

Standard Setting and Equating on the New Generation of New York State Assessments

Overview

In the environment of the Learning Standards, there is a considerable explicit structure to base decisions about what students should know and what students should be able to do. This is an advantage to a criterion-based testing system, because it allows for greater precision in standard setting and test development.

Test equating and scaling and standard setting (setting cut scores) are intricately related in this environment, as explained below, as are the issues of validity, which should be the subject of another paper. The starting point is what is called item calibration or parameterization, which means estimating true item difficulty in a way that does not depend on how well-skilled the children are who took the items.

Calibration

New York State uses item response theory (IRT) to estimate item difficulty. This method places item difficulty on the same scale as person ability. The scale is developed so that the higher the person ability is compared to an item's difficulty, the more likely the person is to answer that item correctly. A person at the same point on the scale as an item has a 50 percent chance of answering correctly. People below the item value have less than a 50 percent chance and people above the item value have greater than a 50 percent chance. In this way, any point on the scale describes both an item's difficulty and a person's score.

Items are chosen from field test data to construct the test that best fit this model and that are most fair to all groups of students. This enables the final forms of the tests to be accurately described by the scale of people and items.

Equating

There is a three-stage model for equating the new generation of assessments:

- (a) Items are chosen from small pretests that have the desired level of difficulty and relation to the test as a whole;
- (b) These items are selected into forms of approximately equal difficulty and field tested to assure they survive the difficulty, discriminability, and bias analyses;
- (c) The forms are calibrated as described above through the use of overlapping items of known difficulty and all put on a common scale, so that the score on any form is of equal difficulty to achieve as the same score on any other test form.

The overlapping items are the key to this process. As much as we try to control the different forms for fluctuation in difficulty, these fluctuations exist. Because different groups of

students take these same overlapping questions, we can adjust the scales to control these fluctuations and thereby bring equity from test form to test form.

Standard Setting

New York State uses an item mapping procedure to set cut scores or standards. Depending upon the test, one or more standards may be set. For the purposes of simplifying this description, the determination of the "proficiency" standard is discussed.

Based on the Learning Standards, the State first establishes what a proficient student should know or should be able to do. This first decision is very broad, is based on the evidence detailed in the Learning Standards.

Experts from around the State, representing communities of different types as well as New York's ethnic and gender diversity are chosen to be panelists in standard setting. The group is given the preliminary definitions and asked to study the Learning Standards and the scoring rubrics for open-ended questions, which are also put on the same scale for each scoring point.

The panelists then decide on the attributes of a proficient student. They then take several forms of the test.

Each judge individually, based on the consensus of the attributes, divides the test questions into those a proficient student would answer correctly and those a proficient student would not answer correctly. To facilitate this process, the test questions are given to the judge in a book in which the easiest appear first and the most difficult appear last. The judge marks the place in the book dividing the items.

Because the items have been put on the same scale as the students' scores, this point of division also corresponds to a scale score. The scale scores for each judge are circulated, anonymously, and discussion ensues about why people made different decisions.

Finally, judges are sent back to consider the contents of test questions students would answer correctly or would not answer correctly. On these grounds, operational definitions are given for what a proficient student knows and is able to do.

At this point, data are given on the impact of the cut scores. Panelists are advised that these data are to inform their decisions, which are not made on the basis of how many could pass or fail, but rather what their conception is of what students know and are able to do. That is in making their decisions, judges may have had an idea that the attributes described more or fewer students, and the data provide an opportunity to reconsider the list of attributes, not to adjust their judgments based on too few passing or failing. This iterative process continues until consensus is reached, both within small groups and across groups. When the judges' scores are averaged, the result constitutes the proficiency standard.

Follow-up

Because the standards are set on field test data, some follow-up analysis is required on operational test data. The follow-up analysis is less concerned with the percent passing or failing as it is with the ordering of test items in difficulty. If the order varies from the field test ordering, then the judge's placement of the passing score would need to be adjusted because some items thought to be passable by a proficient student may be more difficult than originally estimated.

2/9/99