

## Methods of Comparability Studies for Computerized and Paper-Based Tests

In recent years, tests have begun being administered by computer. The potential benefits of computerized testing include more efficient test administration, flexible scheduling, quicker score reporting, more accurate examinee ability estimation, and expanded content/construct coverage. Although computerized testing is an enticing option, most testing programs find it necessary to offer both computer and paper versions of a test at one time, sometimes basing similar decisions on scores derived from both testing media. In this context, a comparability study must be conducted to determine if scores from the computer and paper versions of the test are equivalent. Professional testing standards (APA, 1986; AERA, APA, & NCME, 1999) and the *Standards and Assessments Peer Review Guidance* (USED, 2007) stress the need to study score comparability across test administration media. The first and most critical step in any comparability study is to collect good data for making score comparisons, and three data collection designs are commonly employed.

*“Although computerized testing is an enticing option, most testing programs find it necessary to offer both computer and paper versions of a test at one time.”*

In a *common person* design the same people take the test both on computer and on paper, typically counter-balancing

administration sequence to mitigate potential practice or fatigue effects. The advantages of a common person design are that a smaller sample of students is needed and that it is very powerful in detecting differences. The disadvantages of a common person design are that examinees are required to test twice, and factors associated with testing twice, such as motivation, may influence test performance.

In a *randomly equivalent groups* design, examinees are randomly assigned to test on computer or on paper. Random assignment is employed to minimize the potential impact of groups exhibiting performance differences due to variables other than testing medium (e.g. proficiency or computer familiarity). The advantages of this design are that examinees only need to test once. Additionally, since the two groups of examinees are the same on all important characteristics, no further manipulation of the groups is necessary. The primary disadvantage of random assignment is that it may be difficult to implement. In order to randomly assign examinees, a list of all examinees needs to be available, and examinees have to be willing to test in either condition.

A third study design is the *quasi-experiment design*. In this design, the performance of existing groups of examinees is compared between testing media. For example, in an educational setting, the existing groups may be classrooms of students. There are several potential designs for conducting a quasi-experimental study. In one design, the existing groups would be randomly assigned to testing conditions. In a weaker version of the design, someone, a teacher perhaps, would choose the testing condition, perhaps based on convenience or preference. In both of these examples, the groups to be compared will likely not be

equivalent at the time of test administration. As a result, techniques are utilized to match students from the two groups, sometimes explicitly (e.g., a one-to-one matching) and sometimes statistically (e.g., an Analysis of Covariance). Consequently, the data analyst creates equivalent samples of examinees, and the performance of these two groups is compared between for testing media. The matching can be based simply on a table of score levels by demographic categories or more complicated matching methods (e.g, propensity score matching, Rosenbaum & Rubin, 1985). A quasi-experimental design poses minimal burden on institutions conducting data collection and could easily be part of the regular testing administration. However, this design requires additional demographic information about each student, and the quality of the study results is dependent on the degree of similarity of the samples created.

*“A quasi-experimental design poses minimal burden on institutions conducting data collection.”*

The second step in a good comparability study is to conduct data analyses to answer the bottom line question: are test scores from the computer form interchangeable with scores from the paper form? The analyses typically compare score distributions, reliability estimates, internal structure, test characteristic curves, test information functions, and achievement level percentages based on the computer and paper versions of the test. If a common person design has been used, correlations can be computed across the administration media. Researchers may also examine the

impact of the administration media on different subgroups.

*“If the results of the comparability study show that the test performance across the two testing modes is comparable, then scores on computer and paper can be used interchangeably.”*

A newer method in comparability studies is the *Matched Samples Comparability Analysis* (MSCA) developed by Pearson (Way, Davis, & Fitzpatrick, 2006; Way, Um, Lin, & McClarty, 2007). This method can be carried out in conjunction with a quasi-experimental design. The MSCA first draws a sample of examinees testing by computer and matches them to examinees taking the paper form (or vice versa) so that the two samples have identical profiles of previous test scores and demographic variables. The two samples are then equated under the assumption of a randomly equivalent group design. The procedure is repeated for some number of replications (e.g., 100 or 500) and the equating results are summarized over the replications. The means of the equated results are compared across modes, and the standard deviations over replications are used to estimate sampling variability.

If results of the comparability study show that the test performance across the two testing modes is comparable, then scores on computer and paper can be used interchangeably. If the results show that the scores are not comparable across the modes, equating must be carried out to adjust for the differences. This would result in two sets of raw score to scale

score tables, one set of tables used to score the paper test version, and the other set of tables used to score the computer test version.

To date, many computerized testing programs implemented in educational settings deliver paper test versions linearly on computer. Computer adaptive testing (CAT) is not common. However, because CAT is able to accurately measure examinee ability using fewer items, there is increasing interest in implementing CAT in a variety of testing contexts. It is more challenging to ensure that scores from CAT and paper-based forms are comparable because, in addition to the mode effects, the adaptive nature of CAT may lead to performance differences across the two forms. Comparability studies between CAT and paper forms are usually conducted in one of two paradigms. In the early stage of building a CAT, computer simulation techniques (Wang & Kolen, 2001) are used to set various design features of the CAT so that the CAT and the paper version would be as similar as possible. At this stage, the comparison is not only focused on psychometric properties but also on content specifications across the administration media. In later stages of the CAT development, real subjects are employed in experimental or quasi-experimental designs to evaluate the comparability of the CAT and the paper version of the test (Eignor, 1993; Schaeffer, Bridgeman, Golub-Smith, Lewis, Potenza, & Steffen, 1998; Schaeffer, Steffen, Golub-Smith, Mills, & Durso, 1995). At this stage, the focus is on the psychometric equivalence of the two forms, and equating may be incorporated to adjust for any significant difference.

-- Lei Wan  
Leslie Keng  
Katie McClarty  
Laurie Davis

## REFERENCES

- American Psychological Association Committee on Professional Standards and Committee on Psychological Tests and Assessments (APA) (1986). *Guidelines for computer-based tests and interpretations*. Washington, DC: Author.
- American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.
- Eignor, D. R. (1993). Deriving comparable scores for computer adaptive and conventional tests: An example using the SAT (ETS RR-93-55). Princeton, NJ: Educational Testing Service.
- Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1), 33-38.
- Schaeffer, G. A., Bridgeman, B., Golub-Smith, M. L., Lewis, C., Potenza, M. T., & Steffen, M. (1998). Comparability of paper-and-pencil and computer adaptive test scores on the GRE general test (ETS RR-98-38). Princeton, NJ: Educational Testing Service.
- Schaeffer, G. A., Steffen, M., Golub-Smith, M. L., Mills, C. & Durso, R. (1995). The introduction and comparability

- of the computer adaptive GRE general test (ETS RR 95-20). Princeton, NJ: Educational Testing Service.
- U.S. Department of Education (2007). Standards and Assessments Peer Review Guidance: Information and Examples for Meeting Requirements of the No Child Left Behind Act of 2001.
- Wang, T., & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement, 38*, 19-49.
- Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006, April). *Score comparability of online and paper administrations of the Texas Assessment of Knowledge and Skills*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.
- Way, W. D., Um, K., Lin, C., & McClarty, K. L. (2007, April). *An evaluation of a matched samples method for assessing the comparability of online and paper test performance*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.