Research Report

Principles and Practices of Test Score Equating

Neil J. Dorans

Tim P. Moses

Daniel R. Eignor

December 2010

ETS RR-10-29

Principles and Practices of Test Score Equating

Neil J. Dorans, Tim P. Moses, and Daniel R. Eignor
ETS, Princeton, New Jersey

December 2010

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Technical Review Editor: Don Powers

Technical Reviewers: Sooyeon Kim and Skip Livingston

Abstract

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. Particularly in testing programs that help make high-stakes decisions, it is extremely important that test equating be done carefully and accurately. 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. Because the reported score is so visible, the credibility of a testing organization hinges on activities associated with producing, equating, and reporting scores. This paper addresses the practical implications of score equating by describing aspects of equating and best practices associated with the equating process.

Key words: score equating, score linking, equating methods, data collection, best practices

Table of Contents

List of Tables

Page

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 be 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.

Particularly in testing programs that help make high-stakes decisions, 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. Because the reported score is so visible, the credibility of a testing organization hinges on activities associated with producing, equating, and reporting scores

This paper addresses the practical implications of score equating by describing best (and some not-best) practices associated with the equating process.[1] Even just since 2000, there have been several books published on equating, as well as several review chapters. These are cited in Section 1 of this paper. Readers should refer to these sources for extensive treatments of equating and linking. Section 1 introduces and distinguishes test score equating as a special case of the more general class of procedures called score linking procedures. Section 2 is concerned with the material 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 3 focuses on the most common data collection designs used in the equating of test scores. In Section 4 we describe some common observed-score equating functions. Section 5 describes common data-processing practices that occur prior to computations of equating functions. Section 6 is concerned with how to evaluate an equating function, as well as post-equating activities.

1. Linking and Equating: Foundational Aspects

The term *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. Since the turn of the century, 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 has been provided by Kolen and Brennan (2004). The book by von Davier, Holland, and Thayer (2004) introduced several new ideas of general use in equating, although its focus is on kernel equating. *Uncommon Measures* (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999) and *Embedding Questions* (Koretz, Bertenthal & Green, 1999), two book-length reports from the National Research Council, contain summaries of informed, professional judgment about the issues involved in linking scores on different educational tests. Livingston (2004) has given a straightforward description of many of the major issues and procedures encountered in practice.

Holland and Dorans (2006) provided a historical background for test score linking, building on work by Angoff (1971), Flanagan (1951), and Petersen, Kolen, and Hoover (1989) Holland and Dorans (2006) discussed ways other than test equating that scores on different tests are connected or linked together. Several chapters in Dorans, Pommerich, and Holland (2007) addressed important issues in score equating (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 through 6.

1.1. Classes of Score Linking Methods: Definition of Terms

Holland and Dorans (2006) provided a framework for classes of score linking that built on and clarified earlier work found in Mislevy (1992) and Linn (1993). Holland and Dorans made distinctions between different types of linkages and emphasize that these distinctions are related to how linked scores are used and interpreted. A *link* between scores on two tests is a transformation from a score on one test to a score on another test.

There are different types of links, and the major difference between these types is not procedural, but interpretative. Holland and Dorans (2006) divided linking methods into three basic categories called *predicting*, *scale aligning,* and *equating*. It is essential to understand why these categories differ because they are confused in practice. For the sake of best practices,

understanding the distinctions among these categories and communicating them to test users can prevent violations of professional practices.

Predicting. Predicting is the oldest form of score linking. Since the earliest days of psychometrics, predicting has been confused with equating. An example comes from Embretson and Reise (2000): "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 predict an examinee's score on a test based on other information about that examinee. This information can be multivariate in nature, including scores from several other tests, demographic information, or other types of cognitive information (e.g., grades in selected courses). The goal 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 that will be discussed in the paper, namely to produce scores that can be used interchangeably. Hence, prediction cannot be used to equate scores or produce scores with comparable score properties.

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. Scale aligning 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). The interested reader is urged to consult Holland and Dorans (2006) and Kolen (2006). Scale aligning and score equating are often confused because the statistical procedures used for scale alignment also can be used to equate tests.

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 test design practices that can help ensure the achievement of equating requirements are described in Section 2.

A multi-step process is used to put scores from a new test onto an existing score reporting scale. Best practices should be used in all of these steps. 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), which links raw scores x on a new test, **X**, to those of an old test form **Y**. Equated raw scores are produced via this linking. The final step is to produce a function that converts these equated **X** raw scores to the score reporting scale by composing the equating function, y = e(x) with s(y) to put the raw scores of **X** onto the reporting scale, s(x) = s(e(x)), known as the score conversion function for **X**. We will discuss the final score conversion function that is used in practice because it is the ultimate end-product of the equating process that will be described in Sections 2–6. The raw-to-raw equating is not an end, but rather the means to an end—an appropriate score conversion function. This critical point is sometimes given short shrift in discussions of equating that focus on methods.

1.2. 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.

In an ideal world, test forms would be assembled to be strictly parallel so that they would have identical psychometric properties. Equating would then be unnecessary. In reality, it is virtually impossible to construct multiple forms of a test that are strictly parallel, and equating is necessary to fine-tune the test construction process.

Five requirements are widely viewed as necessary for a linking to be an equating (Holland & Dorans, 2006). Those requirements are:

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 **Y** to those of **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 (sub) population from which it is derived.

Both formal and informal statements of subsets of these five requirements have appeared in a variety of earlier sources, including Angoff (1971), Lord (1950, 1980), Kolen and Brennan (2004), and Petersen et al. (1989). Dorans and Holland (2000) explicitly discussed these five requirements and indicate various ways in which the five "…can be criticized as being vague, irrelevant, impractical, trivial, or hopelessly stringent" (p. 283). For example, Lord (1980) regarded Requirement 4 as the most fundamental, while Livingston (2004) argued that Requirements 4 and 5 are unattainable in practice.

With respect to best practices, Requirements 1 and 2 mean that the tests need to be built to the same specifications, while Requirement 3 precludes regression methods from being a form of test equating. Lord (1980) argued that Requirement 4 implies both Requirements 1 and 2. Requirement 4 is, however, hard to evaluate empirically and its use is primarily theoretical (Hanson, 1991; Lord, 1980). As noted by Holland and Dorans (2006), Requirement 5, which is easy to assess in practice, also can be used to explain why Requirements 1 and 2 are needed. If two tests measure different things or are not equally reliable, then the standard linking methods will not produce results that are invariant across certain subpopulations of examinees. Dorans and Holland (2000) used Requirement 5, rather than Requirement 4, to develop quantitative measures of equatability that indicate the degree to which equating functions depend on the subpopulations used to estimate them. For example, a conversion table relating scores on a mathematics test to scores on a verbal test developed on data for men would be very different from one developed from data on women, since, women tend to do less well than men on

mathematics tests. Articles that study subpopulation issues in equating contexts and more general linking contexts are included in special issues of *Journal of Educational Measurement* (Dorans, 2004) and *Applied Psychological Measurement* (von Davier & Liu, 2008).

2. Test Specifications and Score Linking Plans

2.1 Test Specifications

Based on the equity condition (Requirement 4 in Section 1.2), 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. While there is not much that 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 are composed of 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 can often undermine the quality of tests and test scores by draining pools of pretested items more quickly than these items can be replenished.

2.2 Anchor Test

As noted earlier, an anchor test often plays a crucial role in the equating process (see Sections 3.4, 4.2, and 4.3). An anchor test design allows a new test to be used and equated at each successive operational test administration. This is desirable in high-stakes situations where test reuse may lead to test security problems. The anchor test is used to account for any differences in ability between nonequivalent groups taking the new and old test forms. The anchor needs to be highly related to both tests.

It is generally considered good practice to construct the anchor test according to the test specifications, so that it is a mini-version 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 some cases, the anchor test is administered externally in a separately timed section and performance on this section does not count toward examinee scores. Sometimes when such an external anchor is not available, internal anchors, composed of scored items that are interspersed among other scored items, are used.[2] In this case, 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. For a more complete discussion of context effects, see Section 3.5.

The use of an anchor test design requires users to make one of several possible sets of untestable, missing-data assumptions in order to interpret the linking results as constituting an equating, as will be discussed in Section 4.2. In addition, great care must be taken in collecting and analyzing the data with these anchor test designs, as we will discuss in Section 5.3.

2.3 Score Linking Plans

The raw-to-raw equating is not an end, but rather the means to an end, namely an appropriate score conversion function. As noted earlier, a multi-step process is used to put scores from a new test onto an existing score reporting scale and best practices should be used for all of these steps. 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 itself is problematic. Even tests as widely used as the SAT® could have undesirable new-form scalings that were arising from poor alignment of the score scale with the intended uses of the test score. In the case of the SAT, poor score scale alignment, where 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. This was rectified by recentering the SAT scores (Dorans, 2002). Many score scales suffer from poor construction while others discard useful information because of the way the meaning of the scale changes over time. For example, many of the numerous 800 scores on the GRE ® Quantitative scale would exceed 800 if the scale were not capped at 800 (Dorans,

Yu, & Guo, 2006). 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 using 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 and hope that some magical method of score equating will play the role of *deus ex machina* to set scores straight. Data collection planning, development of linking plans, and maintenance of score scales are crucial best practices.

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, non-strictly parallel tests are given to two different groups of examinees of unequal ability. Assuming that the samples are large enough so that one can ignore sampling error, differences in the distributions of the resulting scores can be due to one or both of two factors. One factor is the relative *difficulty* of the two tests and the other is 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 that 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 those that make 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 can be found in Angoff (1971), Holland and Dorans (2006), Kolen and Brennan (2004), Petersen et al. (1989), and von Davier et al. (2004).

3.1. The Single Group (SG) Design

The single group (SG) 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.

The design table (von Davier et al., 2004) for the SG design is given in Table 1.

Table 1

Design Table for the Single Group (SG) Design

Population	Sample	X	Y
P	1	@	@

Note. **@** indicates examinees in sample for a given row take tests indicated in a given column; lack of **@** indicates score data were not collected for that combination of row and column.

The SG design controls for any possibility of differential examinee proficiency by having the same examinees take both tests. It has several major uses in the practice of scaling and equating. In using this design, however, it is necessary to assume that an examinee's score on the second test form is unaffected by the fact that she or he previously has taken the first form. That is, it must be plausible that practice and other types of order effects can be ignored.

3.2. The Equivalent Groups (EG) Design

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 (EG) design, two equivalent samples are taken from a common population **P** one is tested with **X** and the other with **Y**. The EG design is often used for equating.

The design table for the EG design, Table 2, clearly shows the pattern of missing data (i.e., the cells missing the **@** mark).

Table 2

The Design Table for the Equivalent Groups (EG) Design

Population	Sample	X	Y
P	1	@	
P	2		@

Note. **@** indicates examinees in sample for a given row take tests indicated in a given column; lack of **@** indicates score data were not collected for that combination of row and column.

Because examinees take only one test, the issue of order effects does not arise with the EG design. The problem is to select samples that are equivalent in whatever the tests are supposed to measure. In practice, this is done in two ways. First, it is sometimes possible to take two random samples from **P** and test each with a single test. To reflect this possibility, this design is sometimes called the *random groups design* (Kolen & Brennan, 2004). The two samples are then as equivalent as two random samples from the same population can be. Second, and more commonly, two samples are constructed 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 **X**, then **Y,** and then **X** again, and so on. Certain assumptions must hold in order for spiraling to be feasible. For example, the time limits must be the same for the two tests. Well-executed, spiraled samples are often somewhat more *equivalent* (i.e., less different) than random samples. They are *more equivalent* because they are approximately *stratified* random samples where the strata are the administrative divisions of the tested population (i.e., classrooms, schools).

The EG 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. It also has some limitations. One limitation is that it requires large sample sizes to produce accurate equating results. It also may have some consequences for test security because

in most cases the old form in the design will have been administered previously. However, when samples sizes are large and forms can be reused without security problems, the EG design is usually regarded as a good choice because it avoids the issue of possible order effects that can arise in the SG design, where each examinees takes *both* tests.

3.3. The Counterbalanced (CB) Design

In order to allow for the possibility of order effects in the SG design, the sample is sometimes randomly divided in half and in each half-size subsample the two tests are taken in different orders—**X** first and then **Y** or **Y** first and then **X**. The result is the counterbalanced (CB) data collection design.

If we denote a score from **X** as $\mathbf{X}_1$ when it is taken first and $\mathbf{X}_2$ when it is taken second, and similarly for $\mathbf{Y}_1$ and $\mathbf{Y}_2$, then Table 3 describes the CB design.

Table 3

Design Table for the Counterbalanced (CB) Design

Population	Sample	X_1	X_2	Y_1	Y_2
P	1	@			@
P	2		@	@	

Note. **@** indicates examinees in sample for a given row take tests indicated in a given column; lack of **@** indicates score data were not collected for that combination of row and column.

The CB design contains both the SG and EG designs within it. There are SG designs for both $\mathbf{X}_1$ and $\mathbf{Y}_2$ and $\mathbf{X}_2$ and $\mathbf{Y}_1$. There is an EG design for $\mathbf{X}_1$ and $\mathbf{Y}_1$ and for $\mathbf{X}_2$ and $\mathbf{Y}_2$. The main advantage of the CB design is the same as that of the SG design: accurate equating results from relatively small samples. Its main disadvantage is that it seldom can be fit within an operational administration of a test. Usually, the CB design requires a special study for collecting the data.

3.4. The Anchor Test or Nonequivalent Groups With Anchor Test (NEAT) Design

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. (2004) and call this the *nonequivalent groups with anchor test* (or NEAT) design. Kolen and Brennan (2004) and others

have referred to this as the *common-item nonequivalent groups design* or simply the *common item* or the *anchor test* design.

The NEAT design is used for equating and some forms of scale aligning, as indicated in Holland and Dorans (2006). Table 4 represents the NEAT design.

Table 4

Design Table for the Nonequivalent Groups With Anchor Test (NEAT) Design

Population	Sample	X	A	Y
P	1	@	@	
Q	2		@	@

Note. **@** indicates examinees in sample for a given row take tests indicated in a given column; lack of **@** indicates score data were not collected for that combination of row and column.

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.[3]

Formally, the NEAT design contains two single-group designs within it. The anchor test design is more flexible than the EG design because it allows the two samples taking **X** and **Y** to be different or nonequivalent. It is also more efficient than the SG design because it does not require examinees to take both **X** and **Y**. While 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 not necessary in the use of SG, EG, and CB designs, where common examinees are used; see Sections 2.1 to 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, as reported in Sections 4.2 and 4.3 of this paper.

One way to think about the difference between the NEAT design and the SG, EG, and CB designs is as the difference between observational studies versus experimental designs

(Rosenbaum, 1995). The SG design is like a repeated measures design with a single group and two treatments, the EG design is like a randomized comparison with two treatment groups, and the CB 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 where there are two nonrandomized study groups that are possibly subject to varying amounts of self-selection.

When **P** and **Q** are different or *nonequivalent*, the statistical role of **A** is to remove bias in the equating function that would occur if we presumed the groups were equivalent, as well as to increase precision in the estimation of the equating function. When **A** is a miniature version of **X** and **Y** (i.e., a mini-test that is shorter and less reliable but otherwise measures the same construct as the two tests to be linked), it can be expected to do a good job of removing any bias due to the nonequivalence of **P** and **Q**. When **A** is not really a measure of the same construct as **X** and **Y**, or if it is not highly correlated with them, **A** is less useful for removing bias or for increasing precision.

3.5 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 SG design requires the smallest sample sizes and the EG design requires the largest sample sizes to achieve the same level of accuracy, as measured by the standard error of equating (see Holland & Dorans, 2006; Lord, 1950). The anchor test (i.e., NEAT) designs require sample sizes somewhere in between those of the SG and EG 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 do 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 (EG) 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, and if they are substantial the anchor test can be ignored, leaving a simple EG 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 EG 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 EG design, then the issue of examinee motivation arises, as discussed in Holland and Dorans (2006)

The SG design requires a smaller sample size to achieve the same level of statistical accuracy as that obtained by an EG design with a larger sample, but it brings with it issues of order effects and it requires twice as much time to administer both tests. A particular problem with the SG design is that there is no way to assess whether order effects exist. The CB 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 EG design. Because of the greatly reduced sample size, the resulting EG design may produce equating results that are less accurate than desired. Von Davier et al. (2004) proposed a formal statistical decision process for assessing order effects under the CB design.

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 does the number of anchor tests and the type of score linking plan employed.

External anchor tests. It is often advised that the anchor test be a mini-version of the two tests being equated (Angoff, 1971). Making the anchor test a mini-version 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 mini-version 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 term *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, such as 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 do count towards their score (even though they are instructed not to do this). While this type of behavior may appear to benefit these examinees, because of the way that the anchor test is used in equating, such behavior may actually 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 under-perform 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 relative to the old test than it really is. In score equating, a raw score on an easier test is converted to a lower scaled score than would 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. In practice this effect is likely to be small, and for those examinees who worked on another section during the anchor test, the effect may be canceled by an increased score on the other section. As indicated in Section 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.

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. As noted earlier, 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 4, where the score on the anchor test plays an important role, it is desirable for the anchor test to be a mini-version 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 have been 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.

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 4. To the extent that an anchor test design (Section 3.4) is almost an EG design (Section 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 being equated. The most important properties of the anchor test are its stability across 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. An advantage of internal anchors over external anchors is that their correlations with the tests being equated are usually high because the anchor items contribute to the total score.

The implication of needing highly correlated anchors for best practices is that long anchor tests are generally better than short ones for equating. Longer anchors are usually more reliable and more highly correlated with the tests. In practice, it is desirable that both the anchor test and the tests being equated have high reliability.

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. 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.

4. Procedures for Equating Scores

Many procedures for equating tests have been developed over the years. 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**, while 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 his or her 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) have shown 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 3. Our focus is on observed-score equating because true scores are unobserved and consequently primarily of theoretical interest-only. We provide brief discussions of only the most common observed-score procedures. The reader should consult Holland and Dorans (2006) for more complete treatments of observed-score and true score procedures. Here we first consider procedures used with common-population data collection designs and then procedures used with anchor tests or common-item designs. Within these two types of data collection designs we will look at linear and nonlinear procedures, which we will discuss at the same time. In addition we will be explicit about the population of examinees on which scores are equated. Common to any equating scenario is a population of examinees that we will call the *target population*, **T**, following the usage in von Davier et al.

(2004). In this usage, the target population refers to the source of the samples used to compute the linking function.

4.1 Observed Score Procedures for Equating Scores on Complete Tests Given to a Common Population

Three data collection designs that we described in Section 3 make use of a common population of examinees: the SG, the EG, and the CB 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 $F_{\mathbf{T}}(x)$, and it is defined as the proportion of examinees in **T** who score at or below x on test **X**. More formally, $F_{\mathbf{T}}(x) = P\{\mathbf{X} \leq x \mid \mathbf{T}\}$, where $P\{\ . \mid \mathbf{T}\}$ denotes the population proportion or probability in **T**. Similarly, $G_{\mathbf{T}}(y) = P\{\mathbf{Y} \leq y \mid \mathbf{T}\}$, is the cdf of **Y** over **T**. Cumulative distribution functions 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 4.3, we address the issue of the discreteness of score distributions in detail.

The equipercentile equating function. The equipercentile definition of *comparable* scores is that x (an **X**-score) and y (a **Y**-score) are *comparable* in **T** if $F_{\mathbf{T}}(x) = G_{\mathbf{T}}(y)$. 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 $F_{\mathbf{T}}(x) = G_{\mathbf{T}}(y)$ can always be satisfied and can be solved for y in terms of x. Solving for y leads to the *equipercentile function*, $Equi_{\mathbf{YT}}(x)$, that links x to y on **T**, defined by:

$$y = Equi_{\mathbf{YT}}(x) = G_{\mathbf{T}}^{-1}(F_{\mathbf{T}}(x)). \qquad (1)$$

In Equation 1, $y = G_{\mathbf{T}}^{-1}(p)$ denotes the inverse function of $p = G_{\mathbf{T}}(y)$. 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.

We have followed Dorans and Holland (2000), Holland and Dorans (2006), and von Davier et al. (2004) in explicitly including the target population **T** in the definition of $Equi_{\mathbf{YT}}(x)$. The notation emphasizes that **T** (as well as **X** and **Y**) can influence the form of the equipercentile function.

In general, there is nothing to prevent $Equi_{\mathbf{YT}}(x)$ from varying with the choice of **T**, thereby violating the subpopulation invariance requirement of Section 1.2. 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**.

The linear equating function. If Equation 1 is satisfied, then $Equi_{\mathbf{YT}}(x)$ will transform the distribution of **X**-scores on **T** so that it is the same as the distribution of **Y**-scores on **T**.

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

$$F_{\mathbf{T}}(x) = K[(x—\mu_{\mathbf{XT}})/\sigma_{\mathbf{XT}}] \text{ and } G_{\mathbf{T}}(y) = K[(y—\mu_{\mathbf{YT}})/\sigma_{\mathbf{YT}}], \quad (2)$$

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

When Equation 2 holds, $F_{\mathbf{T}}(x)$ and $G_{\mathbf{T}}(y)$ both have the shape determined by K. In this case, it can be shown that the equipercentile function is the *linear function*, $Lin_{\mathbf{YT}}(x)$, defined as

$$Lin_{\mathbf{YT}}(x) = \mu_{\mathbf{YT}} + (\sigma_{\mathbf{YT}}/\sigma_{\mathbf{XT}})(x—\mu_{\mathbf{XT}}). \quad (3)$$

The linear function may also 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 (c) of Section 1.2.1. This means that $Lin_{\mathbf{XT}}(y) =$

$Lin_{\mathbf{YT}}^{-1}(x)$, and $Equi_{\mathbf{XT}}(y) = Equi_{\mathbf{YT}}^{-1}(x)$ (i.e., equating **Y** to **X** is the inverse of the function for equating **X** to **Y)**. In general, the function, $Equi_{\mathbf{YT}}(x)$, curves around the function, $Lin_{\mathbf{YT}}(x)$.

Two special cases of $Lin_{\mathbf{YT}}(x)$ that follow from very strong assumptions are the *mean linking function* and the *identity function*. When the two standard deviations in Equation 3 are equal, then $Lin_{\mathbf{YT}}(x)$ takes on the form $Mean_{\mathbf{YT}}(x) = x + (\mu_{\mathbf{YT}} — \mu_{\mathbf{XT}})$. The mean linking function adjusts the scores of **X** so that they have the same mean as **Y** does on **T**. When both the means and the standard deviations in Equation 3 are equal, $Lin_{\mathbf{YT}}(x)$ takes on the form $Iden(x) = x$. The identity function makes no adjustment at all to the **X**-scores. It corresponds to assuming that the raw scores on **X** and **Y** are already comparable. Both $Mean_{\mathbf{YT}}(x)$ and $Iden(x)$ are thought to be useful best practices when the samples are very small and cannot support accurate estimates of the moments of **X** and **Y** on **T**. They are discussed in more detail in Kolen and Brennan (2004) and Skaggs (2005). We address them in Section 6.

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 SG and EG designs described in Section 3 (see Angoff, 1971, or Kolen & Brennan, 2004). It is less straightforward to obtain estimates for the CB design, as noted by Holland and Dorans (2006).

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 3.4)—there are two populations, **P** and **Q**, and either one 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** see all items on **Y**. As indicated at the beginning of Section 3, 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.

Drawing from what they saw being done in practice, 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** = w**P** + (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**. When w = 1, **T** = **P** and when w = 0, **T** = **Q**. In practice, w is often taken to be proportional to the two sample sizes from **P** and **Q**. This choice of w is implicit when the data for the anchor test are pooled into a total group, as done in Angoff (1971) and Petersen et al. (1989). Of course, other choices of w are possible, such as w = ½, which gives equal weight to **P** and **Q**. There is considerable evidence that the choice of w has a relatively minor influence on equating results, for example, see von Davier et al. (2004). This insensitivity to w is an example of the population invariance requirement of Section 1.2. 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 *post stratification 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** = w**P** + (1—w)**Q** (i.e., 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. However, they are likely to have parallels for true-score equating as well, a topic worthy of future research. In addition to the PSE and CE types of procedures, classical test theory may be used to derive an additional, less frequently used, *linear observed-score* procedure for the NEAT design—the Levine observed-score equating function (Kolen & Brennan, 2004).

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** = w**P** + (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 EG 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** = w**P** + (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 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 EG 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.

The PSE types of equating procedures. There are both nonlinear and linear PSE procedures. They may be viewed as based on the following two assumptions, which we adopt from von Davier et al. (2004).

> **PSE1**: The conditional distribution of **X** given **A** over **T**, $P\{\mathbf{X} = x \mid \mathbf{A} = a, \mathbf{T}\}$ is the same for any **T** of the form **T** = w **P** + (1—w) **Q**.
>
> **PSE2**: The conditional distribution of **Y** given **A** over **T**, $P\{\mathbf{Y} = y \mid \mathbf{A} = a, \mathbf{T}\}$ is the same for any **T** of the form **T** = w **P** + (1—w) **Q**.

PSE1 and PSE2 are population invariance assumptions because they require that the conditional distributions are the same for any target population of the form **T** = w **P** + (1—w) **Q**.

The clearest examples of procedures of the PSE type are frequency estimation equating (Angoff, 1971, Petersen et al., 1989, and Kolen & Brennan, 2004), and the PSE version of kernel equating (von Davier et al., 2004).

Linear observed-score PSE equating procedures include (a) Tucker equating (Angoff, 1971, Petersen et al., 1989, and Kolen & Brennan, 2004), (b) the Braun-Holland method (Braun & Holland, 1982, and Kolen & Brennan, 2004), and (c) the linear PSE version of kernel equating (von Davier et al., 2004). The linear PSE version of kernel equating is a way to implement the Braun-Holland procedure and both are directly based on PSE1 and PSE2. Ledyard R Tucker was originally motivated by selection theory in the development of the method that bears his name (Angoff, 1971). However, the following versions of PSE1 and PSE2 may also be used to derive Tucker equating with no reference to selection.

> **TUCK1**: (a) The conditional mean of **X** given **A** over **T** is linear in **A** and is the same for any **T** = w **P** + (1—w) **Q**, and (b) the conditional variance of **X** given **A** over **T** is constant in **A** and is the same for any **T**.
>
> **TUCK2**: (a) The conditional mean of **Y** given **A** over **T** is linear in **A** and is the same for any **T** = w **P** + (1—w) **Q**, and (b) the conditional variance of **Y** given **A** over **T** is constant in **A** and is the same for any **T**.

TUCK1 and TUCK2 are population invariance assumptions in the same sense that PSE1 and PSE2 are.

The Braun-Holland and the linear PSE version of kernel equating do not make the more restrictive assumptions of linear conditional means and constant conditional variances that appear in TUCK1 and TUCK2. For this reason, they may give somewhat different results from the Tucker method when the conditional means are nonlinear and/or the conditional variances are not constant.

The CE types of equating procedures. The idea behind the CE procedures is to first link **X** to **A** using the data from **P**, then to link **A** to **Y** using the data from **Q**, and finally to combine these two links to equate **X** to **Y** through **A**. Von Davier et al. (2004) showed that the

following two assumptions are sufficient to interpret chain equating as an observed-score equating function for any target population of the synthetic population form.

CE1: The equipercentile function linking **X** to **A** on **T** is the same for any **T** of the form **T** = w **P** + (1—w) **Q**.

CE2: The equipercentile function linking **A** to **Y** on **T** is the same for any **T** of the form **T** = w **P** + (1—w) **Q**.

There are both linear and nonlinear versions of CE. Linear observed-score CE equating procedures include (a) the chained *linear equating function* (Angoff, 1971) and (b) the linear CE version of kernel equating (von Davier et al., 2004). Because it is derived as a type of equipercentile equating function, the linear CE version of kernel equating is based on assumptions CE1 and CE2. Von Davier, Holland, and Thayer (2004) maintained that the chained *linear equating function* is a linear equating function on a target population **T** as defined in Equation 3 when the linear versions of CE1 and CE2 hold. These are:

CL1: The linear linking function equating **X** to **A** on **T** is the same for any **T** of the form **T** = w **P** + (1—w) **Q**.

CL2: The linear linking function equating **A** to **Y** on **T** is the same for any **T** of the form **T** = w **P** + (1—w) **Q**.

Again, CL1 and CL2 are examples of population invariance assumptions.

4.3 A Linear Observed-Score Equating Procedure From Classical Test Theory

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, $Lev_{\mathbf{YT}}(x)$ (Kolen & Brennan, 2004). $Lev_{\mathbf{YT}}(x)$ 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.

Following Holland and Dorans (2006), we use the formulation of classical test theory described in Holland and Hoskins (2000). The true scores, $\tau_{\mathbf{X}}$ and $\tau_{\mathbf{Y}}$, are defined as latent variables underlying each test that have these properties: $\tau_{\mathbf{X}} = E(\mathbf{X} \mid \tau_{\mathbf{X}}, \mathbf{T})$, and $\tau_{\mathbf{Y}} = E(\mathbf{Y} \mid \tau_{\mathbf{Y}}, \mathbf{T})$, for any target population, **T**. From these two assumptions it follows that $\mu_{\mathbf{XT}} = E(\mathbf{X} \mid \mathbf{T}) = E(\tau_{\mathbf{X}}$

|**T**), and $\mu_{\mathbf{YT}} = E(\mathbf{Y}|\mathbf{T}) = E(\tau_{\mathbf{Y}}|\mathbf{T})$. To formalize the assertion that **X** and **Y** measure the same construct, we assume the true scores are *congeneric*, that is, that they are linearly related by

$$\tau_{\mathbf{Y}} = \alpha\tau_{\mathbf{X}} + \beta, \tag{4}$$

where α and β may depend on the target population, **T**. The idea behind true-score equating is to estimate α and β and to use Equation 4 to find the link between the two sets of true scores. Lord (1980) takes the position that only true scores can ever really be equated, but it can be argued that this is a consequence of his very stringent interpretation of the equity requirement of Section 1.2.

Holland and Dorans (2006) used the form of classical test theory discussed in Holland and Hoskens (2003) to derive the Levine observed score equating function from two assumptions. In addition to using the true score definitions for test scores **X** and **Y** above, they defined an anchor true score as a latent variable, $\tau_{\mathbf{A}}$, that underlies the observed anchor score **A,** and that satisfies: $\tau_{\mathbf{A}} = E(\mathbf{A} \mid \tau_{\mathbf{A}}, \mathbf{T})$, for any target population, **T**. From this definition, it follows that: $\mu_{\mathbf{AT}} = E(\tau_{\mathbf{A}} \mid \mathbf{T})$. To formalize the intuition that **X**, **Y,** and **A** all measure the same construct, we assume their true scores are linearly related in a way that holds for all **T** (i.e., the three measures are congeneric). These assumptions are given below:

LL1: $\tau_{\mathbf{X}} = \alpha\tau_{\mathbf{A}} + \beta$, where α and β do not depend on the target population, **T**.

LL2: $\tau_{\mathbf{Y}} = \gamma\tau_{\mathbf{A}} + \delta$, where γ and δ do not depend on the target population, **T**.

The results of all of these assumptions and definitions are that the Levine observed-score equating function and its assumptions are true score analogues of the Tucker equating method and the Tucker method's assumptions. As shown in Kolen and Brennan (2006) and elsewhere, these assumptions about true scores lead to an equating relationship for observed scores.

5. Data Processing Practices Prior to Computation of Equating Functions.

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.

5.1 Sample Selection

Tests are designed with a target population in mind (defined as **T** throughout Section 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. 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 may have already been excluded if they were not part of the target group. There is an issue as to whether non-native speakers of the language in which the test is administered should also be excluded. One study by Liang, Dorans, and Sinharay (2009) suggests this may not be an issue as long as the proportion of non-native 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 (i.e., 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.

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 (DIF) 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 may also 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 may also 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.

5.3 The Need to Continuize the Discrete Distributions of Scores

The equipercentile function defined in Section 5.2 can depend on how $F_{\mathbf{T}}(x)$ and $G_{\mathbf{T}}(y)$ 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 1 of Section 4.1.1, is not well defined (i.e., for many values of p, there is no score, y, for which $p = G_{\mathbf{T}}(y)$). This is not due to the *finiteness of real samples*, but rather to the *discreteness of real test scores*. To get around this, there are three methods of continuization of $F_{\mathbf{T}}(x)$ and $G_{\mathbf{T}}(y)$ that 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 3 of Section 4.1 is a third continuization method.

There are two primary differences between the first two approaches to continuization. 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 post-smoothing (Fairbank, 1987, Kolen & Brennan, 2004). In contrast, kernel smoothing results in equipercentile functions that are very smooth (i.e., differentiable everywhere) and that do not need further post-smoothing. Second, the equipercentile functions obtained by linear interpolation always map the highest score on **X** into the highest score on **Y** and the same for the lowest scores (unlike kernel smoothing and the linear equating function). While it is sometimes desirable, there are cases where 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. (2004).

5.4. Presmoothing Score Distributions

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 *presmooth* the raw-score frequencies in some way prior to equipercentile equating. 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. If presmoothing is done so as to preserve the essential features of the score frequencies, it will reduce the sampling variability in the estimated frequencies without introducing significant bias. The resulting estimates will be closer to the underlying frequencies in the target population, **T**. When presmoothing is done with a model that does not describe the data well, then the estimated frequencies will be biased estimates of the underlying frequencies in **T**.

A limitation of equipercentile equating is that the equating relationship cannot be computed for any possible scores above the highest observed score or below the lowest observed score. If we could observe the scores of the entire target population, **T**, on both forms of the test, this limitation would not be a problem. Smoothing can help solve this problem because many smoothing methods will produce a smoothed distribution with probabilities (possibly very small) at the highest and lowest score levels, even if no test-takers actually attained those scores.

Kolen and Jarjoura (1987) and Kolen and Brennan (2004) discussed several methods for presmoothing. Von Davier et al. (2004) described the use of loglinear models for presmoothing. Their work is based on Holland and Thayer (1987, 2000), and they gave examples of presmoothing for the SG, EG, CB, and NEAT designs.

The type of data available for presmoothing depends on the data collection design. The EG design is the simplest and results in two independent univariate score distributions, one for **X** and one for **Y**. These may be independently presmoothed. The SG, CB, and NEAT designs result in one or more bivariate distributions containing the joint frequencies for the (**X,Y**)-, (**X,A**)-, or (**Y,A**)-pairs in the sample(s). For these designs, presmoothing should be done on the joint distribution(s). Presmoothing of only the marginal distributions, as if the **X**, **Y,** and **A** scores were all from different samples, ignores the correlations between **X**, **Y,** and **A** and can lead to incorrect standard error estimates.

When presmoothing data, it is important to achieve a balance between a good representation of the original data and smoothness. Smoothness reduces sampling variability while a good representation of the data reduces the possibility of bias. For example, if a loglinear model is used, it needs to preserve the most important features of the data, such as means, variances, and skewness and any other special features. The more parameters that are estimated for the model, the better the model will represent the original data, but the less smooth the fitted model becomes.

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 quality equating and close by discussing directions for additional research.

6.1 Best Practices

While we emphasized the structure of the data collection designs in Section 3, it should be mentioned here that the *amount* of data collected (sample size) has a substantial effect on the usefulness of the resulting equatings. 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 (DIF) 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 EG 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 (DTM), 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) describe the DTM, the standard error of equating and the standard error of the difference in equating or SEED. If the departures from linearity are less than the DTM 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.

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. As Tukey (1963) indicated in his discussion of Rasch's (1960) quest for the best fitting model, "…We must be prepared to use many models, and find their use helpful for many specific purposes, when we already know they are wrong—and in what ways. …In data analysis…we must be quite explicit about the deficiencies of the models with which we work. If we take them at face value, we can—all too frequently—be led to unreasonable and unhelpful actions. If we try to make them 'fit the facts,' we can ensure sufficient mathematical complexity to keep us from any useful guidance" (p. 504).

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.

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 practices are not followed.

Data collection design issues. Some threats to sound equating are related to the choice of data collection design. Test security is an issue for many high-stakes licensure, certification, and

admissions tests. To help maintain test security, many testing programs want to give a new form of the exam at every administration. Consequently, they do not want to re-administer an old form for equating purposes. Instead, they prefer to use an anchor test or common item design so only a subset of items is re-administered for equating purposes. 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.

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. At the extreme, tests may turn out to be too easy or too difficult for the intended population; this results in data that are not amenable for linking to other tests because the distributions are so skewed, and relationships with other scores are attenuated.

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 may also 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.

Samples. Unrepresentative or unmotivated samples undermine equating. Special study data collections need to include incentives that ensure that examinees will take the test seriously. 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 sample 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 is, the more restricted is the class of stable equating methods. It is often useful to pre-smooth the sample frequencies, especially when samples are not large enough to yield small standard errors of equating. 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. For example, it may be necessary to assume that it is the identity or that it differs from the identity by a constant that is estimated by the data. Another alternative is the circle-arc linking method (Livingston & Kim, 2009), which makes arbitrary assumptions about the score points that anchor the end points on an arc that runs through these two points and the mean scores. The authors have used simulation results based on real data to demonstrate that this procedure is superior under certain conditions to other procedures. These findings are normative in nature in that the method did better than other methods. In an absolute sense, however, the results of this method still may fall short of producing satisfactory equating when samples are small in size.

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.

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 (requirement e in Section 1.2.1).[4] The *same construct* and *equal reliability* requirements (Requirements 1 and 2) 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. The invariance can hold, however, if the relative difficulty changes as a function of score level in the same way across subpopulations. If, however, the relative difficulty of the two tests interacts with group membership, or there is an interaction among score level, difficulty and group, then invariance will not hold, and the test construction process may be out of control in that markedly nonparallel test forms are being constructed.

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 (SEA) 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). It uses the subpopulation invariance of linking functions across important subgroups (e.g., gender groups) to assess the degree of score exchangeability. SEA 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).

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 paper. 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 (1983), as well as other

studies cited in Kolen and Brennan (2004). A variety of factors could be manipulated in a series of studies that examines the robustness of both newer approaches like kernel equating and older linear and nonlinear methods. Recent work by Sinharay and Holland (2009) is indicative of the kind of work that can be done to better understand the robustness of various procedures to violation of their assumptions.

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 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. 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 paper has been on best practices for score equating. Score equating is only one aspect of the score reporting process. There are other components of the score reporting process that 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.

References

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington DC: American Council on Education.

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9-49). New York, NY: Academic.

Brennan, R. L. (1992). The context of context effects. *Applied Measurement in Education, 5*, 225-264.

Brennan, R. L. (2006). *Educational measurement* (4th ed.). Westport, CT: Praeger.

Brennan, R. L. (2007). Tests in transition: Synthesis and discussion. In N. J. Dorans, M. Pommerich, & P.W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 161-175). New York, NY: Springer-Verlag.

Cook, L. L. (2007). Practical problems in equating test scores: A practitioner's perspective. In N. J. Dorans, M. Pommerich, & P.W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 73-88). New York, NY: Springer-Verlag.

Dorans, N. J. (2002). Recentering and realigning the SAT score distributions: How and why. *Journal of Educational Measurement, 39*(1), 59-84.

Dorans, N. J. (Ed.). (2004). Population invariance [Special issue]. *Journal of Educational Measurement 41*(1).

Dorans, N. J. (2008). *Holland's advice for the fourth generation of test theory: Blood tests can be contests*. Invited paper presented at the Holland's trip: A conference in honor of Paul W. Holland, Princeton, NJ.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37,* 281-306.

Dorans, N. J, & Liu, J. (2009). *Score equity assessment: Development of a prototype analysis using SAT mathematics test data across several administrations* (ETS Research Rep. No. RR-09-08). Princeton, NJ: ETS.

Dorans, N. J., Yu, L., & Guo, F. (2006). *Evaluating scale fit for broad-ranged admissions tests*. (ETS Research Memorandum No. RM-06-04). Princeton, NJ: ETS.

Dorans, N. J., Pommerich, M., & Holland, P.W. (Eds.). (2007). *Linking and aligning scores and scales.* New York, NY: Springer-Verlag.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates.

Fairbank, B. A. (1987). The use of presmoothing and postsmoothing to increase the precision of equipercentile equating. *Applied Psychological Measurement, 11*, 245-262.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (Eds.). (1999). *Uncommon measures: Equivalence and linkage among educational tests* (Report of the Committee on Equivalency and Linkage of Educational Tests, National Research Council). Washington DC: National Academy Press.

Flanagan, J. C. (1951). Units, scores, and norms. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 695-763). Washington DC: American Council on Education.

Haberman, S., Guo, H., Liu, J., & Dorans, N. J. (2008). *Trend analysis in seasonal time series models. Consistency of SAT reasoning score conversions* (ETS Research Rep. No. RR-08-67). Princeton, NJ: ETS.

Hanson, B. A. (1991). A note on Levine's formula for equating unequally reliable tests using data from the common item nonequivalent groups design. *Journal of Educational Statistics, 16,* 93-100.

Harris, D. J. (2007). Practical issues in vertical scaling. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 233-251). New York, NY: Springer-Verlag.

Harris, D. J., & Gao, X. (2003, April). *A conceptual synthesis of context effect.* In *Context effects: Implications for pretesting and CBT.* Symposium conducted at the annual meeting of the American Educational Research Association, Chicago, IL.

Holland, P. W. (1994). Measurements or contests? Comments on Zwick, Bond and Allen/Donoghue. In *Proceedings of the Social Statistics Section of the American Statistical Association,* 27-29. Alexandria, VA: American Statistical Association.

Holland, P. W. (2007). A framework and history for score linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 5-30). New York, NY: Springer-Verlag.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187-220).Westport, CT: Praeger.

Holland P. W., & Hoskens, M. (2003). Classical test theory as a first-order item response theory: application to true-score prediction from a possibly nonparallel test*. Psychometrika, 68,* 123–149.

Holland, P. W., & Thayer, D. T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (ETS Research Rep. No. RR-87-31). Princeton NJ: ETS.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25,* 133–183.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Kolen, M. J. (2004). Linking assessments: Concept and history. *Applied Psychological Measurement, 28*, 219-226.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111-153). Westport, CT: Praeger.

Kolen, M. J. (2007). Data collection designs and linking procedures. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 31-55). New York, NY: Springer-Verlag.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, linking, and scaling: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Kolen, M. J., & Jarjoura, D. (1987). Analytic smoothing for equipercentile equating under the common item nonequivalent populations design. *Psychometrika, 52*, 43-59.

Koretz, D. M., Bertenthal, M. W., & Green, B. F. (Eds.). (1999). *Embedding questions: The pursuit of a common measure in uncommon tests* (Report of the Committee on Embedding Common Test Items in State and District Assessments, National Research Council). Washington DC: National Academy Press.

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: An historical perspective on an immediate concern. *Review of Educational Research, 55,* 387-413.

Liang, L., Dorans, N. J., & Sinharay, S. (2009). *First language of examinees and its relationship to equating* (ETS Research Rep. No. RR-09-05). Princeton, NJ: ETS.

Lindquist, E. F. (Ed.). (1951). *Educational measurement.* Washington, DC: American Council on Education

Linn, R. L. (1989). *Educational measurement* (3rd ed.). New York, NY: Macmillan.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6*(1), 83–102.

Liou, M., Cheng, P. E., & Li, M. Y. (2001). Estimating comparable scores using surrogate variables. *Applied Psychological Measurement*, *25*, 197–207.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken NJ: Wiley.

Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton, NJ: ETS.

Livingston, S. A., & Kim, S. (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement, 46*, 330-343.

Lord, F. M. (1950). *Notes on comparable scales for test scores* (ETS Research Bulletin No. RB-50-48). Princeton, NJ: ETS.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Marco, G. L., Petersen, N. S., & Stewart, E. E. (1983). *A large-scale evaluation of linear and curvilinear score equating models, Volumes I and II* (ETS Research Memorandum No. RM-83-02). Princeton, NJ: ETS.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects* (Policy Information Rep.). Princeton, NJ: ETS.

Moses, T., & Kim, S. (2007). *Reliability and the nonequivalent groups with anchor test design* (ETS Research Rep. No. RR-07-16). Princeton, NJ: ETS.

Patz, R. J., & Yao, L. (2007). Methods and models for vertical scaling. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 253-272). New York, NY: Springer-Verlag.

Peterson, N. S. (2007). Equating: Best practices and challenges to best practices. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 59-72). New York, NY: Springer-Verlag.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221-262). New York, NY: Macmillan.

Pommerich, M., & Dorans, N. J. (Eds.). (2004). Concordance [Special issue]. *Applied Psychological Measurement 28*(4).

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: The Danish Institute for Educational Research.

Rosenbaum, P. R. (1995). *Observational studies*. New York, NY: Springer-Verlag.

Sinharay, S., & Holland, P. W. (2009). *The missing data assumptions of the NEAT design and their implications for test equating* (ETS Research Rep. No. RR-09-16). Princeton, NJ: ETS.

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement, 42,* 309-330.

Thorndike, R. L. (Ed.). (1971). *Educational measurement* (2nd ed.).Washington, DC: American Council on Education.

Tukey, J. W. (1963). Mathematics 596—An introduction to the frequency analysis of time series. In D. R. Brillinger (Ed.), *The collected works of John W. Tucker, Volume I: Time series, 1949-1964.* London, England: Chapman & Hall.

von Davier, A. A. (2007). Potential solutions to practical equating issues. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 89-106). New York, NY: Springer-Verlag.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating.* New York, NY: Springer.

von Davier, A., A., & Liu, M. (Eds.). (2008). Population invariance [Special issue]. *Applied Psychological Measurement 32*(1).

Yen, W. M. (2007). Vertical scaling and no child left behind. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 273-283). New York, NY: Springer-Verlag.

Wright, N. K., & Dorans, N. J. (1993). *Using the selection variable for matching or equating.* (ETS Research Rep. No. RR-93-04). Princeton, NJ: ETS.

Notes

[1] The intended audience for this report is broad, ranging from high-level graduate students to experienced staff who want to compare their current practices with something closer to the ideal practice. As such, treatments that appear cursory to some readers may appear too detailed for others. In addition, these best practices reflect our perspective which might be called *applied theoretical*, a perspective that emphasizes principled practice over expediency and which attempts to achieve the best results possible in achieving the equating that yields the best scaling solution in a given setting. Others may have different perspectives as to what constitutes best practice.

[2] There are cases where anchor items are interspersed within the same section with items that count toward the total score. These items that are administered internally but are external to the total test score are affected by the same issues that affect internal anchor items.

[3] 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.

[4] 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).