Chapter 2
Equating Test Scores: Toward Best Practices

Neil J. Dorans, Tim P. Moses, and Daniel R. Eignor

Score equating is essential for any testing program that continually produces new editions of a test and for which the expectation is that scores from these editions have the same meaning over time. Different editions may be built to a common blueprint and designed to measure the same constructs, but they almost invariably differ somewhat in their psychometric properties. If one edition is more difficult than another, examinees would be expected to receive lower scores on the harder form. Score equating seeks to eliminate the effects on scores of these unintended differences in test form difficulty. Score equating is necessary to be fair to examinees and to provide score users with scores that mean the same thing across different editions or forms of the test.

In high-stakes testing programs, in particular, it is extremely important that test equating be done carefully and accurately. The reported scores, even though they represent the endpoint of a large test production, administration, and scoring enterprise, are the most visible part of a testing program. An error in the equating function or score conversion can affect the scores for all examinees, which is both a fairness and a validity concern. The credibility of a testing organization hinges on activities associated with producing, equating, and reporting scores because the reported score is so visible.

This chapter addresses the practical implications of score equating. Section 2.1 introduces test score equating as a special case of the more general class of procedures called score linking procedures. Section 2.2 is concerned with the material that is available before data are collected for equating, the tests, the anchor tests, the old form or reference form raw to scale scaling function, and the number of reference forms available. Section 2.3 lists most common data collection designs that are used in the equating of test scores. In Section 2.4, we list some common observed-score equating functions. Section 2.5 describes common data-processing

N.J. Dorans (✉), T.P. Moses, and D.R. Eignor
Educational Testing Service, Rosedale Rd, Princeton, New Jersey 08541, USA
e-mail: ndorans@ets.org

A.A. von Davier (ed.), *Statistical Models for Test Equating, Scaling, and Linking*,
Statistics for Social and Behavioral Sciences,
DOI 10.1007/978-0-387-98138-3_2, © Springer Science+Business Media, LLC 2011

practices that occur prior to computations of equating functions. In Section 2.6, attention is given to how to evaluate an equating function and postequating activities.

2.1 Linking and Equating: Foundational Aspects

Score linking is used to describe the transformation from a score on one test to a score on another test; score equating is a special type of score linking. Much has been written on score equating and linking. The most complete coverage of the entire field of score equating and score linking, in general, is provided by Kolen and Brennan (2004). Other works include: von Davier, Holland, and Thayer (2004b); Feuer, Holland, Green, Bertenthal, and Hemphill (1999); Koretz, Bertenthal, and Green (1999); Livingston (2004); Holland and Dorans (2006); Flanagan (1951); Angoff (1971); Petersen, Kolen, and Hoover (1989); and several chapters in Dorans, Pommerich, and Holland (2007; see Cook, 2007; Holland, 2007; Kolen, 2007; Petersen, 2007; von Davier, 2007). With all this background material available to the reader, we can be brief and incisive in our treatment of the salient issues, first distinguishing different types of linking and then using these distinctions when describing equating issues in Sections 2.2–2.6.

2.1.1 Classes of Score Linking Methods: Definition of Terms

A *link* between scores on two tests is a transformation from a score on one test to a score on another test. The different types of links have been divided into three basic categories called *predicting*, *scale aligning* and *equating* (Holland & Dorans, 2006). It is essential to understand the differences between these categories because they are often confused in practice. Understanding the distinctions among these categories can prevent violations of professional practice.

2.1.1.1 Predicting

Predicting, the oldest form of score linking, has been confused with equating from the earliest days of psychometrics. The confusion still occurs; Ebretson and Reise (2000) wrote, “In linear equating, for example, scores on one test form are regressed on the other test form” (p. 21). The goal of predicting is to minimize errors of prediction of a score on the dependent or criterion variable from information on other predictor variables. This goal guarantees an asymmetry between what is being predicted and what is used to make the prediction. This asymmetry prevents prediction from meeting one of the fundamental prerequisites of equating, the goal of which is to produce scores that can be used interchangeably.

2.1.1.2 Scale Aligning

The goal of scale aligning is to transform the scores from two different tests onto a common scale. Scaling procedures are about 100 years old. Scale aligning is the second category in the Holland and Dorans (2006) framework. It has many subcategories, including activities such as battery scaling (Kolen, 2004), anchor scaling (Holland & Dorans, 2006), vertical scaling (Harris, 2007; Kolen & Brennan, 2004; Patz & Yao, 2007; Yen, 2007), calibration (Holland & Dorans, 2006), and concordance (Pommerich & Dorans, 2004). Scale aligning and score equating are often confused because the statistical procedures used for scale alignment also can be used to equate tests.

2.1.1.3 Equating

Equating is the strongest form of linking between the scores on two tests. Equating may be viewed as a form of scale aligning in which very strong requirements are placed on the tests being linked. The goal of equating is to produce a linkage between scores on two test forms such that the scores from each test form can be used as if they had come from the same test. Strong requirements must be put on the blueprints for the two tests and on the method used for linking scores in order to establish an effective equating. Among other things, the two tests must measure the same construct at almost the same level of difficulty and with the same degree of reliability. Some practices that can help ensure the achievement of equating requirements are described in Section 2.2.

2.1.1.4 What Constitutes an Equating?

The goal of equating is what distinguishes it from other forms of linking. The goal of score equating is to allow the scores from both tests to be used interchangeably. Experience has shown that the scores and tests that produce the scores must satisfy very strong requirements to achieve this demanding goal of interchangeability. There are five requirements that are widely viewed as necessary for a linking to be an equating (Holland & Dorans, 2006):

1. *The equal-construct requirement*: The two tests should both be measures of the same construct (latent trait, skill, ability).
2. *The equal-reliability requirement*: The two tests should have the same level of reliability.
3. *The symmetry requirement*: The equating transformation for mapping the scores of test Y to those of test X should be the *inverse* of the equating transformation for mapping the scores of X to those of Y.
4. *The equity requirement*: It should be a matter of indifference to an examinee as to which of two tests the examinee actually takes.

5. *The population-invariance requirement*: The equating function used to link the scores of X and Y should be the same regardless of the choice of population or subpopulation from which it is derived.

Both formal and informal statements of subsets of these five requirements appear in a variety of earlier sources (Angoff, 1971; Kolen & Brennan, 2004; Lord, 1950, 1980; Petersen et al., 1989). Dorans and Holland (2000) explicitly discussed these five requirements and indicated various ways in which the five "can be criticized as being vague, irrelevant, impractical, trivial or hopelessly stringent" (p. 283).

2.2 Test Specifications and Score Linking Plans

2.2.1 Test Specifications

Based on the equity condition (Requirement 4 in Section 2.1.2.1), Lord (1980) stated that equating was either unnecessary (because it pertains to test forms intended to be parallel) or impossible (because strictly parallel test forms are not likely to be constructed in practice). Even so, equatings are conducted to ensure fair assessment. Although not much can be done about the impossible aspect, best practices can be used to try to make equating as unnecessary as possible. Poor-quality tests cannot be equated properly for several reasons. For one, they may not measure the same construct. Proper test development increases the likelihood that equating will be unnecessary. Well-defined test specifications are a necessary first step. Test editions need to be constructed to the same blueprint. Under proper assembly rules, old and new forms are equally reliable measures of the same construct that are built to the same set of well-specified content and statistical specifications.

Untried or new test questions need to be pretested, and pretested under conditions that reflect actual test administration conditions. When the test forms include unpretested questions or questions pretested in small samples, there is greater likelihood that test forms will not be identical and that equating adjustments will be necessary. Plans for test development should be based on the availability of high-quality pretested material. Continuous testing often can undermine the quality of tests and test scores by draining pools of pretested items quicker than these items can be replenished.

2.2.2 Anchor Test

Often an anchor test plays a crucial role in the equating process. It is generally considered good practice to construct the anchor test according to the test specifications, so that it is a miniversion of the two tests being equated. That means it should have the same difficulty level and contain the same content as the tests to be equated. In the case where the anchor is internal to the test, context effects become a possible issue. To minimize these effects, internal anchor (or common) items are often placed in the same location within each test.

2.2.3 Score Linking Plans

The raw-to-raw equating is not an end, but the means to an end, namely an appropriate score conversion function. This critical point is sometimes given short shrift in discussions of equating that focus on methods. A multistep process is used to put scores from a new test onto an existing score-reporting scale. Before the new test form is administered, there exists a conversion, $s(y)$, for an old test form that takes raw scores, y, on the old test form Y onto the score-reporting scale. This old-form scaling function, $s(y)$, is independent of the new form. Once data are collected on the new form, data from the new form and the old form are used to compute a raw-to-raw equating function, $e(x)$, that links raw scores x on a new test X to those of an old test form Y.

The final step in the process is to produce a function that converts the equated X raw scores to the score-reporting scale by composing the equating function, $y = e(x)$ with $s(y)$. This puts the raw scores of X onto the reporting scale, $ss(e(x))$. The existing score scale for a test limits the quality of the new-form scaling that can be achieved via the equating of a new form. Equatings can produce poor new-form scalings if the old-form scaling is problematic. Even tests as widely used as the SAT® could have undesirable new-form scalings that were affected by poor alignment of the score scale with the intended uses of the test score. In the case of the SAT, poor score scale alignment in which the average Math score was 50 points higher than the average Verbal score led to widespread misinterpretations about a person's relative verbal and mathematical ability and was rectified by recentering of the SAT scores (Dorans, 2002). Many score scales suffer from poor construction, whereas others discard useful information because of the way the meaning of the scale has changed over time. In other words, the value that best equating practices have for reported scores is sometimes constrained by factors that lie outside the domain of equating.

Testing programs that use best practices have well-designed score equating plans and well-aligned score scales that increase the likelihood that scores on different forms can be used interchangeably. Links to multiple old forms are preferable to a link to a single old form. The SAT plan is an example of a sound linking plan that works well, as demonstrated by Haberman, Guo, Liu, and Dorans (2008). Some testing programs link in a haphazard way as if some magical method of score equating might play the role of *deus ex machina* to set scores straight. Data collection planning, developing linking plans, and maintaining score scales are crucial best practices.

2.3 Data Collection Designs Used in Test Score Equating

To obtain the clearest estimates of test-form difficulty differences, all score equating methods must control for differential ability of the examinee groups employed in the linking process. Data collection procedures should be guided by a concern for obtaining equivalent groups, either directly or indirectly. Often, two different tests that are not strictly parallel are given to two different groups of examinees of unequal ability. Assuming that the samples are large enough to ignore sampling error, differences in the distributions of the resulting scores can be due to one or both of two factors: the relative *difficulty* of the two tests and the relative *ability* of the two groups of examinees on these tests. Differences in difficulty are what test score equating is supposed to take care of; difference in ability of the groups is a confounding factor that needs to be eliminated before the equating process can take place.

In practice, there are two distinct approaches for addressing the separation of test difficulty and group ability differences. The first approach is to use a common population of examinees, so there are no ability differences. The other approach is to use an anchor measure of the construct being assessed by X and Y. When the same examinees take both tests, we achieve direct control over differential examinee ability. In practice, it is more common to use two equivalent samples of examinees from a common population instead of identical examinees. The second approach assumes that performance on a set of common items or an anchor measure can quantify the ability differences between two distinct, but not necessarily equivalent, samples of examinees. The use of an anchor measure can lead to more flexible data collection designs than the use of common examinees. However, the use of anchor measures requires users to make various assumptions that are not needed when the examinees taking the tests are either the same or from equivalent samples. When there are ability differences, the various statistical adjustments for ability differences often produce different results.

In all of our descriptions, we will identify one or more populations of examinees and one or more samples from these populations. We will assume that all samples are random samples, even though in practice this may be only an approximation. More extended discussions of data collection designs are given in Angoff (1971), Petersen et al. (1989), von Davier et al. (2004b), Kolen and Brennan (2004), and Holland and Dorans (2006).

The *single-group design* is the simplest data collection design. In the single-group design, all examinees in a single sample of examinees from population P take both tests. The single-group design can provide accurate equating results with relatively small sample sizes.

In most equating situations, it is impossible to arrange for enough testing time for every examinee to take more than one test. The simplest solution is to have two separate samples take each form of the test. In the *equivalent-groups design*, two equivalent samples are taken from a common population P; one is tested with test form X and the other with test form Y. The equivalent-groups design is often used for equating. Sometimes test booklets are assigned randomly to groups, which is why this design is sometimes called the *random-groups design* (Kolen & Brennan, 2004). A more common situation is to construct the two samples by "spiraling" the test booklets for the two tests. The booklets are alternated in the packaging process so that when the tests are distributed to examinees they are alternated, first form X, then form Y, and then form X again, and so on. Well-executed, spiraled samples are often more "equivalent" (i.e., less different) than random samples because they are approximately *stratified* random samples. The equivalent-groups design is fairly convenient to administer. It does not require that the two tests have any items in common, but this design can be used even when they do have items in common. When samples sizes are large and forms can be reused without security problems, the equivalent-groups design is usually regarded as a good choice because it avoids the issue of possible order effects that can arise in the single-group design where each examinee takes *both* tests.

In order to allow for the possibility of order effects in the single-group design, the sample is sometimes randomly divided in half; in each half-sized subsample the two tests are taken in different orders—form X first and then form Y, or vice versa. The result is the *counterbalanced data collection design*. The counterbalanced design contains both the single-group and equivalent-groups designs. Usually, the counterbalanced design requires a special study for collecting the data.

In *anchor test designs* there are two populations P and Q, with a sample of examinees from P taking test X, and a sample from Q taking test Y. In addition, both samples take an anchor test, A. We follow the terminology of von Davier et al. (2004b) and call this the *nonequivalent groups with anchor test* (NEAT) design. Kolen and Brennan (2004) and others referred to this as the *common-item nonequivalent groups design* or simply the *common item* or the *anchor test* design.

The role of the anchor test is to quantify the differences in ability between samples from P and Q that affect their performance on the two tests to be equated, X and Y. The best kind of an anchor for equating is a test that measures the same construct that X and Y measure. The anchor A is usually a shorter and less reliable test than the tests to be equated.[1]

Formally, the NEAT design contains two single-group designs. The anchor test design is more flexible than the equivalent-groups design because it allows the two samples taking X and Y to be different, or nonequivalent. It is also more efficient than the single-group design because it does not require examinees to take both X and Y.

Although the use of anchor tests may appear to be a minor variation of the previous data collection designs, the use of common items involves new assumptions that are

[1] There are exceptions to this general case. For example, sometimes a multiple-choice anchor test is used to link two versions of an all constructed response test. Here the anchor score is more reliable than the scores to be equated. Although the characteristics of anchor tests are usually not specifically described in the requirements of equating or in summaries of these requirements, in practice linkings that utilize anchors that measure different constructs than the tests to be equated are considered unlikely to meet the requirements of equating.

not necessary in the use of single-group, equivalent-groups, and counterbalanced designs, where common examinees are used; see Sections 2.1–2.3 of Holland and Dorans (2006). Some type of assumption, however, is required in the NEAT design to make up for the fact that *X* is never observed for examinees in *Q* and *Y* is never observed for examinees in *P*. For this reason, there are several distinct methods of scaling and equating tests using the NEAT design. Each of these methods corresponds to making different untestable assumptions about the missing data.

One way to think about the difference between the NEAT design and the single-group, equivalent-groups, and counterbalanced designs is as the difference between observational studies versus experimental designs (Rosenbaum, 1995). The single-group design is like a repeated measures design with a single group and two treatments, the equivalent-groups design is like a randomized comparison with two treatment groups, and the counterbalanced design is like a repeated measures design with a single group and counterbalanced order of treatments. In contrast, the NEAT design is like an observational study with two nonrandomized study groups that are possibly subject to varying amounts of self-selection.

2.3.1 Discussion of Data Collection Designs

Data collection is one of the most important aspects of best practices in equating. Each of the data collection designs mentioned in this section has advantages and disadvantages that make it more or less useful for different situations. For equating, the single-group design requires the smallest sample sizes, and the equivalent-groups design requires the largest sample sizes to achieve the same level of accuracy, as measured by the standard error of equating (see Lord, 1950; Holland & Dorans, 2006). The anchor test (i.e., NEAT) designs require sample sizes somewhere in between those of the single- and equivalent-groups designs, although the sample size requirements depend on how strongly correlated the anchor test is with the two tests to be equated and how similar the two populations are. Higher correlations and smaller differences in proficiency between populations require smaller sample sizes than lower correlations and larger differences in proficiency between populations.

We would argue that the ideal design, in theory and in terms of best practice, is a large-sample, equivalent-groups design with an external anchor test. If the anchor test is administered last, only the anchor test can be affected by possible order effects. A comparison of the distributions of the anchor test in the two (equivalent) samples then allows differential order effects to be identified. If they are substantial, the anchor test can be ignored, leaving a simple equivalent-groups design, where no order effects are possible. If the anchor test is internal to the two tests, then context or order (e.g., item location effects) may arise and need to be dealt with.

An important potential drawback of the equivalent-groups design for score equating is that the test form that has been previously equated has to be given at least twice—once when it was originally equated and then again as the old form in the equating of a new form. In some testing programs, it may be problematic for reasons of test security to reuse operational forms. This leads to consideration of special administrations for purposes of equating. However, if special nonoperational test administrations are arranged to collect equating data using the equivalent-groups design, then the issue of examinee motivation arises, as discussed in Holland and Dorans (2006).

The single-group design requires a smaller sample size to achieve the same level of statistical accuracy as that obtained by an equivalent-groups design with a larger sample, but it brings with it issues of order effects and requires twice as much time to administer both tests. A particular problem with the single-group design is that there is no way to assess for order effects. The counterbalanced design, on the other hand, allows order effects to be estimated. However, if they are large and different for the two tests, then there may be no option but to ignore the data from the tests given second and treat the result as an equivalent-groups design. Because of the greatly reduced sample size, the resulting equivalent-groups design may produce equating results that are less accurate than desired. von Davier et al. (2004b) proposed making a formal statistical decision for the counterbalanced design to assess the order effects.

The anchor test design is the most complex design to execute well, especially if differences in ability between the old- and new-form equating samples are large. Whether an equating test is an external anchor or an internal anchor also has an impact, as do the number of anchor tests and the type of score linking plan employed.

2.3.2 Considerations for External Anchor Tests

It is often advised that the anchor test should be a miniversion of the two tests being equated (Angoff, 1971). Making the anchor test a miniversion of the whole test is sometimes in conflict with the need to disguise an external anchor test to make it look like one of the scored sections of the test. For example, to be a miniversion of the test, the anchor test might need to include a variety of item types, whereas, to mirror a specific section of the test, the anchor test might need to include only a limited number of item types. The phrase *external anchor* usually refers to items that are administered in a separately timed section and that do not count towards the examinee's score. One major advantage of external anchors is that they may serve multiple purposes, for example, equating, pretesting, and tryout of new item types. This is accomplished by spiraling versions of the test with different content in this "variable" section. This process also can be used to improve test security by limiting the exposure of the anchor test to a relatively small proportion of the total group tested.

For best practices, it is important to disguise the external anchor test so that it appears to be just another section of the test. One reason for this is that some examinees may identify the anchor test and, knowing that it does not count towards their final score, skip it or use the time to work on sections that count towards their score (even though they are instructed not to do this). Although this type of behavior may appear to benefit these examinees, because of the way that the anchor test is used in equating, such behavior actually may result in lowering the scores of all examinees if enough of them do it. This counterintuitive result can be explained as follows. The anchor test is used to compare the performance of the current group of examinees on the anchor test to that of a previous group. If a substantial number of the current examinees underperform on the anchor test, this will make them appear less able than they really are. As a consequence, the new test will appear to be somewhat easier than it really is relative to the old test. In score equating, a raw score on an easier test is converted to a lower scaled score than that for the same raw score on a harder test. Therefore, the scores reported on the new test will be lower than they would have been had all examinees performed up to their abilities on the anchor test. As indicated in Section 2.5.1, it is best practice to exclude from the equating analysis any examinees whose anchor test performance is inconsistent with their total test performance.

2.3.3 Considerations for Internal Anchor Tests

Items in an internal anchor test are part of the assessment and count towards each examinee's score. Internal anchor items are usually spread throughout the test. Some external anchors (i.e., items that are left out of or are external to the total score) are administered internally and consequently face some of the issues associated with internal anchors. For the observed-score equating methods described in Section 2.4, where the score on the anchor test plays an important role, it is desirable for the anchor test to be a miniversion of the two tests. This may be more feasible for internal anchor tests than for external anchor tests.

Because the items in an internal anchor test count towards the score, examinees are unlikely to skip them. On the other hand, once anchor test items have been used in the test administration of the old form, the items may become susceptible to security breaches and become known by examinees taking the new form to be equated. For anchor items to be effective, they must maintain their statistical properties across the old and new forms. The primary problems with internal anchor tests are context effects, along with the just-mentioned security breaches. Context effects can occur when common items are administered in different locations (e.g., common Item 10 in one form is Item 20 in the other form) or under different testing conditions (i.e., paper and pencil versus computer delivered), or when they are adjacent to different kinds of items in the two tests. These effects are well documented (Brennan, 1992; Harris & Gao, 2003; Leary & Dorans, 1985). Security breaches are an unfortunate reality for many testing programs, and due diligence is required to prevent them or to recognize them when they occur.

2.3.4 Strengthening the Anchor Test

When there are only small differences in ability between the two samples of examinees used in an anchor test design, all linear equating methods tend to give similar results, as do all nonlinear equating methods. Linear and nonlinear equating methods are discussed in Section 2.4. To the extent that an anchor test design (Section 2.3.4) is almost an equivalent-groups design (Section 2.3.2) with an anchor test, the need for the anchor test is minimized and the quality of equating increases.

When the two samples are very different in ability, the use of the anchor test information becomes critical, because it is the only means for distinguishing differences in ability between the two groups of examinees from differences in difficulty between the two tests that are being equated. The most important properties of the anchor test are its stability over occasions when it is used (mentioned above) and its correlation with the scores on the two tests being equated. The correlation should be as high as possible. Long internal and external anchors are generally better for equating than short ones, as longer anchors are usually more reliable and more highly correlated with the tests.

In many settings, there is only one old form. Some tests are equated to two old forms, sometimes routinely, sometimes in response to a possible equating problem with one of the old forms. The SAT links each new form back to four old forms through four different anchor tests (Haberman et al., 2008). This design reduces the influence of any one old form on the determination of the new-form raw-to-scale conversion. It is desirable to have links to multiple old forms, especially in cases where a large ability difference is anticipated between the groups involved in one of the links.

2.4 Procedures for Equating Scores

Many procedures have been developed over the years for equating tests. Holland and Dorans (2006) considered three factors when attempting to develop a taxonomy of equating methods: (a) common-population versus common-item data collection designs, (b) observed-score versus true-score procedures, and (c) linear versus nonlinear methods.

Because equating is an empirical procedure, it requires a data collection design and a procedure for transforming scores on one test form to scores on another. Linear methods produce a linear function for mapping the scores from X to Y, whereas nonlinear methods allow the transformation to be curved. Observed-score procedures directly transform (or equate) the observed scores on X to those on Y. True-score methods are designed to transform the *true scores* on X to the true scores of Y. True score methods employ a statistical model with an examinee's true score defined as their expected observed test score based on the chosen statistical model. The psychometric models used to date are those of classical test theory and item response theory. Holland and Hoskens (2003) showed how these two psychometric models may be viewed as aspects of the same model.

In this section, we will limit our discussion to observed-score equating methods that use the data collection designs described in Section 2.3. Our focus is on observed-score equating because true scores are unobserved and consequently primarily of theoretical interest only. Consult Holland and Dorans (2006) for more complete treatments of observed-score and true-score procedures.

2.4.1 Observed-Score Procedures for Equating Scores in a Common Population

Three data collection designs in Section 2.3 make use of a common population of examinees: the single-group, the equivalent-groups, and the counterbalanced designs. They all involve a single population P, which is also the target population, T.

We will use a definition of observed-score equating that applies to either linear or nonlinear procedures, depending on whether additional assumptions are satisfied. This allows us to consider both linear and nonlinear observed-score equating methods from a single point of view.

Some notation will be used throughout the rest of this chapter. The *cumulative distribution function* (CDF) of the scores of examinees in the target population, T, on test X is denoted by $F_{\mathrm{T}}(x)$, and it is defined as the proportion of examinees in T who score at or below x on test X. More formally, $F_{\mathrm{T}}(x) = P\{X \leq x \mid T\}$, where $P\ \{.\mid T\}$ denotes the population proportion or probability in T. Similarly, $G_{\mathbf{T}}(y) = P\{Y \leq y \mid T\}$, is the CDF of Y over T. CDFs increase from 0 up to 1 as x (or y) moves from left to right along the horizontal axis in a two-way plot of test score by proportion of examinees. In this notation, x and y may be any real values, not necessarily just the possible scores on the two tests. For distributions of observed scores such as number right or rounded formula scores, the CDFs are step functions that have points of increase only at each possible score (Kolen & Brennan, 2004). In Section 2.4.3 we address the issue of the discreteness of score distributions in detail.

2.4.1.1 The Equipercentile Equating Function

The equipercentile definition of *comparable* scores is that x (a score on test form X) and y (a score on test form Y) are comparable in T if $F_{\mathrm{T}}(x) = G_{\mathrm{T}}(y)$. This means that x and y have the same percentile in the target population, T. When the two CDFs are continuous and strictly increasing, the equation $F_{\mathrm{T}}(x) = G_{\mathrm{T}}(y)$ can always be satisfied and can be solved for y in terms of x. Solving for y leads to the *equipercentile function*, $\mathrm{Equi}_{\mathrm{YT}}(x)$, that links x to y on T, defined by

$$y = \mathrm{Equi}_{\mathrm{YT}}(x) = G_{\mathrm{T}}^{-1}(F_{\mathrm{T}}(x)). \tag{2.1}$$

In Equation 2.1, $y = G_{\mathrm{T}}^{-1}(p)$ denotes the inverse function of $p = G_{\mathrm{T}}(y)$. Note that with discrete data, this relationship does not hold because for most x scores there is no y score for which the two cumulative distributions, one for x and one for y are exactly equal. Hence, with most applications, steps are taken to make the data appear continuous, and different steps can yield different answers.

Note that the target population T is explicit in the definition of $\mathrm{Equi}_{\mathrm{YT}}(x)$ (Dorans & Holland, 2000; Holland & Dorans, 2006; von Davier et al., 2004b). In general, there is nothing to prevent $\mathrm{Equi}_{\mathrm{YT}}(x)$ from varying with the choice of T, thereby violating Requirement 5, the subpopulation-invariance requirement, of Section 2.1.2.1. The equipercentile function is used for equating and other kinds of linking. For equating, we expect the influence of T to be small or negligible, and we call the scores *equivalent*. In other kinds of linking, T can have a substantial effect, and we call the scores *comparable in T*.

2.4.1.2 The Linear Equating Function

If Equation 2.1 is satisfied, then $\mathrm{Equi}_{\mathrm{YT}}(x)$ will transform the distribution of X on T so that it is the same as the distribution of Y on T.

It is sometimes appropriate to assume that the two CDFs, $F_{\mathrm{T}}(x)$ and $G_{\mathrm{T}}(y)$, have the same shape and only differ in their means and standard deviations. To formalize the idea of a common shape, suppose that $F_{\mathrm{T}}(x)$ and $G_{\mathrm{T}}(y)$ both have the form,

$$F_{\mathrm{T}}(x) = K[(x - \mu_{\mathrm{XT}})/\sigma_{\mathrm{XT}}] \text{ and } G_{\mathrm{T}}(y) = K[(y - \mu_{\mathrm{YT}})/\sigma_{\mathrm{YT}}], \qquad (2.2)$$

where K is a CDF with mean zero and standard deviation 1.

When Equation 2.2 holds, $F_{\mathrm{T}}(x)$ and $G_{\mathrm{T}}(y)$ both have the shape determined by K. In this case, it can be shown that the equipercentile function is the *linear function*, $\mathrm{Lin}_{\mathrm{YT}}(x)$, defined as

$$\mathrm{Lin}_{\mathrm{YT}}(x) = \mu_{\mathrm{YT}} + (\sigma_{\mathrm{YT}}/\sigma_{\mathrm{XT}})(x - \mu_{\mathrm{XT}}). \qquad (2.3)$$

The linear function also may be derived as the transformation that gives the X scores the same mean and standard deviation as the Y scores on T. Both of the linear and equipercentile functions satisfy the symmetry requirement (Requirement 3) of Section 2.1.2.1. This means that $\mathrm{Lin}_{\mathrm{XT}}(y) = \mathrm{Lin}_{\mathrm{YT}}{}^{-1}(x)$, and $\mathrm{Equi}_{\mathrm{XT}}(y) = \mathrm{Equi}_{\mathrm{YT}}{}^{-1}(x)$, i.e., equating Y to X is the inverse of the function for equating X to Y. In general, the function $\mathrm{Equi}_{\mathrm{YT}}(x)$ curves around the function $\mathrm{Lin}_{\mathrm{YT}}(x)$.

The linear function requires estimates of the means and standard deviations of X and Y scores over the target population, T. It is easy to obtain these estimates for the single-group and equivalent-groups designs described in Section 2.3 (see Angoff, 1971, or Kolen & Brennan, 2004). It is less straightforward to obtain estimates for the counterbalanced design, as noted by Holland and Dorans (2006).

2.4.2 Procedures for Equating Scores on Complete Tests When Using Common Items

The anchor test design is widely used for equating scores because its use of common items to control for differential examinee ability gives it greater operational flexibility than the approaches using common examinees. Examinees need only take one test, and the samples need not be from a common population. However, this flexibility comes with a price. First of all, the target population is less clear-cut for the NEAT design (see Section 2.3.4)—there are two populations, P and Q, and either could serve as the target population. In addition, the use of the NEAT design requires additional assumptions to allow for the missing data—X is never observed in Q and Y is never observed in P. We use the term *complete test* to indicate that everyone in P sees all items on X and that that everyone in Q sees all items on Y. Our use of the term *missing data* is restricted to data that are missing by design. The assumptions needed to make allowances for the missing data are not easily tested with the observed data, and they are often unstated. We will discuss two distinct sets of assumptions that may be used to justify the observed-score procedures that are commonly used with the NEAT design.

Braun and Holland (1982) proposed that the target population for the NEAT design, or what they called the *synthetic population*, be created by weighting P and Q. They denoted the synthetic population by $T = wP + (1 - w)Q$, by which they meant that distributions (or moments) of X or Y over T are obtained by first computing them over P and Q, separately, and then averaging them with w and $(1 - w)$ to get the distribution over T. There is considerable evidence that the choice of w has a relatively minor influence on equating results; for example, see von Davier et al. (2004b). This insensitivity to w is an example of the population-invariance requirement of Section 2.1.2.1. The definition of the synthetic population forces the user to confront the need to create distributions (or moments) for X on Q and Y in P, where there are no data. In order to do this, assumptions must be made about the missing data.

Equating methods used with the NEAT design can be classified into two major types, according to the way they use the information from the anchor. The first type of missing-data assumption commonly employed is of the *poststratification equating* (PSE) type; the second is of the *chain equating* (CE) type. Each of these types of assumptions asserts that an important distributional property that connects scores on X or Y to scores on the anchor test A is the same for any $T = wP + (1 - w)Q$, in other words, is population invariant. Our emphasis here is on the role of such assumptions for observed-score equating because that is where they are the most completely understood at this time.

The PSE types of assumptions all have the form that the conditional distribution of X given A (or of Y given A) is the same for any synthetic population, $T = wP + (1 - w)Q$. In this approach, we estimate, for each score on the anchor test, the distribution of scores on the new form and on the old form in T. We then use these estimates for equating purposes as if they had actually been observed in T. The PSE type of equating assumes that the relationship that generalizes from each equating sample to the target population is a conditional relationship. In terms of the missing data in the NEAT design, this means that conditional on the anchor test score, A, the distribution of X in Q (where it is missing) is the same as in P (where it is not missing). In the special case of an equivalent-groups design with anchor test, $P = Q$ and the PSE assumptions hold exactly. When P and Q are different, the PSE assumptions are not necessarily valid, but there are no data to contradict them.

The CE assumptions all have the form that a linking function from X to A (or from Y to A) is the same for any synthetic population, $T = wP + (1 - w)Q$. In this approach, we link the scores on the new form to scores on the anchor and then link the scores on the anchor to the scores on the old form. The "chain" formed by these two links connects the scores on the new form to those on the old form. The CE type of equating approach assumes that the linking relationship that generalizes from each equating sample to the target population is an equating relationship. It is less clear for the CE assumptions than for the PSE assumptions what is implied about the missing data in the NEAT design (Kolen & Brennan, 2004, p. 146).

In the special case of an equivalent-groups design with anchor test, $P = Q$ and the CE assumptions hold exactly. In this special situation, the corresponding methods based on either the PSE or the CE assumptions will produce identical results. When P and Q are different, the PSE assumptions and CE assumptions can result in equating functions that are different, and there are no data to allow us to contradict or help us choose between either set of assumptions.

In addition to the PSE and CE types of procedures, classical test theory may be used to derive an additional *linear observed-score* procedure for the NEAT design—the Levine observed-score equating function, $\mathrm{Lev}_{\mathrm{YT}}(\mathrm{x})$ (Kolen & Brennan, 2004). $\mathrm{Lev}_{\mathrm{YT}}(\mathrm{x})$ may be derived from two population-invariance assumptions that are different from those that we have considered so far and that are based on classical test theory.

2.5 Data Processing Practices

Prior to equating, several steps should be taken to improve the quality of the data. These best practices of data processing deal with sample selection, item screening, and continuizing and smoothing score distributions.

2.5.1 Sample Selection

Tests are designed with a target population in mind (defined as T throughout Section 2.4). For example, admissions tests are used to gather standardized information about candidates who plan to enter a college or university. The SAT excludes individuals who are not juniors or seniors in high school from its equating samples because they are not considered members of the target population (Liang, Dorans, & Sinharay, 2009). Consequently, junior high school students, for whom the test was not developed but who take the test, are not included in the equating sample. In addition, it is common practice to exclude individuals who may have taken the anchor test (whether internal or external) at an earlier administration. This is done to remove any potential influence of these individuals on the equating results. Examinees who perform well below chance expectation on the test are sometimes excluded, though many of these examinees already might have been excluded if they were not part of the target group. There is an issue as to whether nonnative speakers of the language in which the test is administered should also be excluded. One study by Liang et al. (2009) suggested this may not be an issue as long as the proportion of nonnative speakers does not change markedly across administrations.

Statistical outlier analysis can be used to identify those examinees whose anchor test performance is substantially different from their performance on the operational test, namely the scores are so different that both scores cannot be plausible indicators of the examinee's ability. Removing these examinees from the equating sample prevents their unlikely performance from having an undue effect on the resulting equating function.

2.5.2 Checking That Anchor Items Act Like Common Items

For both internal anchor (anchor items count towards the total score) and external anchor (items do not count towards the score) tests, the statistical properties of the common items should be evaluated to make sure they have not differentially changed from the one test administration to the other. Differential item functioning methods may be used to compare the performance of the common items with the two test administrations treated as the reference and focal groups, and the total score on the common items as the matching criterion (see Holland & Wainer, 1993, especially Chapter 3). Simple plots of item difficulty values and other statistics also may be used to detect changes in items. Internal common items are susceptible to context effects because they may be embedded within different sets of items in the two tests. Changes in widely held knowledge also may lead to changes in performance on anchor test items. For example, a hard question about a new law on a certification exam may become very easy once the law becomes part of the standard training curriculum. There are many examples of this type of "rapid aging" of test questions.

2.5.3 The Need to Continuize the Discrete Distributions of Scores

The equipercentile function defined in Section 2.5.2 can depend on how $F_{\mathrm{T}}(x)$ and $G_{\mathrm{T}}(y)$ are made continuous or *continuized*. Test scores are typically integers, such as number-right scores or rounded formula-scores. Because of this, the inverse function, required in Equation 2.1 of Section 2.4.1.1, is not well defined—for many values of p, there is no score, y, for which $p = G_T(y)$. This is not due to the *finiteness of real samples*, but rather to the *discreteness of real test scores*. To get around this, three methods of continuization of $F_T(x)$ and $G_T(y)$ are in current use. Holland and Dorans (2006) treated two of these methods, the linear interpolation and kernel smoothing methods, in detail. The linear equating function defined in Equation 2.3 of Section 2.4.1.2 is a third continuization method.

The first two approaches to continuization have two primary differences. First, the use of linear interpolation results in an equipercentile function that is piecewise linear and continuous. Such functions may have "kinks" that practitioners feel need to be smoothed out by a further smoothing, often called postsmoothing (Fairbank, 1987; Kolen & Brennan, 2004). In contrast, kernel smoothing results in equipercentile functions that are completely smooth (i.e., differentiable everywhere) and that do not need further postsmoothing. Second, the equipercentile functions obtained by linear interpolation always map the highest score on test form X into the highest score on test form Y and the same for the lowest scores (unlike kernel smoothing and the linear equating function). While it is sometimes desirable, in some cases the highest score on an easier test should not be mapped onto the highest score of a harder test. For more discussion of this point, see Petersen et al. (1989), Kolen and Brennan (2004), and von Davier et al. (2004b).

2.5.4 *Smoothing*

Irregularities in the score distributions can produce irregularities in the equipercentile equating function that do not generalize to other groups of test takers. Consequently, it is generally considered advisable to smooth the raw-score frequencies, the CDFs, or the equipercentile equating function itself (Holland & Thayer, 1987, 2000; Kolen & Jarjoura, 1987; Kolen & Brennan, 2004; von Davier et al., 2004b). The purpose of this step is to eliminate some of the sampling variability present in the raw-score frequencies, in order to produce smoother CDFs for computation of the equipercentile function.

When presmoothing data, it is important to achieve a balance between a good representation of the original data and smoothness. Smoothness reduces sampling variability, whereas a good representation of the data reduces the possibility of bias. For example, if a log-linear model is used, it needs to preserve the most important features of the data, such as means, variances and skewnesses, and any other special features. The more parameters employed in the smoothing, the better the model will represent the original data, but the less smooth the fitted model becomes.

2.6 Evaluating an Equating Function

Quality and similarity of tests to be equated, choice of data collection design, characteristics of anchor test in relation to the total tests, sample sizes and examinee characteristics, screening items, and tests for outliers and choice of analyses all involve best practices that contribute to a successful equating. First, we summarize best practices. Then we discuss challenges to the production of a quality equating and close by discussing directions for additional research.

2.6.1 Best Practices

The amount of data collected (sample size) has a substantial effect on the usefulness of the resulting equating. Because it is desirable for the statistical uncertainty associated with test equating to be much smaller than the other sources of variation in test results, it is important that the results of test equating be based on samples that are large enough to insure this.

Ideally, the data should come from a large representative sample of motivated examinees that is divided in half either randomly or randomly within strata to achieve equivalent groups. Each half is administered either the new form or the old form of a test. If timing is generous and examinees are up to the task of taking both tests, a counterbalanced design could be employed in which each half of the sample is broken into halves again and then both the new and old forms are administered to examinees in a counterbalanced order.

When an anchor test is used, the items are evaluated via differential item functioning procedures to see if they are performing in the same way in both the old- and new-form samples. The anchor test needs to be highly correlated with the total tests. All items on both tests are evaluated to see if they are performing as expected.

It is valuable to equate with several different models, including both linear and equipercentile models. In the equivalent-groups case, the equipercentile method can be compared to the linear method using the standard error of equating, which describes sampling error, and the difference that matters, an effect size that can be used to assess whether differences in equating functions have practical significance or is an artifact of rounding. Holland and Dorans (2006) described the difference that matters, the standard error of equating, and the standard error of equating difference. If the departures from linearity are less than the difference that matters and less than what would be expected due to sampling error, the linear model is often chosen on the grounds of parsimony because it was not sufficiently falsified by the data. Otherwise, the more general, less falsifiable, equipercentile model is selected. Rijmen, Qu, and von Davier (Chapter 19, this volume) provide another approach to choosing among linking functions.

In the anchor test case, it is particularly important to employ multiple models, as each model rests on different sets of assumptions. The search for a single best model that could be employed universally would be unwise data analysis (Tukey, 1963).

An equating should be checked for its reasonableness. How do we determine reasonableness? We compare the raw-to-scale conversion for the new form to those that have been obtained in the past. Is the new form conversion an outlier? Is it consistent with other difficulty information that may be available for that form and other forms that have been administered in the past? Is the performance of the group taking the new form consistent with the performance of other groups that are expected to be similar to it? For example, in testing programs with large volumes and relatively stable populations, it is reasonable to expect that the new-form sample will have a similar scale score distribution to that obtained at the same time the year before. If the test is used to certify mastery, then the pass rates should be relatively stable from year to year, though not necessarily across administrations within a year.

2.6.2 Challenges to Producing High-Quality Equatings

Large, representative, motivated samples that result from a random assignment of test forms to examinees are not always attainable. Reliability is not always as high as desired. Anchor tests may not be very reliable, especially internal anchors with few items. Anchors, especially external anchors, are not always highly related to the tests being equated. Tests are not always appropriate for the group that takes them. These issues often arise when best design and data collection practices are not followed.

2.6.2.1 Data Collection Design Issues

Some threats to sound equating are related to the choice of data collection design. The NEAT design is often used because of the greater flexibility it provides. Statistical procedures are needed to adjust for ability differences between groups when the NEAT design is used. Assumptions need to be made in order to make these adjustments. The assumptions may be flawed.

2.6.2.2 Psychometric Properties of the Tests and Anchors

Characteristics of the test to be equated affect the quality of equating. Pretesting of untried items prior to their operational use produces higher quality exams. The absence of pretesting may result in tests with fewer scorable items than planned. The resulting shorter, less reliable tests are harder to equate because a greater portion of score variability is noise and the resultant equating functions are less stable. More importantly, tests made up of unpretested items can turn out to be different in content and difficulty from the tests to which they are to be equated; these factors increase the difficulty of equating.

The role of the anchor test is to provide a common score that can be used to adjust for group ability differences before adjusting for test difficulty differences via equating. Scores from short anchor tests tend to have inadequate reliabilities and consequently less than desirable correlations with the test scores. Low correlations also may result when the content of the anchor test differs from the test. Context effects can affect the comparability of anchor items. Anchors that are too hard or too easy for the target population produce skewed score distributions that are not helpful for equating.

To disguise the anchor items in a NEAT design, the items are often embedded within sections of scored operational items. Internal anchors or common items may not be located in the same item positions within the old and new forms, making them more susceptible to context effects that may diminish their utility as measures of ability. In addition, the common items may be few in number, making the anchor test relatively unreliable and less useful for identifying differences in ability between the samples.

2.6.2.3 Samples

Unrepresentative or unmotivated samples undermine equating. Special care should be taken to ensure that only members of the population of interest are included in the samples. If possible, the sample should be representative of the population as well.

With the NEAT design, the old- and new-form samples may perform very differently on the anchor test. Large ability differences on the anchor test tend to yield situations where equating is unsatisfactory unless the anchor is highly related to both tests to be equated. In this setting, different equating methods tend to give different answers unless the anchor test is strongly related to both the old and new tests. This divergence of results is indicative of a poor data collection design.

Equating cannot be done effectively in small samples. The smaller the sample size, the more restricted is the class of stable equating methods. Smoothing score distributions works in moderately sized samples but does not help much with very small samples, especially when it is not clear how representative the sample is of the intended population. In these situations, one option may be to make strong assumptions about the equating function (Livingston & Kim, Chapter 7, this volume). For example, it may be necessary to assume the identity is a reasonable approximation to the equating function or that the identity shifted by a constant that is estimated by the data provides a reasonable approximation.

The best practices solution to the small sample size problem may be to report raw scores and state that they cannot be compared across test forms. If the sample size suggested by consideration of standard errors is not achieved, raw scores could be reported with the caveat that they are not comparable to other scores, but that they could be made comparable when adequate data become available. This would protect testing organizations from challenges resulting from the use of either biased linking functions or unstable equating functions. To do otherwise might be problematic over the long term.

2.6.2.4 Lack of Population Invariance

One of the most basic requirements of score equating is that equating functions, to the extent possible, should be subpopulation invariant.[2] The "same construct" and "equal reliability" requirements are prerequisites for subpopulation invariance. One way to demonstrate that two tests are not equatable is to show that the equating functions used to link their scores are not invariant across different subpopulations of examinees. Lack of invariance in a linking function indicates that the differential difficulty of the two tests is not consistent across different groups. Note that subpopulation invariance is a matter of degree. In the situations where equating is usually performed, subpopulation invariance implies that the dependence of the equating function on the subpopulation used to compute it is small enough to be ignored.

Score equity assessment focuses on whether or not test scores on different forms that are expected to be used interchangeably are in fact interchangeable across different subpopulations (Dorans & Liu, 2009). The subpopulation invariance of linking functions is used across important subgroups (e.g., gender groups) to assess the degree of score exchangeability. Score equity assessment focuses on invariance at the reported score level. It is a basic quality control tool that can be used to assess whether a test construction process is under control, as can checks on the consistency of raw-to-scale conversions across forms (Haberman et al., 2008).

2.6.3 Additional Directions for Future Research

There is a need for comprehensive empirical investigations of equating conditions as well as additional theoretical work that can further inform the best practices described in this chapter. The various challenges discussed in previous portions of this section should be explored via systematic investigations of the appropriateness of different equating procedures in a variety of realistic settings. These empirical investigations have their progenitors, such as the comprehensive studies conducted by Marco, Petersen, and Stewart (1983a) as well as other studies cited in Kolen and Brennan (2004). Recent work by Sinharay and Holland (2010) is indicative of the kind of work that can be done to better understand the robustness of various procedures to violation of their assumptions (See also Sinharay, Holland, & von Davier, Chapter 17, this volume.)

Foremost among factors that need to be studied are the effects on equating results of the magnitude of ability differences between P and Q as measured by the anchor items and of the shape of the score distributions. In addition, it would be

[2] Note that these subpopulations should not be defined on the basis of the tests to be equated or the anchor test, because the assumptions made by equating methods are sensitive to direct selection on the test or anchor, as demonstrated by Wright and Dorans (1993).

worthwhile to manipulate difficulty differences between *X*, *Y* and *A* as well as the reliability of the total score and the anchor score, expanding on investigations such as Moses and Kim (2007). Correlations of the anchor score with total score and sample size should also be manipulated and studied. Ideally, real data would be used as the starting point for these studies.

Another area that needs attention is the consistency of equating results over long periods of time, a point made by Brennan (2007) and studied recently on the SAT® by Haberman et al. (2008). These researchers examined the consistency of SAT Math and SAT Verbal equatings between 1995 and 2005 and found them to be very stable. This type of work is especially important in settings where tests are administered on an almost continuous basis (Li, Li, & von Davier, Chapter 20, this volume). In these settings, substantial score drift may occur such that scores may not be comparable across periods as short as one year. The quest to test continuously may subvert one of the basic goals of fair assessment.

Several new methods for equating as well as some new definitions have been and will be introduced. These methods should be stress tested and adapted before they are adopted for use. Procedures that make strong assumptions about the data may give answers that are theoretically pleasing but are difficult to apply in practice and even more difficult to justify to test users. Holland (1994) noted that tests are both measurements and contests. They are contests in the sense that examinees expect to be treated fairly—equal scores for comparable performance. Equating, as discussed by Dorans (2008), can be thought of as a means of ensuring fair contests: An emphasis needs to be placed on fair and equitable treatment of examinees that is commensurate with their actual performance on the test they took. The use of best practices in equating is essential to achieving this goal.

The focus of this chapter has been on best practices for score equating. Score equating is only one aspect of the score reporting process. Other components of the score reporting process affect the final raw-to-scale conversions. Because these components are not as amenable to mathematical treatment as score equating methods, they have not received as much treatment as they should. The best score equating practices can be undermined by a weakness elsewhere in the process, such as poorly defined test specifications or the use of a flawed old-form scaling function. A few of these non-score-equating components have been mentioned in this report, but the treatment has not been as complete as it should be.

Author Note: Any opinions expressed in this chapter are those of the authors and not necessarily of Educational Testing Service.