

CHAPTER 18: EQUATING

Introduction: The Need to Equate

This chapter describes the process for equating the Texas Assessment of Knowledge and Skills (TAKS), the State-Developed Alternative Assessment II (SDAA II), the Reading Proficiency Tests in English (RPTE)/Texas English Language Proficiency Assessment System (TELPAS) reading, and the end-of-course (EOC) assessments. Equating ensures 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) further describe 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 (for example, Form A and Form B) are administered to the 5,000 grade 6 students of a large district. The test forms are 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 randomly equivalent groups of 2,500 students taking each form. After scoring all the tests, the mean raw score on Form A is 32 and the mean raw score on Form B is 34, even though the two test forms were constructed to be parallel in content (i.e., measure the same content in the same manner). Since the two groups taking the forms are assumed to be randomly 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 (for example, 2100); in doing so, the two raw scores have been equated. Both raw scores 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 equated scale scores are comparable even though the raw scores are not (i.e., 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 2002, and the tests were administered for the first time in the spring of 2003; thus, the base scale for reporting was established at that time. All future test forms of these TAKS tests will be equated to this scale, although new TAKS tests (for example, grade 8 science) will have scales established in their first year of implementation. TAKS, SDAA II, RPTE, and EOC are ongoing programs that necessitate annual equating. All four programs 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, RPTE, and EOC 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 (that is, items previously field-tested to obtain student data), TEA staff and psychometricians from PEM 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. 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.

TEA then reviews the newly constructed test form to help ensure that specifications and difficulty levels have been maintained. Although every item on the test has been previously scrutinized by several committees composed of Texas educators and curriculum experts for its match to test specifications, grade and developmental appropriateness, and possible bias, TEA re-examines these factors for each item on the new test. TEA evaluates the difficulty level of the entire test and for each objective while further examining the statistical quality and range of difficulty of every item. Staff members 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 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, it is replaced with a new item and the review process begins again. The process for reviewing each newly constructed test form ensures that each test will be of the highest possible quality.

Post-Equating

After each primary test administration, base items (that is, 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). These difficulties must be transformed, or post-equated, to the existing scale before any direct comparison with previous test forms is appropriate. 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. EOC assessments are pre-equated only.

TAKS, SDAA II, and RPTE

The post-equating phase of the TAKS, SDAA II, and RPTE base tests uses conventional common item procedures, whereby the base test Rasch item difficulties are compared with their previous field-tested values to derive a post-equating constant.

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. Both regional representation and representation from Dallas and/or Houston are required. 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 SDAA II, RPTE, and TAKS (Spanish version) include nearly the entire population of test takers each year because, compared to TAKS, these assessments are 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 2007 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

The relationship between the two forms that is estimated by this equating constant is subject to equating error. Equating error occurs for two reasons. Random equating error occurs when the equating relationship is estimated based on a sample rather than the population. This can be mitigated by using larger sample sizes. As noted above, Texas conducts equating using either almost the entire population of students (with SDAA II, RPTE, and TAKS Spanish) or a large representative sample of the population (as with TAKS English multiple choice tests).

A second source of equating error is systematic error. This can occur when the assumptions of the equating design are violated. For example, if student performance on one or more of the

common items used to equate the test forms has changed across time because of factors such as context effects, fatigue, and examinee inattention, this can cause systematic equating error.

To ensure that discrepant item difficulty values (that is, 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 proficiency are on the same scale, the resulting scale scores are also comparable from year to year.

The specific 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, for TAKS assessments 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. Key-check analyses are run and results are reviewed by PEM psychometricians. Key checks are done both for the base test overall as well as separately by test form in order to detect discrepancies that may only exist on a single test form.
4. Rasch item calibrations are calculated. To facilitate efficient and accurate calibrations across the many tests, the operational calibrations are preceded by a practice run where the program coding, input files, and output files are tested.
5. 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, SDAA II, 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 .3.
6. 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 PEM staff using alternative calibration software. Any significant discrepancies among the various replications are reviewed and resolved by PEM.

Field-Test Equating

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, TELPAS reading, and EOC use both embedded and separate field-test designs to collect data on field-test items. English 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. Additionally, TAKS (Spanish version) uses a combination of embedded and separate field-test designs. EOC uses a separate field test in the initial year of field testing with embedded field-testing thereafter. In 2007, TELPAS reading used both embedded field testing and separate field testing to collect data on the grades 2–12 items administered online.

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 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, TELPAS reading, and EOC tests because of the different field-test designs. For the TELPAS reading and TAKS embedded field tests and Algebra I EOC Assessment, the base-test items that are common to each form are used to equate the items to the original test form after the operational spring administration of the test. For the geometry and biology EOC field tests, the linking items that are common to each form are used to equate the field-test items to each other after the is administered. For TAKS tests utilizing a separate field-test design, the base-test is used as an external common item anchor to equate the field-test items to the common scale. More detail about each of these methods is provided in the next sections.

TELPAS Reading and Algebra I EOC Assessment

TELPAS reading and the Algebra I EOC Assessment use 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 along with the base-test items. (Note: the Algebra I EOC Assessment is offered exclusively online.) The base-test items are common across all test forms and count toward an individual student's score. For TELPAS reading, there are typically between 30 to 40 different forms containing the same base-test items. Each form contains 2 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. These forms are then spiraled across the state so that a representative sample of test takers responds to the field-test items. Typically, 1000–2000 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. Similarly, 12 different forms of the Algebra I EOC Assessment were randomly spiraled during the online assessment. Each form contained the identical base-test items with unique embedded field-test items per form, and approximately 2,500 students responded to each form.

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 previously, 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.

Geometry and Biology EOC Assessments

The Geometry and Biology EOC Assessments used a separate field-test design in the initial year. Newly constructed items that had cleared the review process were assembled into ten forms per subject. The field tests were then spiraled across the state, and each student selected to participate in the geometry or biology EOC field tests was administered a single form of the field test.

Each EOC field-test form contained embedded linking items. Within a subject area, these linking items were common across all field-test forms and served as the basis for a Rasch linking of field-test forms together. Unique field-test items were distributed among the field-test forms. The goal of field-test equating is to take all of the newly field-tested items and move them to a common Rasch scale. Linking of EOC multiple-choice field-test forms was implemented using Wright's common-items equating procedure, as described previously.

TAKS

TAKS uses both an embedded field-test design for tests made up only of multiple choice items and a separate field-test design for tests containing both multiple-choice and open-ended/essay items, as well as some Spanish tests. For multiple-choice-only tests, newly constructed items are embedded in an operational test booklet with 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 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.

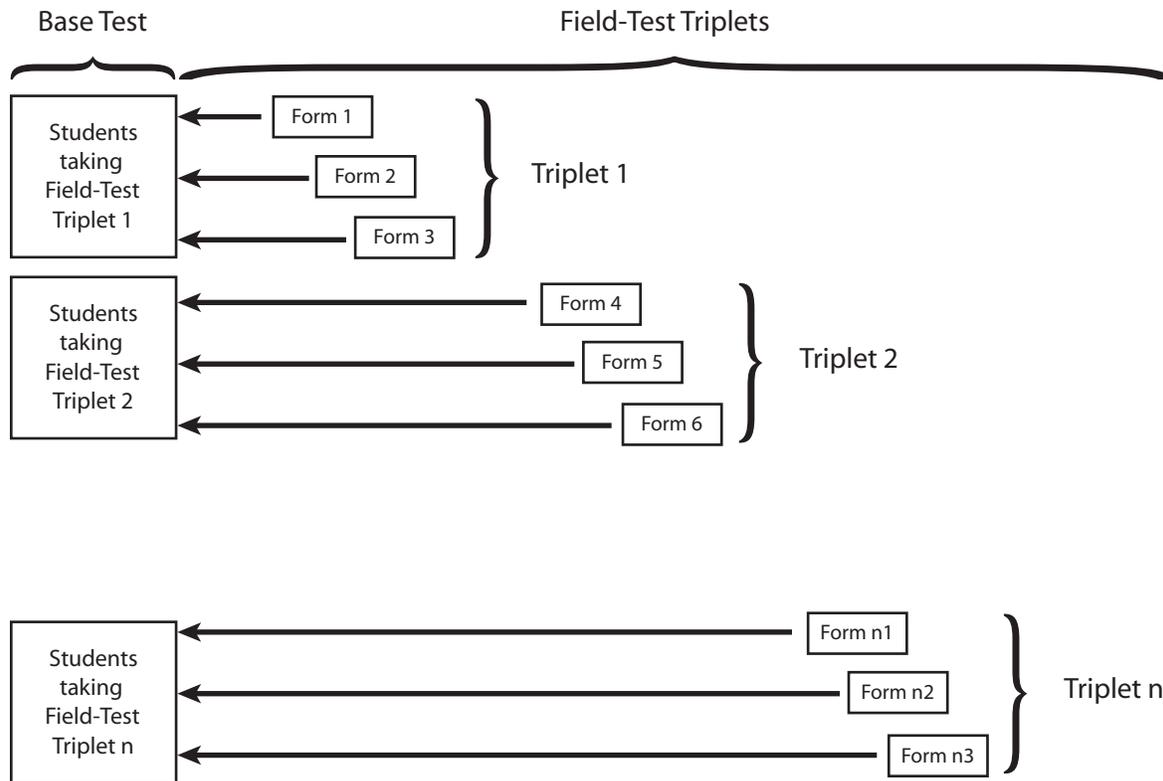
The field-test equating process for TAKS tests that contain only multiple choice items is identical to that described above for TELPAS Reading.

The TAKS writing, grade 9 reading, and grade 10 and exit level English language arts (ELA) tests 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; there are 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.

An external anchor common items 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 process 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 set of thematically linked reading selections (or triplet) is calibrated simultaneously. An anchored calibration is performed using the WINSTEPS Rasch calibration program (Linacre, 2001), 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 Figure 14 on the following page.

Figure 14. Field-Test Triplets



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 described on the appropriate scale. As new tests are constructed and administered, the pre- and post-equating process is repeated.

Comparability Analyses

The issue of comparability between online and paper tests has several facets. When the same test is administered in both delivery modes, studies should be conducted to determine whether the use of the same raw score to scale score table for both online and paper modes is warranted. If mode effects are detected, it may be necessary to use a separate score table for each mode of delivery. The approach used to assess comparability for the TAKS tests was a variation of one outlined by Dorans and Lawrence (1990). Their approach was designed to check the statistical equivalence of nearly identical test forms by evaluating differences in the

raw score to scale score conversion tables. In the context of the TAKS tests, the evaluation was between the online and paper modes. In some equating designs (for example, linear or equipercentile equating), standard errors of equating can be calculated using known formulas. For tests where equating is done using the Rasch model, formulas for calculating standard errors of equating are not available. The bootstrap method (see Kolen & Brennan, 2004, pp. 232–235) is a useful procedure for calculating standard errors of equating using the relevant Rasch model. These standard errors then can be used to evaluate an equating between the online group and a paper group. To accurately examine the comparability of the paper and online versions of a test, the groups of students taking the test in the two modes must be assumed comparable on the skill being measured by the test. If the two groups are not equivalent on the skill being measured, it is not possible to isolate mode differences. There are two ways to achieve group equivalence: one is to randomly assign students to either the paper or online testing process; the other is to match each student participating in the online process to a student in the paper process on the basis of relevant variables such as previous test scores. For TAKS, campuses are allowed to select the mode in which they will test students at the time of test administration. Therefore random assignment is not possible and matching is conducted instead.

The steps used to examine the comparability of the TAKS online and paper tests are outlined below:

1. The paper version of each test is calibrated and equated to the reporting scale using standard procedures for all TAKS assessments. This results in a raw score to scale score conversion table for each paper test. In the case of the July and October retests, a pre-equated raw score to scale score conversion table was used.
2. A random sample of students is drawn with replacement from the online group of students. To estimate sampling error, the sample is the same size as the online group.
3. A sample of students is drawn from the paper group. Each student drawn from the paper administration of the test is matched to a student in the online sample from step 2. The matching variables include gender, ethnicity, and prior or current year test scores.
4. The test items are calibrated separately for the online sample and the paper sample centering on people (that is, the mean ability in each group is set to zero).
5. The theta estimate for each raw score in the online group is used to obtain an estimated raw score using the item parameters from the calibration of the paper test group. These are the equated raw scores for the online group on the scale of the paper test.
6. The equated raw scores for the online group are transformed to scale scores using the raw scores from the sample who took the paper test, corresponding scale scores from step 1, and linear interpolation. These are the scale scores for the online group on the scale of the paper test.

7. Steps 2 through 6 are repeated frequently, typically 100 times (500 for ELA). Note that these bootstrap (see Chapter 13: Sampling for a definition of bootstrap replications) replications incorporate the error in selecting the matched samples as well as the equating error.
8. The average of the equated scale scores at each raw score for the online group over the replications comprises the online scale score table.
9. The standard deviation of online scale score conversions at each raw score represents the conditional bootstrap standard errors of the equating.
10. Raw score points for which the difference between the online and paper scale score conversions exceeds two standard errors of the equating indicate a significant mode effect (Dorans & Lawrence, 1990).

Results of the comparability studies from October 2006, Spring 2007, and July 2007 are included in the 2007 Texas Education Agency Technical Report Series at <http://www.tea.state.tx.us/student.assessment/resources/techdig07/index.html>.

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.

Scoring Table Verification Process for Pre-Equated Tests

The scoring table verification process for pre-equated tests ensures the accuracy of scoring tables prior to any student tests being scored. In this process scoring tables are pulled from the PEM scoring system and compared to scoring tables generated through PEM's test tracking and construction software. Once a Pearson psychometrician has verified that the tables match, these tables are forwarded to the TEA psychometrician for independent verification. If these two reviews concur, then the tables are approved by TEA and PEM personnel and used to

score student tests. This process differs slightly for English language arts tests, in that the scoring tables are generated by a PEM psychometrician, verified, and then loaded to the scoring system, rather than being pulled from the scoring system and then verified.

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. Statistics include omit rates, perfect score rates, p-values, point-biserial correlations, and percentage 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 (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 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 included, scale scores increase as raw scores increase, and that the cut points for the performance standards (such as Met Standard and Commended Performance) are correctly identified. As a check on the reasonableness of the performance standards, 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 percentage of all students in each performance category, and the percentage 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 by PEM using two different programming routines. 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.

