

CHAPTER 15: EQUATING

RATIONALE

In order to maintain the same passing standard across different administrations, TEA constructs each of its tests to be of comparable difficulty from administration to administration at the total test level and, in many cases, at the objective level. TEA uses statistical equating to accomplish this. There are essentially three stages in the item and test development process where equating takes place: pre-equating test forms under construction, post-equating operational test forms after administration, and equating field-test items after administration. This equating design ensures that the established standards for performance on the original test forms are maintained on all subsequent test forms.

Because the new TAKS tests were administered for the first time in the spring of 2003, the scale that all future administrations will be equated to was determined at that time (see chapter 12, Scaling) and no pre- or post-equating was performed. However, TAKS field-test items were administered in the spring of 2002 and were equated to a common scale. Initial TAKS development activities are described in Chapter 2. The SDAA and RPTE are ongoing programs that necessitate annual equating. Both the SDAA and the RPTE use the same pre-equating procedure, but they differ in how they are post-equated and in how field-test items are linked to the original scale.

PRE-EQUATING

The pre-equating process is one in which a newly developed test is linked, before it is administered, to a set of items that were used previously on one or more test forms. In this way the difficulty level of the newly developed test can be determined through this link, and the anticipated raw scores that correspond to scale scores at performance standards can be identified. Each new SDAA and RPTE form is constructed from a pool of items that have been equated to either the original form on which the scale was established or to other forms linked to this original anchor form.

TEST CONSTRUCTION AND REVIEW

Using the items available in the item bank, TEA staff and psychometricians from Pearson Educational Measurement construct new forms by selecting items that meet both the content specifications of the test under construction and the targeted difficulty level for the total test and, in many cases, the targeted difficulty for each objective. Since each item in the item bank is on the same scale as the original test forms, direct comparisons of Rasch item difficulties can be made to ascertain whether the test is of similar difficulty to the original form. In addition, passing raw scores can be estimated to maintain consistency in the passing standard on the raw score scale.

The newly constructed test form is then reviewed by TEA to ensure that specifications and difficulty levels have been maintained. Although every item on the test has been previously scrutinized by several committees made up of Texas educators and curriculum experts for its match to test specifications, grade and developmental appropriateness, and bias, TEA reexamines these factors for each item on the new test. The difficulty level of the entire test and for each objective is also evaluated, and items are further examined for their statistical quality and range of difficulties. Staff members also review forms to ensure that a wide variety of content and situations are presented in the test items to confirm that the test measures a broad sampling of student skills within the test objectives, and to minimize “cuing” of an answer based on the

content of another item on the test. Additional reviews are designed to verify that the keyed answer choice is the only correct answer to an item and that the order of answer choices on the test form varies appropriately.

If any of these procedures uncovers an item that is unsatisfactory, that item is replaced with a new item, and the review process begins again. This process for reviewing each newly constructed test form helps to ensure that each test will be of the highest possible quality.

POST-EQUATING

After a test has been administered, the items are calibrated using a proprietary computer program to obtain Rasch difficulty values for the items. This calibration transforms the metric of the item difficulties to have a mean value of zero (on the logit scale) with a standard deviation of one. These difficulties must be transformed, or post-equated, to an existing scale before any direct comparison with previous test forms is allowed. The method of accomplishing this post-equating differs for the SDAA and RPTE.

EQUATING SAMPLES

The samples used for post-equating the SDAA and RPTE include nearly the entire population of test takers each year. This is because the SDAA and RPTE are administered to relatively few students compared to the TAKS.

POST-EQUATING THE SDAA

The SDAA tests are composed of approximately 50% historical linking items and 50% new, field-tested items. The historical linking items serve as common items in a common-item linking design. For the reading and mathematics tests, these common items are used in a procedure outlined by Wright, (1977) to calculate an equating constant to transform the difficulty metric obtained from the current linking item calibration to the same difficulty scale as that established by the original test form. This constant is defined below:

$$t_{ab} = \sum_i^K (d_{ia} - d_{ib}) / K,$$

where t_{ab} is the equating constant, a is the item on current test, b is the item on previous test, d_{ia} is the item i Rasch Difficulty on current test, d_{ib} is the item i Rasch Difficulty on previous test, and K is the number of common link items, and the summation is over all common items.

To ensure that discrepant item difficulty values (that is, those in error because of some unforeseen effect) are not used in the equating, an iterative stability check procedure is used that eliminates unstable items from the set of common link items. Essentially the process works as follows: (1) derive the linking constant, (2) apply the linking constant, (3) examine the difference between the post-equated difficulty estimates and the field-test difficulty estimates, (4) identify the item (from the common link item set) that is furthest from its post-equated value, (5) if the least stable item is discrepant by less than 0.3 logits, then the process terminates; otherwise the item is eliminated from the common link item set and the process goes back to step 1 and is repeated. This iterative stability check ensures that items used as common link items are stable across time.

Once the equating constant is obtained, it is applied to all the item difficulties, transforming them so that they are on the same difficulty scale as the items from the original form. After this transformation, the item difficulties from the current administration of the reading and mathematics tests are directly comparable with item difficulties from the original forms and item difficulties from past administrations (because such equating was also performed on those items).

The standard Rasch model for multiple-choice items cannot be used with the SDAA writing test because it includes a writing prompt that is scored on a scale from 0 to 4. In this case the partial credit model is used to calibrate the items. An anchored calibration is performed using the Winsteps Rasch calibration program (Linacre, 2003) in which the difficulty values of the historical linking items are held fixed while the difficulties of the new, field-tested items are estimated. This method of calibration results in all of the item difficulties being on the same scale as the historical linking items, and hence they are comparable to the original test form.

POST-EQUATING THE RPTE

Several forms of the RPTE are administered during each operational test administration. This is because an embedded field-test design is used to collect field-test data for new RPTE items. All test forms for each grade-level reporting category contain base-test items that contribute to a student's score. Each form also contains ten, typically unique, field-test items, although some field-test items may appear on more than one form.

A proprietary computer program is used to obtain Rasch difficulty values for all base-test items regardless of the test form. This calibration transforms the metric of these item difficulties to have a mean value of zero (on the logit scale) with a standard deviation of one. These difficulties must be transformed to the existing RPTE scale before any direct comparison with previous test forms can be made. This is done by using a common-items equating as described above for SDAA. The item difficulties for the base-test items are equated to the Rasch values that were used for pre-equating. In essence, the equated field-test difficulties of the base-test items define the scale of the first administration of the items, and the second, or operational administration is post-equated to that scale. To ensure that discrepant item difficulties are not used in the equating, only item difficulty estimates within 0.3 logits of one another are used in the equating. This check ensures that the linking of items remains stable across time.

FIELD-TEST ITEMS

In order to replenish the item bank as new tests are created each year, newly developed items must be field-tested and equated onto the scale of the original form. The SDAA and RPTE use different designs to collect data on field-test items. The SDAA uses a separate early spring field-test design, while the RPTE uses an embedded field-test design. Once the field-test items are administered, it is necessary to place their difficulties onto the same scale as the original form of the test in order to enable pre-equating during the test assembly process.

SDAA

The SDAA uses a separate early spring field test design. Newly constructed items that have cleared the review process are assembled into four to five tests forms per subject depending on instructional level. These test forms each include a common set of historical linking items that were originally administered during the spring 2000 field test and later put on the 2001 scale after it was created. The field tests are then spiraled across the state, and all students eligible to take the SDAA are administered a form of the field test.

RPTE

The RPTE uses an embedded field-test design. Once a newly constructed item has cleared the review process and is ready to be field-tested, it is embedded in an operational test booklet among the base-test items. The base-test items are common across all test forms and count toward the individual student's score. For example, there are approximately 34 different forms containing the same base-test items. Each form

contains one field-test reading passage with up to eight field-test items, and two word identification items which vary by form. The field-test items do not count toward an individual student's score. The 34 test forms are then spiraled across the state so that a large representative sample of test takers responds to the field-test items. One to two thousand students respond to each form. This spiraling design provides a diverse sample of student performance on each field-test item. In addition, because students do not know which items are field-test items and which items are base-test items, no differential motivation effects are expected. To control for fatigue and start-up effects, all field-test items are placed in the same item positions on each test form.

EQUATING OF FIELD-TEST ITEMS

Two variants of the common items equating procedure are used for the SDAA and RPTE because of the different designs for collecting data on field-test items. For the SDAA, the historical linking items that are common to each field-test form are used to equate the field-test items to the original test form after the field test is administered. For the RPTE, the base-test items that are common to each form are used to equate the field test items to the original test form after the operational spring administration of the test.

Each test form is calibrated separately, with both the common items and field-test items combined. A Rasch calibration is used and it centers the resulting item difficulties to have a mean of zero and a standard deviation of one. Wright's common items equating procedure, as described above, is then used to transform the field-test items from each form to the same difficulty scale as the common items. Since the scale of the common items (historical linking items for SDAA and post-equated base test items for RPTE) is already equated to the original, or anchor, form, so too are the equated field-test items. Hence, the field-test items from the various forms are on the same item difficulty scale and are directly comparable to the anchor form item difficulties.

DEVELOPMENT PROCEDURE FOR FUTURE FORMS

Once the field-test items are equated onto the appropriate scale, the statistical item bank is updated with the new information. On occasion, the same field-test item will appear on more than one form. Once the calibration and equating of the field-test data are completed, these items will have multiple Rasch item difficulties. The equated item difficulty from the form that was administered to the largest number of students is the one that serves as the equated Rasch item difficulty value in the item bank.

After the item bank is updated, the difficulties of all field-test items are on the appropriate scale. As new tests are constructed and administered, the pre- and post-equating process is repeated.