Shin, Pearson

When a traditional Paper-Pencil Test (PPT) is delivered by computer, two types of computerization can be implemented. One is a linear Computer-Based Test (CBT) in which the paper version of the test is presented and administered via computers. In a linear CBT, the items on both versions are identical, in general, and scoring methods and procedures are the same. The change from PPT to CBT, therefore, only involves the change of administration mode. The other type of computerization is the Computerized Adaptive Testing (CAT) in which not only the medium of administration changes from paper to computer but also the test delivery algorithm turns from linear to adaptive. This adaptive testing paradigm allows the test items to be selected and administered so that they are tailored to each test taker's ability. Therefore, in comparability studies, both the administration mode and paradigm effect on examinees' performance should be examined to ensure the comparability of the CAT and its PPT counterpart.

Paradigm Effects

The administration mode effect has been widely examined in the comparison of PPT and linear CBT. Although findings are not conclusive, there seems to be a trend indicating that the two versions are comparable across the administration mode (e.g. Paek, 2005, Wang, Jiao, Young, Brooks, & Olson 2007, 2008). When CAT is compared to its PPT counterpart, the mode effect and paradigm effect are confounded with each other. In order to separate the two effects and examine the paradigm effect, some studies have focused on comparability analysis between the linear CBT and CAT.

The administration mode and paradigm effect on examinees' performance should be examined to ensure the comparability of the CAT and its PPT counterpart.

Schaeffer, Steffen, Smith, Mills, and Durso (1995) conducted such a comparability study between linear CBT and CAT on the three GRE measures. They found that analytic CAT and CBT produced incomparable scores which were in favor of the CAT while both versions were comparable for the other two measures. Vispoel, Rocklin, and

Wang (1994) compared CBT and CAT versions of a vocabulary test in terms of measurement precision and efficiency. The results indicate that CAT was more precise generating in ability estimates, especially with a fixed-test length test, and that CAT required fewer items to be administered in order to reach precision similar to the fixed-item CBT. Similar results were found by Vispoel, Wang, and Bleiler (1997) when they compared two versions of tests in assessing music listening skills. In a meta-analysis of the comparability studies for K-12 reading and math achievement (Wang et al., 2007, 2008), the computer delivery algorithm (whether CAT or CBT) was found to be a statistically significant moderator in prediction of mode effect size. The findings from these studies suggest that the test paradigm is one of the factors resulting in the incomparability between the CAT and PPT.

Evaluation Criteria for Comparability

Wang and Kolen (2001) summarized three general categories of criteria to evaluate the comparability between CAT and PPT: (1) validity, (2) psychometric, and (3) statistical assumption/test administration. These criteria are also applicable to evaluating the comparability between the linear computerized tests and PPT. For a CAT, the

evaluation procedures may become more complicated due to various issues relating to the administration procedures of CAT, such as item parameter estimation, item selection, test scoring, and the stopping rule (Green, Bock, Humphreys, Linn, & Reckase 1984).

The essence of the validity criterion is to examine whether the constructs measured by the alternative test versions are the same. Satisfying this criterion is more challenging for a CAT because administration of items in a tailored manner makes differences in content and construct differences across examinees to be more likely. Several techniques have been developed to assess the dimensionality, but direct assessment of dimensionality in CATs needs further research (Wang & Kolen, 2001).

Another important piece of evidence for the validity criterion concerns the relationship between alternative test versions and external criterion variables (Mead & Drasgow, 1993). Neuman and Baydoun (1998) compared the CBT and PPT versions of a clerical battery test which predicted employees' performance based on supervisors' ratings. They did not find differential predictive validity between the two test versions. Likewise, Pomplun, Frey, and

Becker (2002) used scaled scores to examine the predictive validity of the two versions of a reading placement test and found them to have similar predictive power. Subgroup difference across alternative versions of the same test is another comparability concern for the validity criterion. Comparability of the alternative test versions should be consistent across subgroups. Gender and ethnicity are two commonly used grouping variables that are studied in this context. A study by Schaeffer et al. (1995) did find comparability differences across ethnic groups. Specifically, they found that the difference score between CAT and CBT versions of the GRE Analytical test was larger for Asian American examinees than for the other subgroups.

Concerning the psychometric criterion, comparability is evaluated based on the properties such as the shape of the score distribution, reliability, and conditional standard error of measurement (Wang & Kolen, 2001). In the literature of examining the effect of administration mode, mostly comparing CBT and PPT, psychometric properties have been examined at both test and item levels. The most common criterion at the test level is the mean difference, matched samples comparability analysis, and

propensity scores. At the item level, the comparability criterion involves content analysis, mean differences, IRT-based differences, response distributions, and differential item functioning. Detailed information on this issue can be found in the TEA Technical Report Series (2008). Furthermore, the evaluation can be performed for both raw score and scale score distributions.

Reliability is another psychometric criterion that should be examined to evaluate the comparability.

Reliability is another psychometric criterion that should be examined to evaluate the comparability. Green et al. (1984) recommended two types of reliability for the CAT based on different sources of errors. One is the reliability due to random sampling of items from the item pool, and the other is empirical reliability due to "short-run random variation of the trait being measured or to situational variance in the testing conditions" (p. 353) which can be regarded as a counterpart of classical alternate-form reliability. For tests with a passing score or a cut score, the reliability can be examined in terms of the consistency of classification between the two versions. However, in their comparability study of the

adaptive GRE General Test, Schaeffer et al. (1995) found that even though the reliability estimates of the two versions were similar, the measurement precision might not be identical across the ability levels. Therefore, whether the conditional standard error of measurement is the same between the CAT and PPT should be another criterion for evaluating the comparability. Wang and Kolen (2001) referred to this as a second order equity criterion. For tests with a passing score or cut score, they also mentioned the criterion of equal probabilities of achieving passing scores. That is, given a passing score, the examinees with the same true ability should have the same probabilities of meeting or exceeding the given passing score on both tests.

The assumptions used to establish the comparability of the CAT and PPT should be examined.

In addition, the assumptions used to establish the comparability should be examined. For example, if an IRT model is used to compare the CAT and PPT, IRT assumptions like unidimensionality and local independence need to be checked. Many issues specifically related to CAT administration conditions, such

as an item selection algorithm, an item exposure rate, and stopping rules, might impact its comparability with a PPT. The effects of these factors on the comparability need to be further investigated.

Methods to Achieve Comparability

To establish the comparability between CAT and PPT, different procedures which target different aspects of comparability criteria, have been developed. Equating procedures, which are typically applied to equating the alternative forms of PPT, are employed to achieve comparability in terms of psychometric properties at both test and item levels. Kolen (1999-2000) summarized the most common data collection designs that have been used to establish score comparability between CAT and PPT. One is the random group design in which examinees are randomly assigned to either mode of the test and then scores from each mode are equated (e.g. Segall, 1997). Another is the random group with counterbalancing design as used by Eignor (1993). However, that design is typically not recommended due to potential confounding of medium and presentation order effects. The third design is a variation of equating to the IRT-calibrated item pool. In applications of this procedure, the IRT item parameters

are estimated from the paper-and-pencil administrations of the CAT item bank. One important assumption behind this procedure is that items behave in the same way across PPT and CAT (Kolen, 1999-2000), i.e. there is no mode effect. Lunz and Bergstrom (1995) concluded that by using equating with paper calibrated parameters, score comparability can be achieved in terms of ability estimates and passing rate. However, simulations conducted by Pommerich (2007) suggest that using paper calibrated parameters in a CAT administration tends to result in less reliable scores, although this effect was generally small for longer test lengths.

Other procedures involve examining the construct comparability of CAT and PPT. Content balancing, which is an important procedure, has also been researched (Thomasson, 1997). Wang and Kolen (2001) reviewed some algorithms developed for content balancing and used the Stocking and Swanson (1993) weighted deviation algorithm in a simulation example. They found that that procedure worked well. Furthermore, making the CAT comparable to its PPT counterpart is related to a variety of aspects in the CAT design such as item selection algorithm, exposure control rate, and stopping rules. Research on these factors is also helpful for

achieving comparability. For example, Thompson and Way (2007) compared maximum information item selection and targeted information selection which selects items that best match a series of intermediate information targets. They concluded the targeted information selection produced results that are comparable to a PPT version in terms of psychometric equivalence.

Conclusions

The comparability between the alternative test versions cannot be taken for granted and related investigations have to be done to ensure that the examinees are not treated unfairly due to the testing mode. Particularly, when a computerized adaptive test (CAT) is adopted along with a PPT, both the mode and the paradigm effects, as well as issues related to adaptiveness, should be examined.

References

- Eignor, D. R. (1993). *Deriving comparable scores for computer adaptive and conventional tests: An example using the SAT* (ETS Research Report RR-93-5).
- Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of*

- Educational Measurement*, 21(4), 347-360.
- Kolen, M. J. (1999-2000). Threats to score comparability with applications to performance assessments and computerized adaptive tests. *Educational Assessment*, 6, 73-96.
- Lunz, M. E., & Bergstrom, B. A. (1995). *Equating computerized adaptive certification examinations*. Paper presented at the Annual Meeting of the National Council on Measurement in Education. San Francisco, CA.
- 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.pearsonedmeasurment.com/downloads/research/RR_05_05.pdf.
- Pommerich, M. (2007). The effect of using item parameters calibrated from paper administrations in computer adaptive test administrations. *Journal of Technology, Learning, and Assessment*, 5(7). Retrieved from <http://www.jtla.org>.
- 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.
- Schaeffer, G. A., Steffen, M. Golub-Smith, M. L., Mills, C. N., & Durso, R. (1995). *The introduction and comparability of the computer-adaptive GRE General Test* (Research Rep. No. 95-20). Princeton NJ: Educational Testing Service.
- Segall, D. O. (1997). Equating the CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 181-198). Washington, DC: American Psychological Association.
- Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, 17, 277-292.

- 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_literature_review_of_comparability_report.pdf.
- Thomasson G. L. (1997). *The goal of equity within and between computerized adaptive tests and paper and pencil forms*. Paper presented at the Annual Meeting of the National Council on Measurement in Education. Chicago, IL.
- Thompson, T. & Way, D. (2007). *Investigating CAT designs to achieve comparability with a paper test*. In D. J. Weiss (Ed.). Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing. Retrieved from www.psych.umn.edu/psylabs/CATCentral/.
- Vispoel, W. P., Rocklin, T. R., & Wang, T. (1994). Individual differences and test administration procedures: A comparison of fixed-item, computerized-adaptive, and self-adapted testing. *Applied Measurement in Education, 53*, 53-79.
- Vispoel W. P., Wang T., & Bleiler T. (1997). Computerized adaptive and fixed-item testing of music listening skill: A comparison of efficiency, precision, and concurrent validity. *Journal of Educational Measurement, 34*, 43-63.
- Wang, T., & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement, 38*, 19-49.
- 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.