

Chapter 18: Equating

Overview

This chapter describes the process for equating the Texas Assessment of Knowledge and Skills (TAKS), the TAKS–Modified (TAKS–M), the 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. (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

two 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. 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, the Texas Education Agency (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. In addition, TEA uses statistical equating to adjust for any small differences in test form difficulty. There are essentially three stages in the item and test development process where equating takes place:

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

Such an equating design allows the established standards of performance on the original test forms to be maintained on all subsequent test forms. For TAKS, the established standards of performance were set by the State Board of Education (SBOE) in November 2002, and the tests were administered for the first time in spring 2003; thus, the base scale for reporting was established at that time. All subsequent test forms of these TAKS tests would be equated to this scale, although new TAKS tests (for example, grade 8 science) would have scales established in their first year of implementation. TAKS, TAKS–M, TELPAS reading, and EOC are ongoing programs that require annual equating. All 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 Technical Advisory Committee (TAC). Any planned modifications to the original procedures are first presented to and discussed with the TAC 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, TELPAS reading, and EOC form is constructed from a bank 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 form.

Using the items available in the item bank (that is, items previously field-tested to obtain student data), TEA staff and psychometricians from Pearson 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 also are 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.

TAKS–M

The pre-equating process is also done for TAKS–M. However, each new TAKS–M test form is constructed from TAKS–M items that are modified TAKS items. All item modifications are reviewed by Pearson, TEA, and educator committees. For more information on the item modification procedures, see [Chapter 3: Assessments for Students with Disabilities](#).

The TAKS–M test forms are constructed using the modified items. Staff from TEA and Pearson construct test forms by verifying that the content specifications for the test are met. In addition, the field-test information regarding item difficulties for the modified items is reviewed.

Once the new TAKS–M test forms are built, the review process for each form follows the same process described above. Any unsatisfactory items are replaced with a new item and the review process is repeated.

Post-equating

After each primary test administration, base items (that is, items that are not field-test items) are calibrated using a proprietary computer program (in the case of tests composed of multiple-choice items only) to obtain Rasch item difficulty values. 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 passing 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–M and the EOC assessments are pre-equated only.

TAKS

The post-equating phase of the TAKS tests used conventional common item equating procedures, whereby the base/live test Rasch item difficulties were compared with their previous field-tested values to derive a post-equating constant.

The samples used for post-equating the TAKS English multiple-choice-only assessments were typically in excess of 100,000 students per grade and subject. Both regional representation and representation from Dallas and/or Houston were required. The raw score distribution was also monitored, and the sample was not pulled until it had stabilized. Essentially the entire student population was used in equating tests with open-ended and/or essay items. The samples used for post-equating TAKS Spanish assessments included nearly the entire population of test takers each year because, compared to TAKS English versions, these assessments were administered to relatively few students.

The post-equating constant ($t_{a,b}$) was calculated as the difference in mean Rasch item difficulty of 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, a
 $d_{i,b}$ = Rasch Difficulty of Item i on Previous Test, b
 k = Number of Common Link Items

The relationship between the two forms estimated by this equating constant is subject to equating error. Equating error occurs for two reasons. Random equating error can occur 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 conducted equating by using either nearly the entire population of students (TAKS Spanish) or a large representative sample of the population (TAKS English multiple choice-only 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, these 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) were not used in equating, an iterative stability check procedure and other checks were used to eliminate unstable items from the set of common-link items.

Once the equating constant was obtained, it was applied to all item difficulties, transforming them so that they were 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 were 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 were on the same scale, the resulting scale scores were also comparable from year to year.

The specific equating procedures involved the following steps:

1. Tests were assembled and evaluated using Rasch-based objective level and overall targets. The resulting tests had pre-equated score conversions, which in

some cases were used for live test administrations. For example, for TAKS assessments in grades 3, 5, 8, and 11, pre-equated score tables were used for retest forms assembled to give students who had not previously demonstrated a Met Standard of proficiency additional testing opportunities.

2. Data from the test administrations were sampled according to the criteria mentioned above.
3. Key-check analyses were run and results were reviewed by Pearson psychometricians. Key checks were done both for the base test overall as well as separately by test form to detect discrepancies that may only exist on a single test form.
4. Rasch item calibrations were conducted. To facilitate efficient and accurate calibrations across the many tests, the calibrations for live tests were preceded by a practice run where the program coding, input files, and output files were tested.
5. A post-equating constant ($t_{a,b}$) was 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 equating procedures used an iterative post-equating stability check procedure to eliminate from the calculation of the equating constant test items whose Rasch item difficulty calibration differed from the pre-equated value by more than a specified value. Historically, this threshold was an absolute value of .3.
6. The post-equating constant was applied to the base form item parameter estimates and raw to scale score conversion tables were produced.

The full equating process (item calibration, post-equating stability check, and final raw score to scale score conversion tables) was independently replicated for verification by at least three independent parties (one or more of whom was external to Pearson). Any significant discrepancies among the various replications were reviewed and resolved by Pearson.

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 used both embedded and separate field-test designs to collect data on field-test items. TAKS English tests that contained only multiple-choice items and TELPAS reading both used an embedded field-test design, while TAKS tests containing open-ended or essay items and all TAKS–M tests used a separate field-test design. Additionally, TAKS Spanish used a combination of embedded and separate field-test designs. EOC used a separate field test in the initial year of field testing with embedded field testing thereafter.

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 were used for the TAKS, TAKS–M, TELPAS reading, and EOC tests because of the different field-test designs.

- In tests where field-test items were embedded into a base/live test form (such as the TAKS embedded field tests, TELPAS reading, and Algebra I, geometry, and biology EOC assessments), live test items common to each form were used to equate the items to the original test form after the live spring administration of the test.
- In tests where no operational/live test form existed (such as for the chemistry and U.S. history EOC and TAKS-M field tests), a set of linking items common to each form were used to equate the field-test items to each other after the test was administered.
- In TAKS tests that utilized a separate field-test design, a common person equating design was used to link the scale for the field-test items with the scale of the live/base test items. This design was possible because students taking the separate field test also participated in the live administration of the test. The base/live test then was used as an external common item anchor to equate the field-test items to the common scale.

Details about each of the field-test designs are provided in the following sections.

Assessments with Embedded-Only Field-Test Design

TELPAS reading and the Algebra I, geometry, and biology EOC assessments used an embedded field-test design exclusively. Once a newly constructed item had cleared the review process and was ready to be field-tested, it was embedded in an operational test booklet along with the base-test items. (Note: the EOC assessments were offered exclusively online.) The base-test items were common across all test forms and counted toward an individual student's score. For TELPAS reading, there were typically between 30 to 40 different forms containing the same base-test items. Each form contained two field-test reading passages with up to 15 field-test items, which varied by form. The field-test items did not count toward an individual student's score. These forms were then spiraled across the state so that a representative sample of test takers responded to the field-test items. Between 2,000 and 5,000 students responded to each form. This spiraling design provided a diverse sample of student performance on each field-test item. In addition, because students did not know which items were field-test items and which were base-test items, no differential motivation effects were expected. Similarly, 15 different forms of the Algebra I, geometry, and biology EOC assessments 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 2,000–3,300 students responded to each form.

Each test form was calibrated separately, with both the base-test items and field-test items combined. A Rasch calibration was used, which centered the resulting item difficulties to a mean of zero. Wright's common-items equating procedure, as described previously, was 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 had already been equated to the original form, so too were the equated field-test items. Therefore, the field-test items from the various forms were on the same item difficulty scale and were directly comparable to the original form's item difficulties.

Assessments with Stand Alone-Only Field-Test Design

Because the chemistry and U.S. history EOC assessments were offered for the first time in May 2008, these tests used a separate field-test design. Newly constructed items that had cleared the review process were assembled into 11 forms per subject. The field tests then were spiraled across the state, and each student selected to participate in the chemistry or U.S. history 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 placing all field-test forms onto a common Rasch scale. Unique field-test items were distributed among the field-test forms. The goal of field-test equating was 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.

The TAKS–M tests also used a separate field-test design. Modified items that had been through educator review were used in conjunction with linking items to construct between two and four field-test forms per grade and subject. The embedded linking items were common across all field-test forms within a grade and subject area. The same procedure employed for the chemistry and U.S. history EOC assessments was used to put all TAKS–M field-test forms in each grade and subject area onto the same Rasch scale. The field-test form taken by the most students (that is, with the largest n-count) became the base scale and the items from the other field-test forms were moved onto that scale using the linking items. All future TAKS–M tests, both base and field tests, will be equated through the linking items appearing on the field-test forms.

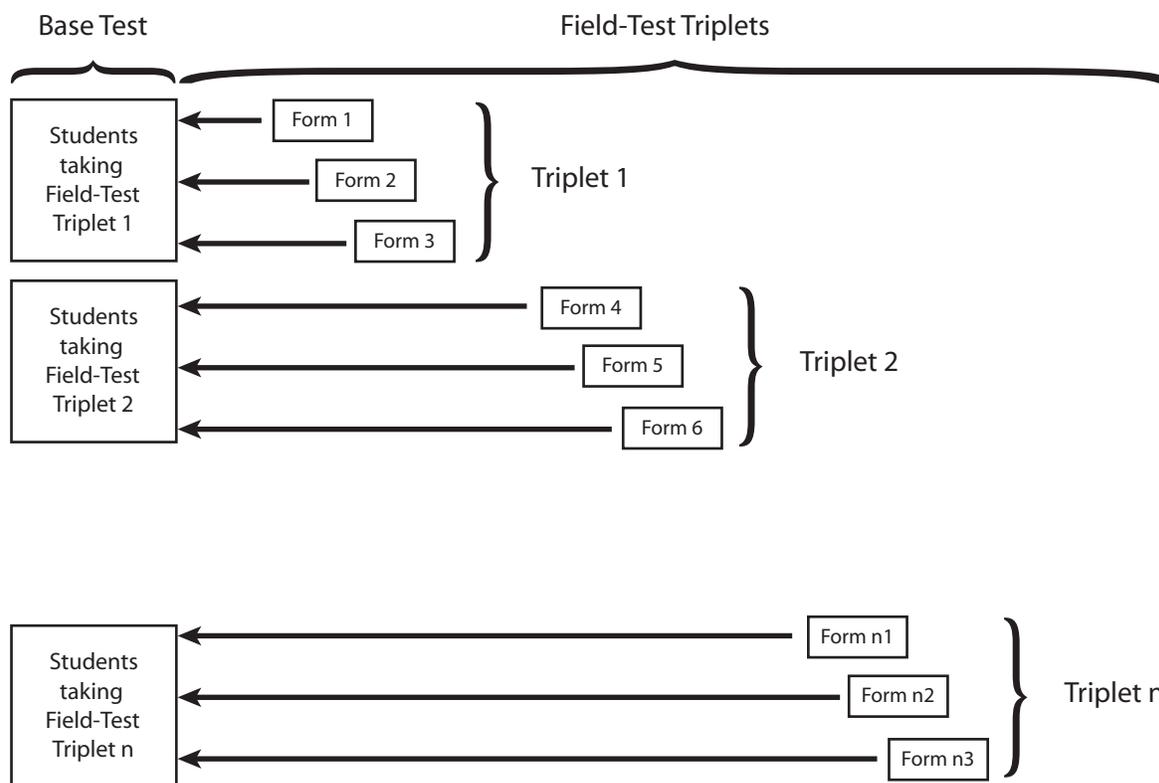
Assessments with Embedded and Stand-Alone Field-Test Designs

TAKS (English and Spanish) used both an embedded field-test design for tests composed only of multiple-choice items 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 were embedded in an operational test booklet with the base-test items. The base-test items were common across all test forms and counted toward the individual student's score. For TAKS English, there were typically 30 to 60 different forms containing the same base-test items per subject, depending on grade level. Each form contained eight to ten embedded field-test items. The field-test items did not count toward an

individual student's score. The test forms were spiraled across the state so a large representative sample of test takers responded to the field-test items. Between 5,000 and 10,000 students responded to each form. This spiraling design provided a diverse sample of student performance on each field-test item. In addition, because students did not know which items were field-test items and which items were base-test items, no differential motivation effects were expected. To control for fatigue and start-up effects, all field-test items were placed in the same item positions on each test form.

The TAKS writing, grade 9 reading, and grade 10 and exit level English language arts (ELA) tests contained open-ended and/or essay items. A separate field-test design was used for these tests. Newly constructed items that had cleared the review process were assembled into separate test forms; there were typically between 10 and 30 forms per subject, depending on the grade level. The test forms were distributed across the state so that a large representative sample of test takers responded to each field-test form. An external anchor common items equating design was used for the separate TAKS field tests. The base-test items from the operational test form acted as the common items, and the same students took both the base test and a field-test form. This process allowed the field-test items to be equated to the original test form through the operational spring base test. Test forms were calibrated one at a time for grades 4 and 7 writing. For grade 9 reading and grades 10 and 11 ELA, forms containing a common set of thematically linked reading selections (or triplet) were calibrated simultaneously. An anchored calibration was performed using the WINSTEPS Rasch calibration program (Linacre, 2001), in which the difficulty values of the base-test items were held fixed while the difficulties of the new field-tested items were estimated. This method of calibration resulted in all item difficulties being on the same scale as the base-test items, and, hence, were comparable to the original test form. Intact field-test forms that were field-tested in prior years were also included in the set of field-test forms each year to account for parameter drift. An example of this anchored calibration with field-test triplets is illustrated in Figure 11.

Figure 11. Field-Test Triplets



Development Procedure for Future Forms

Once the field-test items were equated onto the appropriate scale, the statistical item bank was updated with the new information. On occasion, the same field-test item might appear on more than one form. For the triplet TAKS field tests, the responses to these items from all forms on which they appeared were combined and calibrated together as part of the simultaneous calibration procedure. For field-test forms that were calibrated separately, these items would 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 was updated, the difficulties of all field-test items were described on the appropriate scale. As new tests are constructed and administered each year, the pre- and post-equating process is repeated. For TAKS–M and EOC assessments, once modified items have been field tested and moved onto the base scale, the new test forms are constructed and pre-equated before the operational administration.

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 exit level 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. When differences exceeded a given threshold, the use of separate raw score to scale score conversion tables may have been warranted. This threshold was defined by using the standard error of equating. 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 were allowed to select the mode in which they test students at the time of test administration. Therefore, random assignment was not possible and matching was conducted instead.

For the TAKS program, Texas used a comparability method known as Matched Samples Comparability Analysis (MSCA; Way, Davis, & Fitzpatrick, 2006). MSCA combines the evaluation of bootstrap standard errors of equating with a matching procedure. The detailed steps of the procedure are discussed below. A stratified MSCA approach was used to do the comparability analyses beginning with the October 2007 exit level test administration. Students in the online mode with previous TAKS scores on grade 11 primary tests were matched with those testing on paper who also had grade 11 primary TAKS scores. Students who were first-time testers were matched with their counterparts (the first-time testers) from the paper mode.

In 2007–2008 Texas began offering online administrations of the exit level retests at all four testing opportunities—October, March, April, and July (previously only October and July were offered online). To conduct the comparability analyses using the MSCA approach, there must be a sufficient number of students participating in both the online and paper modes to allow matching to occur. For the April administrations of exit level ELA and social studies, there were not a sufficient number of students participating in the online mode. Therefore, no comparability analysis was conducted. For the April administrations of exit level mathematics and science, as well as all subjects in the October, March, and July administrations, comparability analyses were conducted using the stratified MSCA method described below.

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

1. The paper version of each test was calibrated and equated to the reporting scale using standard pre- or post-equating procedures for all TAKS assessments. This resulted 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. Pre-equated conversions were also used for the March 2008 retest in mathematics, science, and social studies, and the April 2008 ELA retest. Post-equated tables were used for the March ELA and April mathematics, science, and social studies retests.
2. Students were divided into two strata based upon whether or not they had previous exit level test scores.
3. A random sample of students was drawn with replacement from the online group of students. To estimate sampling error, the sample was the same size as the online group.
4. A sample of students was drawn from the paper group. Each student drawn from the paper administration of the test was matched to a student in the online sample from step 2. The matching variables included gender, ethnicity, and prior or current year test scores.
5. Steps 3 and 4 were repeated for both strata: students who had previous exit level test scores and those who did not.
6. The matched data from the two strata were combined.
7. The test items were calibrated separately for the online sample and the paper sample centering on people (that is, the mean ability in each group was set to zero). Note that although samples were drawn by stratum, the calibration was done by combining the two strata.
8. A raw score to raw score equating was conducted. The theta (ability) estimate for each raw score in the online group was used to obtain an estimated raw score using the item parameters from the calibration of the paper test group. These were the equated raw scores for the online group on the scale of the paper test.
9. The equated raw scores for the online group were 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 were the scale scores for the online group on the scale of the paper test.
10. Steps 2 through 7 were repeated 100 times (500 for ELA). Note that these bootstrap replications incorporated the error in selecting the matched samples

as well as the equating error. (See [Chapter 13: Sampling](#) for a definition of bootstrap replications.)

11. The average of the equated scale scores at each raw score for the online group over the replications comprised the online scale score table.
12. The standard deviation of online scale score conversions at each raw score represented the conditional bootstrap standard errors of the equating.
13. Raw score points at which the difference between the online and paper scale score conversions exceeded two standard errors of the equating (statistical significance) and the raw score differences between online and paper that were greater than half a raw score point (practical significance) were flagged.

Results of the comparability studies from October 2007, Spring 2008, and July 2008 are included in the 2008 Texas Education Agency Technical Report Series at <http://www.tea.state.tx.us/student.assessment/resources/techdigest/>.

In preparation for moving the TELPAS reading assessment to an exclusively online assessment, a comparability study for the TELPAS reading was also conducted in spring 2008. Although the methodology was similar (using the MSCA methodology), the primary mode was online—as opposed to the TAKS retests, where the primary mode was paper. In addition, the matching variables were different than those used in the TAKS. There was a mode effect, with the paper mode being easier, across all grade clusters except grade cluster 10–12, where no mode effect was found. Detailed results of the TELPAS comparability studies can also be found by following the link mentioned above.

Quality Assurance

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

Test Construction Review

Test developers from TEA and Pearson selected items from a pool of items that have followed a two-year development process. This process included multiple internal and external reviews, field testing, and data review (including screening for differential item functioning or potential item bias). During test construction, test builders selected items to be parallel, in both content and statistical parameters, to the base test upon which the passing standards were established. This helped to ensure that comparable high-quality test questions were selected. Once the test developers were satisfied that the currently constructed test met all requirements, it was passed on to TEA and Pearson staff for additional review. Items that appeared on TAKS–M tests had been through the test

construction review as a TAKS item. Once the items were modified, committees reviewed the modifications and test forms to verify that the modified item was of the same high quality as its unmodified counterpart.

Scoring Table Verification Process for Pre-equated Tests

The scoring table verification process for pre-equated tests evaluated the accuracy of scoring tables prior to any student tests being scored. In this process scoring tables were pulled from the Pearson scoring system and compared to scoring tables generated through Pearson's test tracking and construction software. Once a Pearson psychometrician verified that the tables match, these tables were forwarded to TEA for approval. Once approved, the tables were used to score student tests. This process differed slightly for the ELA tests, in that the scoring tables were generated by a Pearson psychometrician, verified, and loaded to the scoring system, rather than being pulled from the scoring system and then verified.

Statistical Key-Check Procedure

For both pre- and post-equated test forms, Pearson performed a statistical key-check procedure when a sufficient sample size had been obtained. Through this procedure, statistics were generated by subject, grade, and form. Statistics included omit rates, p -values, point-biserial correlations, and percentage of students choosing each option. These statistics were reviewed to identify any possible scoring key problems. If items were flagged, content experts reviewed the test questions, and the keys were verified.

Verification of the Post-equating Process

Once enough test materials had been returned (see the TAKS section in this chapter), data were provided so that the post-equating process could begin. The post-equating process for TAKS was conducted using at least three different programming routines (two by Pearson and at least one by an external independent psychometrician). Prior to the actual equating, each psychometrician conducted 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 were 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, estimated student proficiencies, standard errors for estimated student proficiency values, and the equating constant. Quality assurance checks included a review of these same values from the previous year.

Once each of the psychometricians (Pearson and an external independent contractor) completed his or her equating activities and generated preliminary raw score to scale score conversion tables, the separate results were compiled. Compiled results for the item difficulties, the raw to scale score conversions, and equating constants were reviewed for differences. If any differences were detected, the outlying results and procedures were

reviewed until consensus was reached. When generating the raw score to scale score conversion table, psychometricians verified that all raw scores were included, scale scores increased as raw scores increased, and that the cut points for the performance standards (such as Commended Performance and Met Standard) were correctly identified. As a check on the reasonableness of the cut scores associated with the performance standards, psychometricians compared results from the current year with results from the past year for the raw score cut points, 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 (e.g., gender, ethnicity, economically disadvantaged).

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

Verification of the Field-Test Equating Process

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

