Evaluating the Comparability of Results of Test Variations

In October of 2006, the US Department of Education awarded an Enhanced Assessment Grant to North Carolina, as the lead state for a consortium of states and CCSSO, to support the development and evaluation of methods for determining the comparability of scores from test variations to scores from the general assessments. A research team has been investigating methods for evaluating the comparability of scores from (1) translations; (2) a modified (2%) assessment that incorporated linguistic simplification; (3) alternative formats such as portfolios or non-parallel native language forms; and (4) computer-delivered versions of the paper-and-pencil test. Because the degree to which scores on a modified assessment are comparable to scores on the general test is of interest to the field and will help in understanding comparability issues, the test has been included in this series of studies.] The ultimate goal of the project is to produce and disseminate a handbook of best practice and research-based procedures for evaluating and documenting the comparability of test variations. To date, studies have been completed on scores from computer delivered tests and translations:

Computer delivered tests

Sue Lottridge and her colleagues have analyzed data from an administration of paper-based-test (PBTs) and computer-based versions (CBTs) of the paper-based state tests in Algebra I and English II. The tests were administered to a large sample of students in a state as part of the operational administration, with the sample taking both a PBT and a CBT. Other students took only the PBT. Data from the repeated measures design were analyzed to determine how comparable the scores were on two tests and whether items were functioning differently in the two formats. Propensity score matching (PSM), in which a sample of students who took only the PBTs were selected to match the group of students who took the CBT first, was used to create two comparable groups of students (a between subjects design). The matched groups were compared using the same statistics (as appropriate) as the repeated-measures design to determine whether using PSM, which does not require double-testing students, could approximate the results of the repeated-measures design.

Translations

Steve Sireci and his colleagues have conducted structural analysis (confirmatory factor analysis and multidimensional scaling) and item analysis (evaluation of differential item functioning) of a video read-aloud Spanish version of a grade 3 mathematics test, comparing it to the paper and pencil version and a video read-aloud English version. Techniques for dealing with small sample sizes and for following up on results, such as multiple replications to detect spurious findings, were applied to the data.

These techniques will be further evaluated using additional tests. The analysis of data from studies of clarified language forms and alternative formats is in progress. In addition to reviewing empirical data, researchers are developing judgment-based procedures for evaluating comparability. These techniques will be tried out this fall.

The handbook will include literature reviews on evaluating the comparability of the four types of test variations. Three of the literature reviews are complete, those for translations, computer-delivered tests, and clarified language.

Reports of all study results and a draft of the handbook will be available in late February. For more information or to request a copy of draft reports, contact Phoebe Winter, phoebe_winter@sbcglobal.net.