

CHAPTER 16: EQUATING

Introduction: The Need to Equate

The following chapter provides information regarding the process for equating the Texas Assessment of Knowledge and Skills (TAKS), the State-Developed Alternative Assessment II (SDAA II), and the Reading Proficiency Tests in English (RPTE) to ensure the comparability of passing scores from one administration to the next. The need to perform statistical equating is described by Kolen and Brennan (2004):

The process of equating is used in situations where such alternate forms of a test exist and scores earned on different forms are compared to each other. Even though test developers attempt to construct test forms that are as similar as possible to one another in content and statistical specifications, the forms typically differ somewhat in difficulty. Equating is intended to adjust for these difficulty differences, allowing the forms to be used interchangeably. Equating adjusts for differences in difficulty, not for differences in content. After successful equating, for example, examinees who earn an equated score of, say, 26 on a test form could be considered, on average, to be at the same achievement level as examinees who earn an equated score of 26 on a different test form (p. 3).

The Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999) also outline the need for equating:

Many test uses involve different versions of the same test, which yield scores that can be used interchangeably even though they are based on different sets of items (p. 51).

The process of placing scores from such alternative forms on a common scale is called equating. Equating is analogous to the calibration of different balances so that they indicate the same weight for any given object. However, the equating process for test scores is more complex. It involves small statistical adjustments to account for minor differences in the difficulty and statistical properties of the alternate test forms (p. 51).

Consider the following example. Suppose two different forms of a 50-item test (e.g., Form A and Form B) are administered to the 5,000 sixth graders of a large district in the following way: the test is spiraled so that every other student sitting in a classroom is administered Form A, and the other students are administered Form B. The result is two equivalent groups of 2,500 students taking each form. Now, suppose that after scoring all the tests, the mean raw score on Form A is 32 and the mean raw score on Form B is 34, although the two test forms were constructed to be parallel in content. Since the two groups taking the forms are equivalent, it would be natural to conclude that Form A is 2 items more difficult than Form B. As such, the score of 32 on the more difficult Form A is equivalent to the score of 34 on the easier Form B. Hence, both the 32 on Form A and the 34 on Form B are assigned the same scale score (e.g., 2100); in doing so, the two raw scores have been equated. Both raw scores actually

represent the same achievement, or performance, level. Therefore, a score of 32 on Form A would receive a scale score of 2100, and a score of 34 on Form B would also receive a scale score of 2100. Obviously, the scale scores are comparable even though the raw scores are not (a raw score of 32 on Form A does not represent the same achievement, or performance, level as a raw score of 32 on Form B).

From this example it is evident that the principle behind equating is very simple: equitability. The how to of equating, particularly for every possible raw score on two forms, is not always so mathematically simple, but the basic principle of equitability still drives the process. For a more detailed explanation, see Kolen and Brennan (2004) or Petersen, Kolen, and Hoover (1989).

Rationale

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, where possible, 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:

1. Pre-equating test forms under construction
2. Post-equating operational test forms after administration
3. Equating field-test items after administration

Such an equating design helps to ensure that the established standards of performance on the original test forms are maintained on all subsequent test forms. For TAKS, the established standards of performance were set by the State Board of Education in November of 2002, and the tests were administered for the first time in the spring of 2003; thus, the scale of record was established at that time. All future test forms of these TAKS tests will be equated to this scale, although new TAKS tests (e.g., Grade 8 science) will have scales established in their year of implementation. The TAKS, SDAA II, and RPTE are ongoing programs that necessitate annual equating. (The SDAA II was administered for the first time in 2005. Further details are provided in Chapter 13.) All three tests 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. The equating procedures used for these programs have been presented to and endorsed by the Texas Technical Advisory Committee (TTAC); any planned modifications to the original procedures are first presented to and discussed with the TTAC prior to implementation.

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 appeared previously on one or more test forms. In this way the difficulty level of newly developed tests can be determined through this link, and the

anticipated raw scores that correspond to scale scores at performance standards can be identified. Each new TAKS, SDAA II, 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 base tests linked to this original anchor form.

Using the items available in the item bank (i.e., items previously field-tested to obtain actual student data), 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. In addition, targeted difficulty for each objective is maintained where possible. Since each item in the item bank has been placed on the same scale as the original base test, direct comparisons of 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. Finally, classical item statistics are also reviewed, providing another indicator of constructed test difficulty.

The newly constructed test form is then reviewed by TEA to help 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 re-examines 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 help 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 “cueing” 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 each primary test administration, base items (i.e., non-field-test items) are calibrated using a proprietary computer program (in the case of tests composed of multiple-choice items only) to obtain Rasch difficulty values for the items. In the case of “mixed-model” assessments (those containing both multiple-choice and open-ended/essay items requiring hand-scoring), the calibration is performed using the commercially available software program WINSTEPS (Linacre, 2001). These calibrations force the metric of the item difficulties to have a mean value of zero (on the logit scale) and a standard deviation of one. These difficulties must be transformed, or post-equated, to the existing scale before any direct comparison with

previous test forms is allowed. Some TAKS tests are administered multiple times during an academic year to allow students who did not meet the standard on their first attempt additional opportunities to do so. Since the retest population is not representative of the general population, a pre-equated scoring table is used for all retest administrations

TAKS and RPTE

The post-equating phase of the TAKS and RPTE base tests uses conventional common item procedures whereby the base test Rasch item difficulties are compared with their previous year's values to derive a post-equating constant. Typically, only those items that were field-tested the previous year are included in the common item set. TAKS test construction practices call for test forms to be built from the previous year's field-test items whenever possible. Although it is sometimes necessary to pull some items from prior years, such items are typically not part of the common item set.

The samples used for post-equating the TAKS assessment (multiple-choice tests only; English versions only) are typically in excess of 100,000 students per grade and subject. In addition, regional representation is required as well as representation from Dallas and/or Houston. The raw score distribution is also monitored and the sample is not pulled until it has stabilized. Essentially the entire student population is used in equating tests with open-ended and/or essay scores. The samples used for post-equating RPTE include nearly the entire population of test takers each year, because compared to TAKS, RPTE is administered to relatively few students.

The post-equating constant ($t_{a,b}$) is calculated as the difference in mean Rasch item difficulty of items in the common item set on the baseline (2003) scale versus the 2005 Rasch calibrated scale. The exact procedure is explained in the paragraphs that follow.

Wright (1977) outlines the procedure performed on the common-item set to calculate an equating constant in order 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 as follows:

$$t_{a,b} = \frac{\sum_{i=1}^k (d_{i,a} - d_{i,b})}{k}$$

- where $t_{a,b}$ = Equating Constant
 $d_{i,a}$ = Rasch Difficulty of Item i on Current Test
 $d_{i,b}$ = Rasch Difficulty of Item i on Previous Test
 k = Number of Common Link Items

To ensure that discrepant item difficulty values (i.e., those in error because of factors such as context effects, fatigue, and examinee inattention) are not used in the equating, an iterative stability check procedure and other checks are used to eliminate unstable items from the set of common-link items.

Once the equating constant is obtained, it is applied to all 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 test are directly comparable with the item difficulties from the original form and with the item difficulties from all past administrations of the test (because such equating was also performed on those items). Since, under the Rasch model, both item difficulty and person ability are on the same scale, the resulting scale scores are also comparable from year to year.

The specific TAKS equating procedures involve the following steps:

1. Tests are assembled and evaluated using Rasch-based targets. The resulting tests have pre-equated score conversions, which in some cases are used for operational test administrations. For example in Grades 3, 5, and 11, pre-equated score tables are used for retest forms assembled to give students who have not previously demonstrated a “met standard” of proficiency additional testing opportunities.
2. Data from the test administrations are sampled according to the criteria mentioned above.
3. Students are dropped from the sample if they have not met an attemptedness check of having valid responses to five or more items.
4. Key-check analyses are run and results are reviewed by PEM psychometricians. Key checks are done both for the base test as well as separately by test form (including braille and large print forms) in order to detect discrepancies that may only exist on a single test form.
5. Rasch item calibrations are carried out. To facilitate efficient and accurate calibrations across the many tests, the operational calibrations are preceded by a dry run where the program coding, input files, and output files are tested.
6. A post-equating constant ($t_{a,b}$) is calculated as the difference in mean Rasch item difficulty of items in the common item set on the base form versus their field-tested values. The TAKS and RPTE equating procedures use an iterative post-equating stability check procedure to eliminate from the calculation of the equating constant test items whose Rasch item difficulty calibration differ from the pre-equated value by more than a specified value. Historically, this threshold has been an absolute value of 0.3.
7. The post-equating constant is applied to the base form item parameter estimates and raw-to-scale score conversion tables are produced.

The full equating process (item calibration, post-equating stability check, and final raw-to-scale score conversion tables) is independently replicated for verification by a TEA staff member, an independent contractor, and Pearson Educational Measurement staff using alternative calibration software. Any significant discrepancies among the various replications are reviewed and resolved by PEM.

SDAA II

Because the new SDAA II tests were administered for the first time in spring 2005, the scale that all future administrations will be equated to was determined at that time (see Chapter 13: Scaling) and no post-equating was performed. However, SDAA II field-test items were administered in spring 2004 and were equated to a common scale at that time. After the 2005 test administration, all field-test items used in 2004 were placed onto the 2005 scale by using similar common-item equating procedures as previously described.

Field-Test Equating

In order to replenish the item bank as new tests are created each year, newly developed items must be field-tested and equated to the scale of the original form. TAKS, SDAA II, and RPTE use different designs to collect data on field-test items. RPTE uses an embedded field-test design and SDAA II uses a separate field-test design. TAKS tests that contain only multiple-choice items use an embedded field-test design while TAKS tests containing open-ended or essay items use a separate 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 to be done during the test assembly process.

Three variants of the common-items equating procedure are used for the TAKS, SDAA II, and RPTE tests because of the different field-test designs. For the RPTE and TAKS embedded field tests, 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. For SDAA II, the 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 TAKS tests utilizing a separate field-test design, previously-calibrated items are included in field test forms and used to equate the field-test items to the common scale.

RPTE

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 an individual student's score. For RPTE, there are typically between 30 and 40 different forms containing the same base-test items. Each form contains two field-test reading passages with up to 15 field-test items, which vary by form. The field-test items do not count toward an individual student's score. The field-test forms are then spiraled across the state so that a representative sample of test takers responds to the field-test items. Typically, 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.

Each test form is calibrated separately, with both the base test items and field-test items combined. A Rasch calibration is used, which centers the resulting item difficulties to a mean of zero. 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 is already post-equated to the original form, so too are the equated field-test items. Therefore, the field-test items from the various forms are on the same item difficulty scale and are directly comparable to the original form's item difficulties.

SDAA II

SDAA II uses a separate field-test design. Newly constructed items that have cleared the review process are assembled into four or five forms per subject, depending on the instructional level. These test forms each include a common set of linking items. The field tests are then spiraled across the state, and all students eligible to participate in the SDAA II are administered a single form of the field test. Three separate field-test equating designs are used for SDAA II: one for multiple-choice tests, one for writing tests, and one for Reading 9 and ELA 10 tests.

Multiple-Choice Tests

Each SDAA II field-test form contains embedded anchor items in separate section blocks. Within a subject area and instructional level these anchor items are common across all field-test forms and serve as the basis for a Rasch linking of field test forms together. Unique field-test items are distributed among the field-test forms. The goal of field-test equating is to take all of the newly field-tested items that exist on a local Rasch scale after calibration and find the adjustment (linking constant) that will move them to the baseline (base year's) Rasch scale. Linking of SDAA II multiple-choice field-test forms is implemented using a standard Rasch common item linking design.

Writing Tests

For SDAA II writing tests, an incomplete data matrix approach is used to simultaneously calibrate all field-test forms and move all items onto the baseline scale. An incomplete data matrix for each writing instructional level is created that contains the student responses to the anchor items and the unique field-test items.

Reading 9 and ELA 10 Tests

SDAA II Instructional Level 9 Reading and Instructional Level 10 ELA field tests are constructed using the thematically linked triple literary selection (or triplet) format. Each triplet is field

tested over two forms so that a sufficient number of items may be field-tested. The SDAA II program administers four field-test forms for each of these tests, resulting in two triplets being field tested at a time. The field-test forms equating process for these tests differs from the process used with other SDAA II tests. Due to the unique nature and variability of the special education population, a common items approach is used to equate within the two forms of a triplet. The SDAA II Reading 9 and ELA 10 field-test forms each contain embedded anchor items in two or three separate section blocks. These anchor items serve as the basis for the Rasch linking of field-test forms within triplets. This results in two calibration runs, one for each triplet being field-tested. Then the two forms of the triplet are be anchored to the base scale. All items on the two different triplets will be on the same scale because both triplets are anchored to the base test.

TAKS

TAKS uses both an embedded field-test design (for multiple-choice only tests) and a separate field-test design (for tests containing both multiple-choice and open-ended/essay items). For multiple-choice-only tests, newly constructed items are 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 TAKS there are typically 30 to 60 different forms containing the same base-test items per subject, depending on grade level. Each form also contains eight to ten field-test items. The field-test items do not count toward an individual student's score. The test forms are then spiraled across the state so that a large representative sample of test takers responds to the field-test items. Five to ten 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.

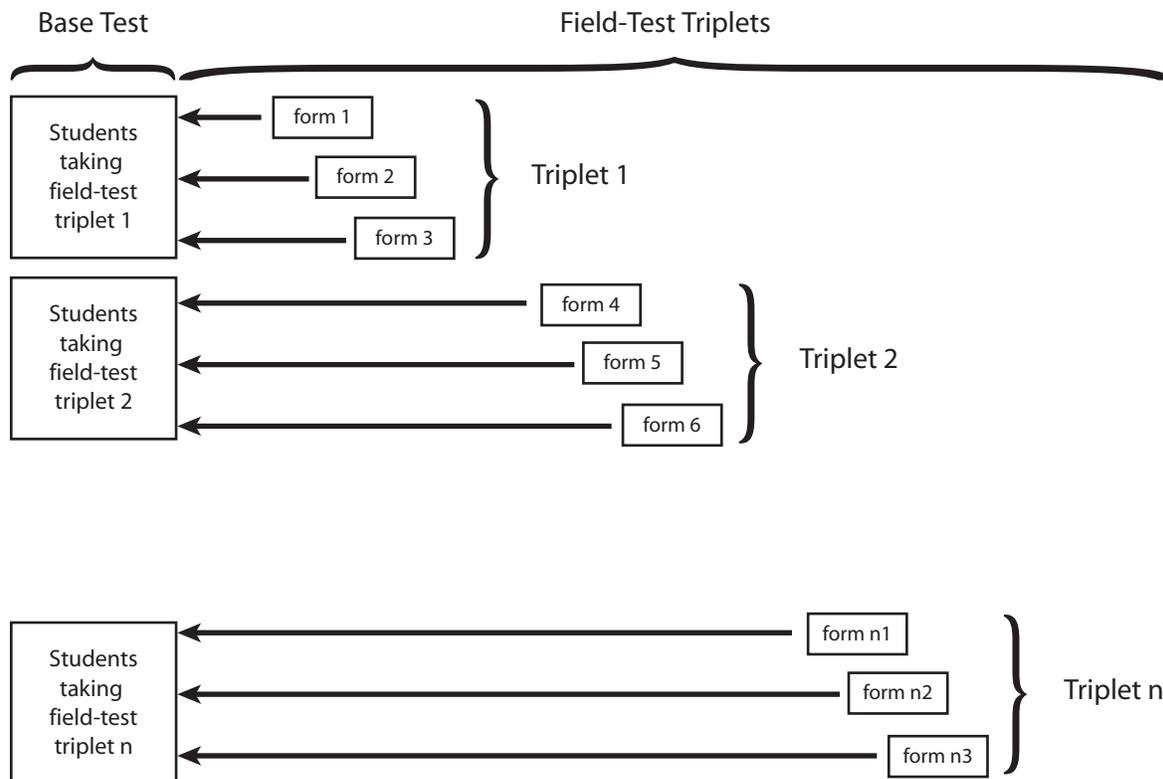
The field-test equating process for TAKS tests that contain only multiple choice items is identical to that described above for RPTE.

The TAKS writing tests and the Grade 9 reading test contain open-ended and/or essay items. A separate field-test design is used for these tests. Newly constructed items that have cleared the review process are assembled into separate test forms; typically between 10 and 30 forms per subject depending on the grade level. The test forms are then distributed across the state so that a large representative sample of test takers responds to each field-test form.

A combined common items and common persons equating design is used for the separate TAKS field tests. The base-test items from the operational test form act as the common items and the same students take both the base test and a field-test form. This allows the field-test items to be equated to the original test form through the operational spring base test. Test forms are calibrated one at a time for Grades 4 and 7 writing. For Grade 9 reading and Grades 10 and 11 ELA, each set of three forms that contain a common triplet is calibrated simultaneously. An anchored calibration is performed using the WINSTEPS Rasch calibration

program (Linacre, 2003), in which the difficulty values of the base-test items are held fixed while the difficulties of the new field-tested items are estimated. This method of calibration results in all item difficulties being on the same scale as the base-test items, and hence they are comparable to the original test form. Intact field-test forms that were field tested in prior years are included in the set of field-test forms each year to ensure that parameter drift does not occur. An example of this anchored calibration with field-test triplets is illustrated in the diagram below.

Diagram 1



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. For the separate TAKS field tests, the responses to these items from all forms on which they appear are combined and calibrated together as part of the simultaneous calibration procedure. For field-test forms that are calibrated separately, these items will have multiple Rasch item difficulties. The equated item difficulty from the form that was administered to the largest number of students 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.

Quality Assurance

During the equating process, many steps are taken to maximize the accuracy of the data collected and the quality of the processes employed. While many of these steps are not strictly related to equating, they do potentially affect the outcome of the equating and are listed in this section.

Pre-Equating Review

Test developers from TEA and PEM select items from a pool of items that have followed a two-year development process. This process includes multiple internal and external reviews, field-testing, and data review (including screening for differential item functioning or potential item bias). During pre-equating test construction, test builders select items to be parallel, in both content and statistical parameters, to the base test upon which the passing standards were established. This helps to ensure that comparable high-quality test questions are selected. Once the test developers are satisfied that the currently constructed test meets all requirements, it is passed on to TEA and PEM staff for additional review.

Statistical Key-Check Procedure

After a significant quantity of test materials has been returned but prior to post-equating, PEM performs a statistical key-check procedure. Through this procedure, statistics are generated by subject, grade, and form and by regular print, large print, and braille. Statistics include omit rates, perfect score rates, p-values, point-biserial correlations, and percent of students choosing each option. These statistics are then reviewed to identify any possible scoring key problems. If items are flagged, content experts review the test questions, and the keys are verified.

Verification of the Post-Equating Process

Once enough test materials have been returned (see “TAKS and RPTE section” in this chapter), data are provided so that the post-equating process may begin. The post-equating process for TAKS is conducted using four different programming routines (two by PEM, one by TEA, and one by an external independent psychometrician). For SDAA II and RPTE, the post-equating process is conducted using two independent programming routines (by PEM). Prior to the actual equating, each psychometrician conducts a check to verify the number of students used in the equating sample, the unique item numbers of the test items, the number of total test items, and the number of options allowed per item. During the equating process, checks are made on the number of common items, the average item difficulty for the common items, the number of items dropped during the stability check, Rasch item difficulties, standard errors for the Rasch item difficulties, theta values, standard errors for theta values, and the equating constant. Quality assurance checks include a review of these same values from the previous year.

Once each of the psychometricians (two PEM, TEA, and an independent contractor) completes his or her equating activities and generates preliminary raw-score-to-scale-score conversion tables, the separate results are then compiled. Compiled results for the item difficulties, the raw-to-scale score conversions, and the equating constants are reviewed for differences. If any differences are detected, the outlying results and procedures are reviewed until consensus is reached. When generating the raw-to-scale score conversion table, psychometricians verify that all raw scores are accounted for, that scale scores increase as raw scores increase, and that the cut points for the performance standards (Met the Standard and Commended Performance) are correctly identified. As a reasonableness check, psychometricians compare results from the current year with results from the past year for the raw score cut points, the number of items on the test, the raw score mean, the raw score standard deviation, the number of students used in the equating dataset, the percent of all students in each performance category, and the percent of students in each performance category for groups (i.e., gender, ethnicity, economically disadvantaged).

After all quality control steps are completed and any differences are resolved, PEM's main analyses (and associated raw-score-to-scale-score conversion tables) are used for the scoring and reporting of student results.

Verification of the Field-Test Equating Process

The field-test equating process is conducted using two different programming routines (by PEM). Once the parties complete their respective field-test equating activity, the separate results are compiled. These compiled results are reviewed for differences. If any differences are detected, the outlying results and procedures are reviewed until consensus is reached. Once any differences are resolved, PEM's main analyses are used for generation of statistical data for uploading into the item bank.

