Chapter 4
Contributions to Score Linking Theory and Practice

Neil J. Dorans and Gautam Puhan

Test score equating is essential for testing programs that use multiple editions of the same test and for which scores on different editions are expected to have the same meaning. 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 were more difficult than another, test takers would tend 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 test takers.

ETS statisticians and psychometricians have contributed indirectly or directly to the wealth of material in the chapters on score equating or on score linking that have appeared in the four editions of *Educational Measurement*. ETS's extensive involvement with the score equating chapters of these editions of *Educational Measurement* highlights the impact that ETS has had in this important area of psychometrics.

At the time of publication, each of the four editions of *Educational Measurement* represented the state of the art in domains that are essential to the purview of the National Council on Measurement in Education. Experts in each domain wrote a chapter in each edition. Harold Gulliksen was one of the key contributors to the Flanagan (1951) chapter on units, scores, and norms that appeared in the first edition. Several of the issues and problems raised in that first edition are still current, which shows their persistence. Angoff (1971), in the second edition, provided a comprehensive introduction to scales, norms, and test equating. Petersen et al. (1989) introduced new material developed since the Angoff chapter. Holland and Dorans (2006) included a brief review of the history of test score linking. In addition to test equating, Holland and Dorans (2006) discussed other ways that scores on different tests are connected or linked together.

N.J. Dorans (✉) • G. Puhan
Educational Testing Service, Princeton, NJ, USA
e-mail: ndorans@ets.org

R.E. Bennett, M. von Davier (eds.), *Advancing Human Assessment*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-58689-2_4

The purpose of this chapter is to document ETS's involvement with score linking theory and practice. This chapter is not meant to be a book on score equating and score linking.[1] Several books on equating exist; some of these have been authored by ETS staff, as is noted in the last section of this chapter. We do not attempt to summarize all extant research and development pertaining to score equating or score linking. We focus on efforts conducted by ETS staff. We do not attempt to pass judgment on research or synthesize it. Instead, we attempt to describe it in enough detail to pique the interest of the reader and help point him or her in the right direction for further exploration on his or her own. We presume that the reader is familiar enough with the field so as not to be intimidated by the vocabulary that has evolved over the years in this area of specialization so central to ETS's mission to foster fairness and quality.

The particular approach to tackling this documentation task is to cluster studies around different aspects of score linking. Section 4.1 lists several examples of score linking to provide a motivation for the extent of research on score linking. Section 4.2 summarizes published efforts that provide conceptual frameworks of score linking or examples of scale aligning. Section 4.3 deals with data collection designs and data preparation issues. In Sect. 4.4, the focus is on the various procedures that have been developed to link or equate scores. Research describing processes for evaluating the quality of equating results is the focus of Sect. 4.5. Studies that focus on comparing different methods are described in Sect. 4.6. Section 4.7 is a brief chronological summary of the material covered in Sects. 4.2, 4.3, 4.4, 4.5 and 4.6. Section 4.8 contains a summary of the various books and chapters that ETS authors have contributed on the topic of score linking. Section 4.9 contains a concluding comment.

4.1 Why Score Linking Is Important

Two critical ingredients are needed to produce test scores: the test and those who take the test, the test takers. Test scores depend on the blueprint or specifications used to produce the test. The specifications describe the construct that the test is supposed to measure, how the items or components of the test contribute to the measurement of this construct (or constructs), the relative difficulty of these items for the target population of test takers, and how the items and test are scored. The definition of the target population of test takers includes who qualifies as a member of that population and is preferably accompanied by an explanation of why the test

[1] The term *linking* is often used in an IRT context to refer to procedures for aligning item parameter and proficiency metrics from one calibration to another, such as those described by M. von Davier and A. A. von Davier (2007). We do not consider this type of IRT linking in this chapter; it is treated in the chapter by Carlson and von Davier (Chap. 5, this volume). We do, however, address IRT true-score linking in Sect. 4.6.4 and IRT preequating in Sect. 4.4.4.

is appropriate for these test takers and examples of appropriate and inappropriate use.

Whenever scores from two different tests are going to be compared, there is a need to link the scales of the two test scores. The goal of scale aligning is to transform the scores from two different tests onto a common scale. The types of linkages that result depend on whether the test scores being linked measure different constructs or similar constructs, whether the tests are similar or dissimilar in difficulty, and whether the tests are built to similar or different test specifications. We give several practical examples in the following.

When two or more tests that measure different constructs are administered to a common population, the scores for each test may be transformed to have a common distribution for the target population of test takers (i.e., the reference population). The data are responses from (a) administering all the tests to the same sample of test takers or (b) administering the tests to separate, randomly equivalent samples of test takers from the same population. In this way, all of the tests are taken by equivalent groups of test takers from the reference population. One way to define comparable scores is in terms of comparable percentiles in the reference population.

Even though the scales on the different tests are made comparable in this narrow sense, the tests do measure different constructs. The recentering of the *SAT® I* test scale is an example of this type of scale aligning (Dorans 2002a, b). The scales for the SAT Verbal (SAT-V) and SAT Mathematical (SAT-M) scores were redefined so as to give the scaled scores on the SAT-V and SAT-M the same distribution in a reference population of students tested in 1990. The recentered score scales enable a student whose SAT-M score is higher than his or her SAT-V score to conclude that he or she did in fact perform better on the mathematical portion than on the verbal portion, at least in relation to the students tested in 1990.

Tests of skill subjects (e.g., reading) that are targeted for different school grades may be viewed as tests of similar constructs that are intended to differ in difficulty—those for the lower grades being easier than those for the higher grades. It is often desired to put scores from such tests onto a common overall scale so that progress in a given subject, such as mathematics or reading, can be tracked over time. A topic such as mathematics or reading, when considered over a range of school grades, has several subtopics or dimensions. At different grades, potentially different dimensions of these subjects are relevant and tested. For this reason, the constructs being measured by the tests for different grade levels may differ somewhat, but the tests are often similar in reliability.

Sometimes tests that measure the same construct have similar levels of difficulty but differ in reliability (e.g., length). The classic case is scaling the scores of a short form of a test onto the scale of its full or long form.

Sometimes tests to be linked all measure similar constructs, but they are constructed according to different specifications. In most cases, they are similar in test length and reliability. In addition, they often have similar uses and may be taken by the same test takers for the same purpose. Score linking adds value to the scores on both tests by expressing them as if they were scores on the other test. Many colleges and universities accept scores on either the ACT or SAT for the purpose of admissions decisions, and they often have more experience interpreting the results from one of these tests than the other.

Test equating is a necessary part of any testing program that produces new test forms and for which the uses of these tests require the meaning of the score scale be maintained over time. Although they measure the same constructs and are usually built to the same test specifications or test blueprint, different editions or forms of a test almost always differ somewhat in their statistical properties. For example, one form may be harder than another, so without adjustments, test takers would be expected to receive lower scores on this harder form. A primary goal of test equating for testing programs is to eliminate the effects on scores of these unintended differences in test form difficulty. The purpose of equating test scores is to allow the scores from each test to be used interchangeably, as if they had come from the same test. This purpose puts strong requirements on the tests and on the method of score linking. Most of the research described in the following pages focused on this particular form of scale aligning, known as *score equating*.

In the remaining sections of this chapter, we focus on score linking issues for tests that measure characteristics at the level of the individual test taker. Large-scale assessments, which are surveys of groups of test takers, are described in Beaton and Barone (Chap. 8, this volume) and Kirsh et al. (Chap. 9, this volume).

4.2 Conceptual Frameworks for Score Linking

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 (2006) made distinctions between different types of linkages and emphasized 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. Each type of score linking uses either equivalent groups of test takers or common items for linkage purposes. It is essential to understand why these types differ because they can be confused in practice, which can lead to violations of the standards that guide professional practice. Section 4.2.1 describes frameworks used for score linking. Section 4.2.2 contains a discussion of score equating frameworks.

4.2.1 Score Linking Frameworks

Lord (1964a, b) published one of the early articles to focus on the distinction between test forms that are actually or rigorously parallel and test forms that are nominally parallel—those that are built to be parallel but fall short for some reason.

This distinction occurs in most frameworks on score equating. Lord (1980) later went on to say that equating was either unnecessary (rigorously parallel forms) or impossible (everything else).

Mislevy (1992) provided one of the first extensive treatments of different aspects of what he called linking of educational assessments: *equating, calibration, projection, statistical moderation,* and *social moderation.*

Dorans (1999) made distinctions between three types of linkages or score correspondences when evaluating linkages among SAT scores and ACT scores. These were equating, scaling, and prediction. Later, in a special issue of *Applied Psychological Measurement*, edited by Pommerich and Dorans (2004), he used the terms *equating*, *concordance*, and *expectation* to refer to these three types of linkings and provided means for determining which one was most appropriate for a given set of test scores (Dorans 2004b). This framework was elaborated on by Holland and Dorans (2006), who made distinctions between *score equating, scale aligning*, and *predicting*, noting that scale aligning was a broad category that could be further subdivided into subcategories on the basis of differences in the construct assessed, test difficulty, test reliability, and population ability.

Many of the types of score linking cited by Mislevy (1992) and Dorans (1999, 2004b) could be found in the broad area of scale aligning, including concordance, vertical linking, and calibration. This framework was adapted for the public health domain by Dorans (2007) and served as the backbone for the volume on linking and aligning scores and scales by Dorans et al. (2007).

4.2.2 Equating Frameworks

Dorans et al. (2010a) provided an overview of the particular type of score linking called score equating from a perspective of best practices. After defining equating as a special form of score linking, the authors described the most common data collection designs used in the equating of test scores, some common observed-score equating functions, common data-processing practices that occur prior to computations of equating functions, and how to evaluate an equating function.

A.A. von Davier (2003, 2008) and A.A. von Davier and Kong (2005), building on the unified statistical treatment of score equating, known as *kernel equating*, that was introduced by Holland and Thayer (1989) and developed further by A.A. von Davier et al. (2004b), described a new unified framework for linear equating in a nonequivalent groups anchor test design. They employed a common parameterization to show that three linear methods, Tucker, Levine observed score, and chained,[2] can be viewed as special cases of a general linear function. The concept of a method function was introduced to distinguish among the possible forms that a linear equating function might take, in general, and among the three equating methods, in particular. This approach included a general formula for the standard error of equating

[2] These equating methods are described in Sect. 4.4.

for all linear equating functions in the nonequivalent groups anchor test design and advocated the use of the standard error of equating difference (SEED) to investigate if the observed differences in the equating functions are statistically significant.

A.A. von Davier (2013) provided a conceptual framework that encompassed traditional observed-score equating methods, kernel equating methods, and item response theory (IRT) observed-score equating, all of which produce one equating function between two test scores, along with local equating or local linking, which can produce a different linking function between two test scores given a score on a third variable (Wiberg et al. 2014). The notion of multiple conversions between two test scores is a source of controversy (Dorans 2013; Gonzalez and von Davier 2013; Holland 2013; M. von Davier et al. 2013).

4.3 Data Collection Designs and Data Preparation

Data collection and preparation are prerequisites to score linking.

4.3.1 Data Collection

Numerous data collection designs have been used for score linking. To obtain unbiased estimates of test form difficulty differences, all score equating methods must control for differential ability of the test-taker 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, nonstrictly parallel tests are given to two different groups of test takers of unequal ability. Assuming that the samples are large enough to ignore sampling error, differences in the distributions of the resulting scores can be due to one or both of two factors. One factor is the relative difficulty of the two tests, and the other is the relative ability of the two groups of test takers 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, two distinct approaches address the separation of test difficulty and group ability differences. The first approach is to use a common population of test takers so that there are no ability differences. The other approach is to use an anchor measure of the construct being assessed by the tests to be equated. Ideally, the data should come from a large representative sample of motivated test takers that is divided in half either randomly or randomly within strata to achieve equivalent groups. Each half of this sample is administered either the new form or the old form of a test. It is typical to assume that all samples are random samples from populations of interest, even though, in practice, this may be only an approximation. When the same test takers take both tests, we achieve direct control over differential test-taker ability. In practice, it is more common to use two equivalent samples of test takers from a common population instead of identical test takers.

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 test takers. The use of an anchor measure can lead to more flexible data collection designs than those that require common test takers. However, the use of anchor measures requires users to make various assumptions that are not needed when the test takers taking the tests are either the same or from equivalent samples. When there are ability differences between new and old form samples, the various statistical adjustments for ability differences often produce different results because the methods make different assumptions about the relationships of the anchor test score to the scores to be equated. In addition, assumptions are made about the invariance of item characteristics across different locations within the test.

Some studies have attempted to link scores on tests in the absence of either common test material or equivalent groups of test takers. Dorans and Middleton (2012) used the term *presumed linking* to describe these situations. These studies are not discussed here.

It is generally considered good practice to have the anchor test be a mini-version of the total tests being equated. That means it should have the same difficulty and similar content. Often an external anchor is not available, and internal anchors are used. In this case, context effects become a possible issue. To minimize these effects, anchor (or common) items are often placed in the same location within each test. When an anchor test is used, the items should be evaluated via procedures for assessing whether items are functioning in the same way in both the old and new form samples. All items on both total tests are evaluated to see if they are performing as expected. If they are not, it is often a sign of a quality-control problem. More information can be found in Holland and Dorans (2006).

When there are large score differences on the anchor test between samples of test takers given the two different test forms to be equated, equating based on the nonequivalent-groups anchor test design can often become problematic. Accumulation of potentially biased equating results can occur over a chain of prior equatings and lead to a shift in the meaning of numbers on the scores scale.

In practice, the true equating function is never known, so it is wise to look at several procedures that make different assumptions or that use different data. Given the potential impact of the final score conversion on all participants in an assessment process, it is important to check as many factors that can cause problems as possible. Considering multiple conversions is one way to do this.

Whereas many sources, such as Holland and Dorans (2006), have focused on the structure of data collection designs, the amount of data collected 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 ensure this. This fact should always be kept in mind when selecting a data collection design. Section 4.4 describes procedures that have been developed to deal with the threats associated with small samples.

4.3.2 Data Preparation Activities

Prior to equating and other forms of linking, several steps can be taken to improve the quality of the data. These best practices of data preparation often deal with sample selection, smoothing score distributions, excluding outliers, repeaters, and so on. These issues are the focus of the next four parts of this section.

4.3.2.1 Sample Selection

Before conducting the equating analyses, testing programs often filter the data based on certain heuristics. For example, a testing program may choose to exclude test takers who do not attempt a certain number of items on the test. Other programs might exclude test takers based, for example, on repeater status. ETS researchers have conducted studies to examine the effect of such sample selection practices on equating results. Liang et al. (2009) examined whether nonnative speakers of the language in which the test is administered should be excluded and found that this may not be an issue as long as the proportion of nonnative speakers does not change markedly across administrations. Puhan (2009b, 2011c) studied the impact of repeaters in the equating samples and found in the data he examined that inclusion or exclusion of repeaters had very little impact on the final equating results. Similarly, Yang et al. (2011) examined the effect of repeaters on score equating and found no significant effects of repeater performance on score equating for the exam being studied. However, Kim and Walker (2009a, b) found in their study that when the repeater subgroup was subdivided based on the particular form test takers took previously, subgroup equating functions substantially differed from the total-group equating function.

4.3.2.2 Weighted Samples

Dorans (1990c) edited a special issue of *Applied Measurement in Education* that focused on the topic of equating with samples matched on the anchor test score (Dorans 1990a). The studies in that special issue used simulations that varied in the way in which real data were manipulated to produce simulated samples of test takers. These and related studies are described in Sect. 4.6.3.

Other authors used demographic data to achieve a form of matching. Livingston (2014a) proposed the demographically adjusted groups procedure, which uses demographic information about the test takers to transform the groups taking the two different test forms into groups of equal ability by weighting the test takers unequally. Results indicated that although this procedure adjusts for group differences, it does not reduce the ability difference between the new and old form samples enough to warrant use.

Qian et al. (2013) used techniques for weighting observations to yield a weighted sample distribution that is consistent with the target population distribution to achieve true-score equatings that are more invariant across administrations than those obtained with unweighted samples.

Haberman (2015) used adjustment by minimum discriminant information to link test forms in the case of a nonequivalent-groups design in which there are no satisfactory common items. This approach employs background information other than scores on individual test takers in each administration so that weighted samples of test takers form pseudo-equivalent groups in the sense that they resemble samples from equivalent groups.

4.3.2.3 Smoothing

Irregularities in score distributions can produce irregularities in the equipercentile equating adjustment that might not generalize to different groups of test takers because the methods developed for continuous data are applied to discrete data. Therefore it is generally advisable to presmooth the raw-score frequencies in some way prior to equipercentile equating.

The idea of smoothing score distributions prior to equating goes far back to the 1950s. Karon and Cliff (1957) proposed the Cureton–Tukey procedure as a means for reducing sampling error by mathematically smoothing the sample score data before equating. However, the differences among the linear equating method, the equipercentile equating method with no smoothing of the data, and the equipercentile equating method after smoothing by the Cureton–Tukey method were not statistically significant. Nevertheless, this was an important idea, and although Karon and Cliff's results did not show the benefits of smoothing, currently most testing programs using equipercentile equating use some form of pre- or postsmoothing to obtain more stable equating results.

Ever since the smoothing method using loglinear models was adapted by ETS researchers in the 1980s (for details, see Holland and Thayer 1987; Rosenbaum and Thayer 1987) smoothing has been an important component of the equating process. The new millennium saw a renewed interest in smoothing research. Macros using the statistical analysis software SAS loglinear modeling routines were developed at ETS to facilitate research on smoothing (Moses and von Davier 2006, 2013; Moses et al. 2004). A series of studies were conducted to assess selection strategies (e.g., strategies based on likelihood ratio tests, equated score difference tests, Akaike information criterion (AIC) for univariate and bivariate loglinear smoothing models and their effects on equating function accuracy (Moses 2008a, 2009; Moses and Holland 2008, 2009a, b, c, 2010a, b).

Studies also included comparisons of traditional equipercentile equating with various degrees of presmoothing and kernel equating (Moses and Holland 2007) and smoothing approaches for composite scores (Moses 2014) as well as studies that compared smoothing with pseudo-Bayes probability estimates (Moses and Oh 2009).

There has also been an interest in smoothing in the context of systematic irregularities in the score distributions that are due to scoring practice and scaling issues (e.g., formula scoring, impossible scores) rather than random irregularities (J. Liu et al. 2009b; Puhan et al. 2008b, 2010).

4.3.2.4 Small Samples and Smoothing

Presmoothing the data before conducting an equipercentile equating has been shown to reduce error in small-sample equating. For example, Livingston and Feryok (1987) and Livingston (1993b) worked with small samples and found that presmoothing substantially improved the equating results obtained from small samples. Puhan (2011a, b), based on the results of an empirical study, however, concluded that although presmoothing can reduce random equating error, it is not likely to reduce equating bias caused by using an unrepresentative small sample and presented other alternatives to the small-sample equating problem that focused more on improving data collection (see Sect. 4.4.5).

4.4 Score Equating and Score Linking Procedures

Many procedures for equating tests have been developed by ETS researchers. In this section, we consider equating procedures such as linear, equipercentile equating, kernel equating, and IRT true-score linking.[3] Equating procedures developed to equate new forms under special circumstances (e.g., preequating and small-sample equating procedures) are also considered in this section.

[3] We have chosen to use the term *linking* instead of *equating* when it comes to describing the IRT true-score approach that is in wide use. This linking procedure defines the true-score equating that exists between true scores on Test X and true scores on Test Y, which are perfectly related to each other, as both are monotonic transformations of the same IRT proficiency estimate. Typically, this true-score equating is applied to observed scores as if they were true scores. This application produces an observed-score linking that is not likely to yield equated scores, however, as defined by Lord (1980) or Holland and Dorans (2006); hence our deliberate use of linking instead of equating.

4.4.1 Early Equating Procedures

Starting in the 1950s, ETS researchers have made substantial contributions to the equating literature by proposing new methods for equating, procedures for improving existing equating methods, and procedures for evaluating equating results.

Lord (1950) provided a definition of comparability wherein the score scales of two equally reliable tests are considered comparable with respect to a certain group of test takers if the score distributions of the two tests are identical for this group. He provided the basic formulas for equating means and standard deviations (in six different scenarios) to achieve comparability of score scales. Tucker (1951) emphasized the need to establish a formal system within which to consider scaling error due to sampling. Using simple examples, he illustrated possible ways of defining the scaling error confidence range and setting a range for the probability of occurrence of scaling errors due to sampling that would be considered within normal operations. Techniques were developed to investigate whether regressions differ by groups. Schultz and Wilks (1950) presented a technique to adjust for the lack of equivalence in two samples. This technique focused on the intercept differences from the two group regressions of total score onto anchor score obtained under the constraint that the two regressions had the same slope. Koutsopoulos (1961) presented a linear practice effect solution for a counterbalanced case of equating, in which two equally random groups (alpha and beta) take two forms, X and Y, of a test, alpha in the order X, Y and beta in the order Y, X. Gulliksen (1968) presented a variety of solutions for determining the equivalence of two measures, ranging from a criterion for strict interchangeability of scores to factor methods for comparing multifactor batteries of measures and multidimensional scaling. Boldt (1972) laid out an alternative approach to linking scores that involved a principle for choosing objective functions whose optimization would lead to a selection of conversion constants for equating.

Angoff (1953) presented a method of equating test forms of the American Council on Education (ACE) examination by using a miniature version of the full test as an external anchor to equate the test forms. Fan and Swineford (1954) and Swineford and Fan (1957) introduced a method based on item difficulty estimates to equate scores administered under the nonequivalent anchor test design, which the authors claimed produced highly satisfactory results, especially when the two groups taking the two forms were quite different in ability.

Assuming that the new and old forms are equally reliable, Lord (1954, 1955) derived maximum likelihood estimates of the population mean and standard deviation, which were then substituted into the basic formula for linear equating.

Levine (1955) developed two linear equating procedures for the common-item nonequivalent population design. Levine observed-score equating relates observed scores on a new form to the scale of observed scores on an old form. Levine true-score equating equates true scores. Approximately a half-century later, A.A. von Davier et al. (2007) introduced an equipercentile version of the Levine linear observed-score equating function, which is based on assumptions about true scores.

Based on theoretical and empirical results, Chen (2012) showed that linear IRT observed-score linking and Levine observed-score equating for the anchor test design are closely related despite being based on different methodologies. Chen and Livingston (2013) presented a new equating method for the nonequivalent groups with anchor test design: poststratification equating based on true anchor scores. The linear version of this method is shown to be equivalent, under certain conditions, to Levine observed-score equating.

4.4.2 *True-Score Linking*

As noted in the previous section, Levine (1955) also developed the so-called Levine true-score equating procedure that equates true scores.

Lord (1975) compared equating methods based on item characteristic curve (ICC) theory, which he later called item response theory (IRT) in Lord (1980), with nonlinear conventional methods and pointed out the effectiveness of ICC-based methods for increasing stability of the equating near the extremes of the data, reducing scale drift, and preequating. Lord also included a chapter on IRT preequating. (A review of research related to IRT true-score linking appears in Sect. 4.6.4.)

4.4.3 *Kernel Equating and Linking With Continuous Exponential Families*

As noted earlier, Holland and Thayer (1989) introduced the kernel method of equating score distributions. This new method included both linear and standard equipercentile methods as special cases and could be applied under most equating data collection designs.

Within the Kernel equating framework, Chen and Holland (2010) developed a new curvilinear equating for the nonequivalent groups with anchor test (NEAT) design which they called curvilinear Levine observed score equating.

In the context of equivalent-groups design, Haberman (2008a) introduced a new way to continuize discrete distribution functions using exponential families of functions. Application of this linking method was also considered for the single-group design (Haberman 2008b) and the nonequivalent anchor test design (Haberman and Yan 2011). For the nonequivalent groups with anchor test design, this linking method produced very similar results to kernel equating and equipercentile equating with loglinear presmoothing.

4.4.4 Preequating

Preequating has been tried for several ETS programs over the years. Most notably, the computer-adaptive testing algorithm employed for the *GRE*® test, the *TOEFL*® test, and GMAT examination in the 1990s could be viewed as an application of IRT preequating. Since the end of the twentieth century, IRT preequating has been used for the *CLEP*® examination and with the GRE revised General Test introduced in 2011. This section describes observed-score preequating procedures. (The results of several studies that used IRT preequating can be found in Sect. 4.6.5.)

In the 1980s, section preequating was used with the GMAT examination. A preequating procedure was developed for use with small-volume tests, most notably the *PRAXIS*® assessments. This approach is described in Sect. 4.4.5. Holland and Wightman (1982) described a preliminary investigation of a linear section preequating procedure. In this statistical procedure, data collected from equivalent groups via the nonscored variable or experimental section(s) of a test were combined across tests to produce statistics needed for linear preequating of a form composed of these sections. Thayer (1983) described the maximum likelihood estimation procedure used for estimating the joint covariance matrix for sections of tests given to distinct samples of test takers, which was at the heart of the section preequating approach.

Holland and Thayer (1981) applied this procedure to the GRE test and obtained encouraging results. Holland and Thayer (1984, 1985) extended the theory behind section preequating to allow for practice effects on both the old and new forms and, in the process, provided a unified account of the procedure. Wightman and Wightman (1988) examined the effectiveness of this approach when there is only one variable or experimental section of the test, which entailed using different missing data techniques to estimate correlations between sections.

After a long interlude, section preequating with a single variable section was studied again. Guo and Puhan (2014) introduced a method for both linear and nonlinear preequating. Simulations and a real-data application showed the proposed method to be fairly simple and accurate. Zu and Puhan (2014) examined an observed-score preequating procedure based on empirical item response curves, building on work done by Livingston in the early 1980s. The procedure worked reasonably well in the score range that contained the middle 90th percentile of the data, performing as well as the IRT true-score equating procedure.

4.4.5 Small-Sample Procedures

In addition to proposing new methods for test equating in general, ETS researchers have focused on equating under special circumstances, such as equating with very small samples. Because equating with very small samples tends to be less stable, researchers have proposed new approaches that aim to produce more stable equating results under small-sample conditions. For example, Kim et al. (2006, 2007, 2008c, 2011) proposed the synthetic linking function (which is a weighted average of the small-sample equating and the identity function) for small samples and conducted several empirical studies to examine its effectiveness in small-sample conditions. Similarly, the circle-arc equating method, which constrains the equating curve to pass through two prespecified endpoints and an empirically determined middle point, was also proposed for equating with small samples (Livingston and Kim 2008, 2009, 2010a, b) and evaluated in empirical studies by Kim and Livingston (2009, 2010). Finally, Livingston and Lewis (2009) proposed the empirical Bayes approach for equating with small samples whereby prior information comes from equatings of other test forms, with an appropriate adjustment for possible differences in test length. Kim et al. (2008d, 2009) conducted resampling studies to evaluate the effectiveness of the empirical Bayes approach with small samples and found that this approach tends to improve equating accuracy when the sample size is 25 or fewer, provided the prior equatings are accurate.

The studies summarized in the previous paragraph tried to incorporate modifications to existing equating methods to improve equating under small-sample conditions. Their efficacy depends on the correctness of the strong assumptions that they employ to affect their proposed solutions (e.g., the appropriateness of the circle arc or the identity equatings).

Puhan (2011a, b) presented other alternatives to the small-sample equating problem that focused more on improving data collection. One approach would be to implement an equating design whereby data conducive to improved equatings can be collected to help with the small-sample equating problem. An example of such a design developed at ETS is the single-group nearly equivalent test design, or the SiGNET design (Grant 2011), which introduces a new form in stages rather than all at once. The SiGNET design has two primary merits. First, it facilitates the use of a single-group equating design that has the least random equating error of all designs, and second, it allows for the accumulation of data to equate the new form with a larger sample. Puhan et al. (2008a, 2009) conducted a resampling study to compare equatings under the SiGNET and common-item equating designs and found lower equating error for the SiGNET design than for the common-item equating design in very small sample size conditions (e.g., $N = 10$).

4.5 Evaluating Equatings

In this part, we address several topics in the evaluation of links formed by scale alignment or by equatings. Section 4.5.1 describes research on assessing the sampling error of linking functions. In Sect. 4.5.2, we summarize research dealing with measures of the effect size for assessing the invariance of equating and scale-aligning functions over subpopulations of a larger population. Section 4.5.3 is concerned with research that deals with scale continuity.

4.5.1 Sampling Stability of Linking Functions

All data based linking functions are statistical estimates, and they are therefore subject to sampling variability. If a different sample had been taken from the target population, the estimated linking function would have been different. A measure of statistical stability gives an indication of the uncertainty in an estimate that is due to the sample selected. In Sect. 4.5.1.1, we discuss the standard error of equating (SEE). Because the same methods are also used for concordances, battery scaling, vertical scaling, calibration, and some forms of anchor scaling, the SEE is a relevant measure of statistical accuracy for these cases of test score linking as well as for equating.

In Sects. 4.5.1.1 and 4.5.1.2, we concentrate on the basic ideas and large-sample methods for estimating standard error. These estimates of the SEE and related measures are based on the delta method. This means that they are justified as standard error estimates only for large samples and may not be valid in small samples.

4.5.1.1 The Standard Error of Equating

Concern about the sampling error associated with different data collection designs for equating has occupied ETS researchers since the 1950s (e.g., Karon 1956; Lord 1950). The SEE is the oldest measure of the statistical stability of estimated linking functions. The SEE is defined as the conditional standard deviation of the sampling distribution of the equated score for a given raw score over replications of the equating process under similar conditions. We may use the SEE for several purposes. It gives a direct measure of how consistently the equating or linking function is estimated. Using the approximate normality of the estimate, the SEE can be used to form confidence intervals. In addition, comparing the SEE for various data collection designs can indicate the relative advantage some designs have over others for particular sample sizes and other design factors. This can aid in the choice of a data collection design for a specific purpose.

The SEE can provide us with statistical caveats about the instability of linkings based on small samples. As the size of the sample(s) increases, the SEE will decrease. With small samples, there is always the possibility that the estimated linking function is a poor representation of the population linking function.

The earliest work on the SEE is found in Lord (1950) and reproduced in Angoff (1971). These papers were concerned with linear-linking methods and assumed normal distributions of scores. Zu and Yuan (2012) examined estimates for linear equating methods under conditions of nonnormality for the nonequivalent-groups design. Lord (1982b) derived the SEE for the equivalent- and single-group designs for the equipercentile function using linear interpolation for continuization of the linking functions. However, these SEE calculations for the equipercentile function did not take into account the effect of presmoothing, which can produce reductions in the SEE in many cases, as demonstrated by Livingston (1993a). Liou and Cheng

(1995) gave an extensive discussion (including estimation procedures) of the SEE for various versions of the equipercentile function that included the effect of presmoothing. Holland et al. (1989) and Liou et al. (1996, 1997) discussed the SEE for kernel equating for the nonequivalent-groups anchor test design.

A.A. von Davier et al. (2004b) provided a system of statistical accuracy measures for kernel equating for several data collection designs. Their results account for four factors that affect the SEE: (a) the sample sizes; (b) the effect of presmoothing; (c) the data collection design; and (d) the form of the final equating function, including the method of continuization. In addition to the SEE and the SEED (described in Sect. 4.5.1.2), they recommend the use of percent relative error to summarize how closely the moments of the equated score distribution match the target score distribution that it is striving to match. A.A. von Davier and Kong (2005) gave a similar analysis for linear equating in the non-equivalent-groups design.

Lord (1981) derived the asymptotic standard error of a true-score equating by IRT for the anchor test design and illustrated the effect of anchor test length on this SEE. Y. Liu et al. (2008) compared a Markov chain Monte Carlo (MCMC) method and a bootstrap method in the estimation of standard errors of IRT true-score linking. Grouped jackknifing was used by Haberman et al. (2009) to evaluate the stability of equating procedures with respect to sampling error and with respect to changes in anchor selection with illustrations involving the two-parameter logistic (2PL) IRT model.

4.5.1.2 The Standard Error of Equating Difference Between Two Linking Functions

Those who conduct equatings are often interested in the stability of differences between linking functions. A.A. von Davier et al. (2004b) were the first to explicitly consider the standard error of the distribution of the difference between two estimated linking functions, which they called the SEED. For kernel equating methods, using loglinear models to presmooth the data, the same tools used for computing the SEE can be used for the SEED for many interesting comparisons of kernel equating functions. Moses and Zhang (2010, 2011) extended the notion of the SEED to comparisons between kernel linear and traditional linear and equipercentile equating functions, as well.

An important use of the SEED is to compare the linear and nonlinear versions of kernel equating. von Davier et al. (2004b) combined the SEED with a graphical display of the plot of the difference between the two equating functions. In addition to the difference, they added a band of ±2SEED to put a rough bound on how far the two equating functions could differ due to sampling variability. When the difference curve is outside of this band for a substantial number of values of the X-scores, this is evidence that the differences between the two equating functions exceed what might be expected simply due to sampling error. The ±2SEED band is narrower for larger sample sizes and wider for smaller sample sizes.

Duong and von Davier (2012) illustrated the flexibility of the observed-score equating framework and the availability of the SEED in allowing practitioners to compare statistically the equating results from different weighting schemes for distinctive subgroups of the target population.

In the special situation where we wish to compare an estimated equating function to another nonrandom function, for example, the identity function, the SEE plays the role of the SEED. Dorans and Lawrence (1988, 1990) used the SEE to create error bands around the difference plot to determine whether the equating between two section orders of a test was close enough to the identity. Moses (2008a, 2009) examined a variety of approaches for selecting equating functions for the equivalent-groups design and recommended that the likelihood ratio tests of loglinear models and the equated score difference tests be used together to assess equating function differences overall and also at score levels. He also encouraged a consideration of the magnitude of equated score differences with respect to score reporting practices.

In addition to the statistical significance of the difference between the two linking functions (the SEED), it is also useful to examine whether this difference has any important consequences for reported scores. This issue was addressed by Dorans and Feigenbaum (1994) in their notion of a difference that matters (DTM). They called a difference in reported score points a DTM if the testing program considered it to be a difference worth worrying about. This, of course, depends on the test and its uses. If the DTM that is selected is smaller than 2 times an appropriate SEE or SEED, then the sample size may not be sufficient for the purposes that the equating is intended to support.

4.5.2 Measures of the Subpopulation Sensitivity of Score Linking Functions

Neither the SEE nor the SEED gives any information about how different the estimated linking function would be if the data were sampled from other populations of test takers. Methods for checking the sensitivity of linking functions to the population on which they are computed (i.e., subpopulation invariance checks) serve as diagnostics for evaluating links between tests (especially those that are intended to be test equatings). The most common way that population invariance checks are made is on subpopulations of test takers within the larger population from which the samples are drawn. Subgroups such as male and female are often easily identifiable in the data. Other subgroups are those based on ethnicity, region of the country, and so on. In general, it is a good idea to select subgroups that are known to differ in their performance on the tests in question.

Angoff and Cowell (1986) examined the population sensitivity of linear conversions for the GRE Quantitative test (GRE-Q) and the specially constituted GRE Verbal-plus-Quantitative test (GREV+Q) using equivalent groups of approximately

13,000 taking each form. The data clearly supported the assumption of population invariance for GRE-Q but not quite so clearly for GREV+Q.

Dorans and Holland (2000a, b) developed general indices of population invariance/sensitivity of linking functions for the equivalent groups and single-group designs. To study population invariance, they assumed that the target population is partitioned into mutually exclusive and exhaustive subpopulations. A.A. von Davier et al. (2004a) extended that work to the nonequivalent-groups anchor test design that involves two populations, both of which are partitioned into similar subpopulations.

Moses (2006, 2008b) extended the framework of kernel equating to include the standard errors of indices described in Dorans and Holland (2000a, b). The accuracies of the derived standard errors were evaluated with respect to empirical standard errors.

Dorans (2004a) edited a special issue of the *Journal of Educational Measurement*, titled "Assessing the Population Sensitivity of Equating Functions," that examined whether equating or linking functions relating test scores achieved population invariance. A. A. von Davier et al. (2004a) extended the work on subpopulation invariance done by Dorans and Holland (2000a, b) for the single-population case to the two-population case, in which the data are collected on an anchor test as well as the tests to be equated. Yang (2004) examined whether the multiple-choice (MC) to composite linking functions of the *Advanced Placement*® examinations remain invariant over subgroups by region. Dorans (2004c) examined population invariance across gender groups and placed his investigation within a larger fairness context by introducing score equity analysis as another facet of fair assessment, a complement to differential item functioning and differential prediction.

A.A. von Davier and Liu (2007) edited a special issue of *Applied Psychological Measurement*, titled "Population Invariance," that built on and extended prior research on population invariance and examined the use of population invariance measures in a wide variety of practical contexts. A.A. von Davier and Wilson (2008) examined IRT models applied to Advanced Placement exams with both MC and constructed-response (CR) components. M. Liu and Holland (2008) used Law School Admission Test (LSAT) data to extend the application of population invariance methods to subpopulations defined by geographic region, whether test takers applied to law school, and their law school admission status. Yang and Gao (2008) investigated the population invariance of the one-parameter IRT model used with the testlet-based computerized exams that are part of CLEP. Dorans et al. (2008) examined the role that the choice of anchor test plays in achieving population invariance of linear equatings across male and female subpopulations and test administrations.

Rijmen et al. (2009) compared two methods for obtaining the standard errors of two population invariance measures of equating functions. The results indicated little difference between the standard errors found by the delta method and the grouped jackknife method.

Dorans and Liu (2009) provided an extensive illustration of the application of score equity assessment (SEA), a quality-control process built around the use of population invariance indices, to the SAT-M exam. Moses et al. (2009, 2010b) developed a SAS macro that produces Dorans and Liu's (2009) prototypical SEA analyses, including various tabular and graphical analyses of the differences between scaled score conversions from one or more subgroups and the scaled score conversion based on a total group. J. Liu and Dorans (2013) described how SEA can be used as a tool to assess a critical aspect of construct continuity, the equivalence of scores, whenever planned changes are introduced to testing programs. They also described how SEA can be used as a quality-control check to evaluate whether tests developed to a static set of specifications remain within acceptable tolerance levels with respect to equitability.

Kim et al. (2012) illustrated the use of subpopulation invariance with operational data indices to assess whether changes to the test specifications affected the equatability of a redesigned test to the current test enough to change the meaning of points on the score scale. Liang et al. (2009), also reported in Sinharay et al. (2011b), used SEA to examine the sensitivity of equating procedures to increasing numbers of nonnative speakers in equating samples.

4.5.3 Consistency of Scale Score Meaning

In an ideal world, measurement is flawless, and score scales are properly defined and well maintained. Shifts in performance on a test reflect shifts in the ability of test-taker populations, and any variability in the raw-to-scale conversions across editions of a test is minor and due to random sampling error. In an ideal world, many things need to mesh. Reality differs from the ideal in several ways that may contribute to scale inconsistency, which, in turn, may contribute to the appearance or actual existence of scale drift. Among these sources of scale inconsistency are inconsistent or poorly defined test-construction practices, population changes, estimation error associated with small samples of test takers, accumulation of errors over a long sequence of test administrations, inadequate anchor tests, and equating model misfit. Research into scale continuity has become more prevalent in the twenty-first century. Haberman and Dorans (2011) made distinctions among different sources of variation that may contribute to score-scale inconsistency. In the process of delineating these potential sources of scale inconsistency, they indicated practices that are likely either to contribute to inconsistency or to attenuate it.

Haberman (2010) examined the limits placed on scale accuracy by sample size, number of administrations, and number of forms to be equated. He demonstrated analytically that a testing program with a fixed yearly volume is likely to experience more substantial scale drift with many small-volume administrations than with fewer large volume administrations. As a consequence, the comparability of scores across different examinations is likely to be compromised from many small-volume administrations. This loss of comparability has implications for some modes of continuous testing. Guo (2010) investigated the asymptotic accumulative SEE for linear equating methods under the nonequivalent groups with anchor test design. This tool measures the magnitude of equating errors that have accumulated over a series of equatings.

Lee and Haberman (2013) demonstrated how to use harmonic regression to assess scale stability. Lee and von Davier (2013) presented an approach for score-scale monitoring and assessment of scale drift that used quality-control charts and time series techniques for continuous monitoring, adjustment of customary variations, identification of abrupt shifts, and assessment of autocorrelation.

With respect to the SAT scales established in the early 1940s, Modu and Stern (1975) indicated that the reported score scale had drifted by almost 14 points for the verbal section and 17 points for the mathematics section between 1963 and 1973. Petersen et al. (1983) examined scale drift for the verbal and mathematics portions of the SAT and concluded that for reasonably parallel tests, linear equating was adequate, but for tests that differed somewhat in content and length, 3PL IRT-based methods lead to greater stability of equating results. McHale and Ninneman (1994) assessed the stability of the SAT scale from 1973 to 1984 and found that the SAT-V score scale showed little drift. Furthermore, the results from the Mathematics scale were inconsistent, and therefore the stability of this scale could not be determined.

With respect to the revised SAT scales introduced in 1995, Guo et al. (2012) examined the stability of the SAT Reasoning Test score scales from 2005 to 2010. A 2005 old form was administered along with a 2010 new form. Critical Reading and Mathematics score scales experienced, at most, a moderate upward scale drift that might be explained by an accumulation of random equating errors. The Writing score scale experienced a significant upward scale drift, which might reflect more than random error.

Scale stability depends on the number of items or sets of items used to link tests across administrations. J. Liu et al. (2014) examined the effects of using one, two, or three anchor tests on scale stability of the SAT from 1995 to 2003. Equating based on one old form produced persistent scale drift and also showed increased variability in score means and standard deviations over time. In contrast, equating back to two or three old forms produced much more stable conversions and had less variation.

Guo et al. (2013) advocated the use of the conditional standard error of measurement when assessing scale deficiencies as measured by gaps and clumps, which were defined in Dorans et al. (2010b).

Using data from a teacher certification program, Puhan (2007, 2009a) examined scale drift for parallel equating chains and a single long chain. Results of the study indicated that although some drift was observed, the effect on pass or fail status of test takers was not large.

Cook (1988) explored several alternatives to the scaling procedures traditionally used for the College Board Achievement Tests. The author explored additional scaling covariates that might improve scaling results for tests that did not correlate highly with the SAT Reasoning Test, possible respecification of the sample of students used to scale the tests, and possible respecification of the hypothetical scaling population.

4.6 Comparative Studies

As new methods or modifications to existing methods for data preparation and analysis continued to be developed at ETS, studies were conducted to evaluate the new approaches. These studies were diverse and included comparisons between newly developed methods and existing methods, chained versus poststratification methods, comparisons of equatings using different types of anchor tests, and so on. In this section we attempt to summarize this research in a manner that parallels the structure employed in Sects. 4.3 and 4.4. In Sect. 4.6.1, we address research that focused on data collection issues, including comparisons of equivalent-groups equating and anchor test equating and comparisons of the various anchor test equating procedures. Section 4.6.2 contains research pertaining to anchor test properties. In Sect. 4.6.3, we consider research that focused on different types of samples of test takers. Next, in Sect. 4.6.4, we consider research that focused on IRT equating. IRT preequating is considered in Sect. 4.6.5. Then some additional topics are addressed. Section 4.6.6 considers equating tests with CR components. Equating of subscores is considered in Sect. 4.6.7, whereas Sect. 4.6.8 considers equating in the presence of multidimensional data. Because several of the studies addressed in Sect. 4.6 used simulated data, we close with a caveat about the strengths and limitations of relying on simulated data in Sect. 4.6.9.

4.6.1 Different Data Collection Designs and Different Methods

Comparisons between different equating methods (e.g., chained vs. poststratification methods) and different equating designs (e.g., equivalent groups vs. nonequivalent groups with anchor test design) have been of interest for many ETS researchers. (Comparisons that focused on IRT linking are discussed in Sect. 4.6.4.)

Kingston and Holland (1986) compared alternative equating methods for the GRE General Test. They compared the equivalent-groups design with two other designs (i.e., nonequivalent groups with an external anchor test and equivalent groups with a preoperational section) and found that the equivalent groups with preoperational section design produced fairly poor results compared to the other designs.

After Holland and Thayer introduced kernel equating in 1989, Livingston (1993b) conducted a study to compare kernel equating with traditional equating methods and concluded that kernel equating and equipercentile equating based on smoothed score distributions produce very similar results, except at the low end of the score scale, where the kernel results were slightly more accurate. However, much of the research work at ETS comparing kernel equating with traditional equating methods happened after A.A. von Davier et al. (2004b) was published. For example, A.A. von Davier et al. (2006) examined how closely the kernel equating (KE) method approximated the results of other observed-score equating methods under the common-item equating design and found that the results from kernal equating (KE) and the other methods were quite similar. Similarly, results from a study by Mao et al. (2006) indicated that the differences between KE and the traditional equating methods are very small (for most parts of the score scale) for both the equivalent-groups and common-item equating design. J. Liu and Low (2007, 2008) compared kernel equating with analogous traditional equating methods and concluded that KE results are comparable to the results of other methods. Similarly, Grant et al. (2009) compared KE with traditional equating methods, such as Tucker, Levine, chained linear, and chained equipercentile methods, and concluded that the differences between KE and traditional equivalents were quite small. Finally, Lee and von Davier (2008) compared equating results based on different kernel functions and indicated that the equated scores based on different kernel functions do not vary much, except for extreme scores.

There has been renewed interest in chained equating (CE) versus poststratification equating (PSE) research in the new millennium. For example, Guo and Oh (2009) evaluated the frequency estimation (FE) equating method, a PSE method, under different conditions. Based on their results, they recommended FE equating when neither the two forms nor the observed conditional distributions are very different. Puhan (2010a, b) compared Tucker, chained linear, and Levine observed equating under conditions where the new and old form samples were either similar in ability or not and where the tests were built to the same set of content specifications and concluded that, for most conditions, chained linear equating produced fairly accurate equating results. Predictions from both PSE and CE assumptions were compared using data from a special study that used a fairly novel approach (Holland et al. 2006, 2008). This research used real data to simulate tests built to the same set of content specifications and found that that both CE and PSE make very similar predictions but that those of CE are slightly more accurate than those of PSE, especially where the linking function is nonlinear. In a somewhat similar vein as the preceding studies, Puhan (2012) compared Tucker and chained linear equating in two scenarios. In the first scenario, known as rater comparability scoring and equating, chained linear equating produced more accurate results. Note that although rater comparability scoring typically results in a single-group equating design, the study evaluated a special case in which the rater comparability scoring data were used under a common-item equating design. In the second situation, which used a common-item equating design where the new and old form samples were randomly equivalent, Tucker equating produced more accurate results. Oh and Moses (2012) investigated differences between uni- and bidirectional approaches to chained equipercentile equating and concluded that although the bidirectional results were slightly less erratic and smoother, both methods, in general, produce very similar results.

4.6.2 The Role of the Anchor

Studies have examined the effect of different types of anchor tests on test equating, including anchor tests that are different in content and statistical characteristics. For example, Echternacht (1971) compared two approaches (i.e., using common items or scores from the GRE Verbal and Quantitative measures as the anchor) for equating the GRE Advanced tests. Results showed that both approaches produce equating results that are somewhat different from each other. DeMauro (1992) examined the possibility of equating the *TWE*® test by using TOEFL as an anchor and concluded that using TOEFL as an anchor to equate the TWE is not appropriate.

Ricker and von Davier (2007) examined the effects of external anchor test length on equating results for the common-item equating design. Their results indicated that bias tends to increase in the conversions as the anchor test length decreases, although FE and kernel poststratification equating are less sensitive to this change than other equating methods, such as chained equipercentile equating. Zu and Liu (2009, 2010) compared the effect of discrete and passage-based anchor items on common-item equating results and concluded that anchor tests that tend to have more passage-based items than discrete items result in larger equating errors, especially when the new and old samples differ in ability. Liao (2013) evaluated the effect of speededness on common-item equating and concluded that including an item set toward the end of the test in the anchor affects the equating in the anticipated direction, favoring the group for which the test is less speeded.

Moses and Kim (2007) evaluated the impact of unequal reliability on test equating methods in the common-item equating design and noted that unequal and/or low reliability inflates equating function variability and alters equating functions when there is an ability difference between the new and old form samples.

Sinharay and Holland (2006a, b) questioned conventional wisdom that an anchor test used in equating should be a statistical miniature version of the tests to be equated. They found that anchor tests with a spread of item difficulties less than that of a total test (i.e., a midi test) seem to perform as well as a mini test (i.e., a miniature version of the full test), thereby suggesting that the requirement of the anchor test to mimic the statistical characteristics of the total test may not be optimal. Sinharay et al. (2012) also demonstrated theoretically that the mini test may not be the optimal anchor test with respect to the anchor test–total test correlation. Finally, several empirical studies by J. Liu et al. (2009a, 2011a, b) also found that the midi anchor performed as well or better than the mini anchor across most of the score scale, except the top and bottom, which is where inclusion or exclusion of easy or hard items might be expected to have an effect.

For decades, new editions of the SAT were equated back to two past forms using the nonequivalent-groups anchor test design (Holland and Dorans 2006). Successive new test forms were linked back to different pairs of old forms. In 1994, the SAT equatings began to link new forms back to four old forms. The rationale for this new scheme was that with more links to past forms, it is easier to detect a poor past conversion function, and it makes the final new conversion function less reliant on any particular older equating function. Guo et al. (2011) used SAT data collected from 44 administrations to investigate the effect of accumulated equating error in equating conversions and the effect of the use of multiple links in equating. It was observed that the single-link equating conversions drifted further away from the operational four-link conversions as equating results accumulated over time. In addition, the single-link conversions exhibited an instability that was not obvious for the operational data. A statistical random walk model was offered to explain the mechanism of scale drift in equating caused by random equating error. J. Liu et al. (2014) tried to find a balance point where the needs for equating, control of item/form exposure, and pretesting could be satisfied. Three equating scenarios were examined using real data: equating to one old form, equating to two old forms, or equating to three old forms. Equating based on one old form produced persistent score drift and showed increased variability in score means and standard deviations over time. In contrast, equating back to two or three old forms produced much more stable conversions and less variation in means and standard deviations. Overall, equating based on multiple linking designs produced more consistent results and seemed to limit scale drift.

Moses et al. (2010a, 2011) studied three different ways of using two anchors that link the same old and new form tests in the common-item equating design. The overall results of this study suggested that when using two anchors, the poststratification approach works better than the imputation and propensity score matching approaches. Poststratification also produced more accurate SEEDs, quantities that are useful for evaluating competing equating and scaling functions.

4.6.3 Matched-Sample Equating

Equating based on samples with identical anchor score distributions was viewed as a potential solution to the variability seen across equating methods when equating samples of test takers were not equivalent (Dorans 1990c). Cook et al. (1988) discussed the need to equate achievement tests using samples of students who take the new and old forms at comparable points in the school year. Stocking et al. (1988) compared equating results obtained using representative and matched samples and concluded that matching equating samples on the basis of a fallible measure of ability is not advisable for any equating method, except possibly the Tucker equating method. Lawrence and Dorans (1988) compared equating results obtained using a representative old-form sample and an old-form sample matched to the new-form sample (matched sample) and found that results for the five studied equating methods tended to converge under the matched sample condition.

Lawrence and Dorans (1990), using the verbal anchor to create differences from the reference or base population and the pseudo-populations, demonstrated that the poststratification methods did best and the true-score methods did slightly worse than the chained method when the same verbal anchor was used for equating. Eignor et al. (1990a, b) used an IRT model to simulate data and found that the weakest results were obtained for poststratification on the basis of the verbal anchor and that the true-score methods were slightly better than the chained method. Livingston et al. (1990) used SAT-M scores to create differences in populations and examined the equating of SAT-V scores via multiple methods. The poststratification method produced the poorest results. They also compared equating results obtained using representative and matched samples and found that the results for all equating methods in the matched samples were similar to those for the Tucker and FE methods in the representative samples. In a follow-up study, Dorans and Wright (1993) compared equating results obtained using representative samples, samples matched on the basis of the equating set, and samples matched on the basis of a selection variable (i.e., a variable along which subpopulations differ) and indicated that matching on the selection variable improves accuracy over matching on the equating test for all methods. Finally, a study by Schmitt et al. (1990) indicated that matching on an anchor test score provides greater agreement among the results of the various equating procedures studied than were obtained under representative sampling.

4.6.4 Item Response Theory True-Score Linking

IRT true-score linking[4] was first used with TOEFL in 1979. Research on IRT-based linking methods received considerable attention in the 1980s to examine their applicability to other testing programs. ETS researchers have focused on a wide variety of research topics, including studies comparing non-IRT observed-score and IRT-based linking methods (including IRT true-score linking and IRT observed-score equating methods), studies comparing different IRT linking methods, studies examining the consequences of violation of assumptions on IRT equating, and so on. These studies are summarized here.

Marco et al. (1983a) examined the adequacy of various linear and curvilinear (observed-score methods) and ICC (one- and three-parameter logistic) equating models when certain sample and test characteristics were systematically varied. They found the 3PL model to be most consistently accurate. Using TOEFL data, Hicks (1983, 1984) evaluated three IRT variants and three conventional equating methods (Tucker, Levine and equipercentile) in terms of scale stability and found that the true-score IRT linking based on scaling by fixing the *b* parameters produces the least discrepant results. Lord and Wingersky (1983, 1984) compared IRT true-score linking with equipercentile equating using observed scores and concluded that the two methods yield almost identical results.

Douglass et al. (1985) studied the extent to which three approximations to the 3PL model could be used in item parameter estimation and equating. Although

[4] Several of the earlier studies cited in this section used the phrase IRT equating to describe the application of an IRT true-score equating function to linking two sets of observed scores. We are using the word linking because this procedure does not ensure that the linked scores are interchangeable in the sense described by Lord (1980) and Holland and Dorans (2006).

these approximations yielded accurate results (based on their circular equating criteria), the authors recommended further research before these methods are used operationally. Boldt (1993) compared linking based on the 3PL IRT model and a modified Rasch model (common nonzero lower asymptote) and concluded that the 3PL model should not be used if sample sizes are small. Tang et al. (1993) compared the performance of the computer programs LOGIST and BILOG (see Carlson and von Davier, Chap. 5, this volume, for more on these programs) on TOEFL 3PL IRT-based linking. The results indicated that the BILOG estimates were closer to the true parameter values in small-sample conditions. In a simulation study, Y. Li (2012) examined the effect of drifted (i.e., items performing differently than the remaining anchor items) polytomous anchor items on the test characteristic curve (TCC) linking and IRT true-score linking. Results indicated that drifted polytomous items have a relatively large impact on the linking results and that, in general, excluding drifted polytomous items from the anchor results in an improvement in equating results.

Kingston et al. (1985) compared IRT linking to conventional equating of the GMAT and concluded that violation of local independence had a negligible effect on the linking results. Cook and Eignor (1985) indicated that it was feasible to use IRT to link the four College Board Achievement tests used in their study. Similarly, McKinley and Kingston (1987) investigated the use of IRT linking for the GRE Subject Test in Mathematics and indicated that IRT linking was feasible for this test. McKinley and Schaefer (1989) conducted a simulation study to evaluate the feasibility of using IRT linking to reduce test form overlap of the GRE Subject Test in Mathematics. They compared double-part IRT true-score linking (i.e., linking to two old forms) with 20-item common-item blocks to triple-part linking (i.e., linking to three old forms) with 10-item common-item blocks. On the basis of the results of their study, they suggested using more than two links.

Cook and Petersen (1987) summarized a series of ETS articles and papers produced in the 1980s that examined how equating is affected by sampling errors, sample characteristics, and the nature of anchor items, among other factors. This summary added greatly to our understanding of the uses of IRT and conventional equating methods in suboptimal situations encountered in practice. Cook and Eignor (1989, 1991) wrote articles and instructional modules that provided a basis for understanding the process of score equating through the use of IRT. They discussed the merits of different IRT equating approaches.

A.A. von Davier and Wilson (2005, 2007) used data from the *Advanced Placement Program*® examinations to investigate the assumptions made by IRT true-score linking method and discussed the approaches for checking whether these assumptions are met for a particular data set. They provided a step-by-step check of how well the assumptions of IRT true-score linking are met. They also compared equating results obtained using IRT as well as traditional methods and showed that IRT and chained equipercentile equating results were close for most of the score range.

D. Li et al. (2012) compared the IRT true-score equating to chained equipercentile equating and observed that the sample variances for the chained equipercentile equating were much smaller than the variances for the IRT true-score equating, except at low scores.

4.6.5 Item Response Theory Preequating Research

In the early 1980s, IRT was evaluated for its potential in preequating tests developed from item pools. Bejar and Wingersky (1981) conducted a feasibility study for preequating the TWE and concluded that the procedure did not exhibit problems beyond those already associated with using IRT on this exam. Eignor (1985) examined the extent to which item parameters estimated on SAT-V and SAT-M pretest data could be used for equating purposes. The preequating results were mixed; three of the four equatings examined were marginally acceptable at best. Hypotheses for these results were posited by the author. Eignor and Stocking (1986) studied these hypotheses in a follow-up investigation and concluded that there was a problem either with the SAT-M data or the way in which LOGIST calibrated items under the 3PL model. Further hypotheses were generated. Stocking and Eignor (1986) investigated these results further and concluded that difference in ability across samples and multidimensionality may have accounted for the lack of item parameter invariance that undermined the preequating effort. While the SAT rejected the use of preequating on the basis of this research, during the 1990s, other testing programs moved to test administration and scoring designs, such as computer-adaptive testing, that relied on even more restrictive invariance assumptions than those that did not hold in the SAT studies.

Gao et al. (2012) investigated whether IRT true-score preequating results based on a Rasch model agreed with equating results based on observed operational data (postequating) for CLEP. The findings varied from subject to subject. Differences among the equating results were attributed to the manner of pretesting, contextual/order effects, or the violations of IRT assumptions. Davey and Lee (2011) examined the potential effect of item position on item parameter and ability estimates for the GRE revised General Test, which would use preequating to link scores obtained via its two-stage testing model. In an effort to mitigate the impact of position effects, they recommended that questions be pretested in random locations throughout the test. They also recommended considering the impact of speededness in the design of the revised test because multistage tests are more subject to speededness compared to linear forms of the same length and testing time.

4.6.6 Equating Tests With Constructed-Response Items

Large-scale testing programs often include CR as well as MC items on their tests. Livingston (2014b) listed some characteristics of CR tests (i.e., small number of tasks and possible raw scores, tasks that are easy to remember and require judgment for scoring) that cause problems when equating scores obtained from CR tests. Through the years, ETS researchers have tried to come up with innovative solutions to equating CR tests effectively.

When a CR test form is reused, raw scores from the two administrations of the form may not be comparable due to two different sets of raters among other reasons. The solution to this problem requires a rescoring, at the new administration, of test-taker responses from a previous administration. The scores from this "rescoring" are used as an anchor for equating, and this process is referred to as rater comparability scoring and equating (Puhan 2013b). Puhan (2013a, b) challenged conventional wisdom and showed theoretically and empirically that the choice of target population weights (for poststratification equating) has a predictable impact on final equating results obtained under the rater comparability scoring and equating scenario. The same author also indicated that chained linear equating produces more accurate equating results than Tucker equating under this equating scenario (Puhan 2012).

Kim et al. (2008a, b, 2010a, b) have compared various designs for equating CR-only tests, such as using an anchor test containing either common CR items or rescored common CR items or an external MC test and an equivalent-groups design incorporating rescored CR items (no anchor test). Results of their studies showed that the use of CR items without rescoring results in much larger bias than the other designs. Similarly, they have compared various designs for equating tests containing both MC and CR items such as using an anchor test containing only MC items, both MC and CR items, both MC and rescored CR items, and an equivalent-groups design incorporating rescored CR items (no anchor test). Results of their studies indicated that using either MC items alone or a mixed anchor without CR item rescoring results in much larger bias than the other two designs and that the equivalent-groups design with rescoring results in the smallest bias. Walker and Kim (2010) examined the use of an all-MC anchor for linking mixed-format tests containing both MC and CR items in a nonequivalent-groups design. They concluded that a MC-only anchor could effectively link two such test forms if either the MC or CR portion of the test measured the same knowledge and skills and if the relationship between the MC portion and the total test remained constant across the new and reference linking groups.

Because subpopulation invariance is considered a desirable property for equating relationships, Kim and Walker (2009b, 2012a) examined the appropriateness of the anchor composition in a mixed-format test, which includes both MC and CR items, using subpopulation invariance indices. They found that the mixed anchor was a better choice than the MC-only anchor to achieve subpopulation invariance between males and females. Muraki et al. (2000) provided an excellent summary describing issues and developments in linking performance assessments and included comparisons of common linking designs (single group, equivalent groups, nonequivalent groups) and linking methodologies (traditional and IRT).

Myford et al. (1995) pilot-tested a quality-control procedure for monitoring and adjusting for differences in reader performance and discussed steps that might enable different administrations of the TWE to be equated. Tan et al. (2010)

compared equating results using different sample sizes and equating designs (i.e., single group vs. common-item equating designs) to examine the possibility of reducing the rescoring sample. Similarly, Kim and Moses (2013) conducted a study to evaluate the conditions under which single scoring for CR items is as effective as double scoring in a licensure testing context. Results of their study indicated that under the conditions they examined, the use of single scoring would reduce scoring time and cost without increasing classification inconsistency. Y. Li and Brown (2013) conducted a rater comparability scoring and equating study and concluded that raters maintained the same scoring standards across administrations for the CRs in the *TOEFL iBT*® test Speaking and Writing sections. They recommended that the TOEFL iBT program use this procedure as a tool to periodically monitor Speaking and Writing scoring.

Some testing programs require all test takers to complete the same common portion of a test but offer a choice of essays in another portion of the test. Obviously there can be a fairness issue if the different essays vary in difficulty. ETS researchers have come up with innovative procedures whereby the scores on the alternate questions can be adjusted based on the estimated total group mean and standard deviation or score distribution on each alternate question (Cowell 1972; Rosenbaum 1985). According to Livingston (1988), these procedures tend to make larger adjustments when the scores to be adjusted are less correlated with scores on the common portion. He therefore suggested an adjustment procedure that makes smaller adjustments when the correlation between the scores to be adjusted and the scores on the common portion is low. Allen et al. (1993) examined Livingston's proposal, which they demonstrate to be consistent with certain missing data assumptions, and compared its adjustments to those from procedures that make different kinds of assumptions about the missing data that occur with essay choice.

In an experimental study, Wang et al. (1995) asked students to identify which items within three pairs of MC items they would prefer to answer, and the students were required to answer both items in each of the three pairs. The authors concluded that allowing choice will only produce fair tests when it is not necessary to allow choice. Although this study used tests with MC items only and involved small numbers of items and test takers, it attempted to answer via an experiment a question similar to what the other, earlier discussed studies attempted to answer, namely, making adjustments for test-taker choice among questions.

The same authors attempted to equate tests that allowed choice of questions by using existing IRT models and the assumption that the ICCs for the items obtained from test takers who chose to answer them are the same as the ICCs that would be obtained from the test takers who did not answer them (Wainer et al. 1991, 1994). Wainer and Thissen (1994) discussed several issues pertaining to tests that allow a choice to test takers. They provided examples where equating such tests is impossible and where allowing choice does not necessarily elicit the test takers' best performance.

4.6.7 Subscores

The demand for subscores has been increasing for a number of reasons, including the desire of candidates who fail the test to know their strengths and weaknesses in different content areas and because of mandates by legislatures to report subscores. Furthermore, states and academic institutions such as colleges and universities want a profile of performance for their graduates to better evaluate their training and focus on areas that need remediation. However, for subscores to be reported operationally, they should be comparable across the different forms of a test. One way to achieve comparability is to equate the subscores.

Sinharay and Haberman (2011a, b) proposed several approaches for equating augmented subscores (i.e., a linear combination of a subscore and the total score) under the nonequivalent groups with anchor test design. These approaches only differ in the way the anchor score is defined (e.g., using subscore, total score or augmented subscore as the anchor). They concluded that these approaches performed quite accurately under most practical situations, although using the total score or augmented subscore as the anchor performed slightly better than using only the subscore as the anchor. Puhan and Liang (2011a, b) considered equating subscores using internal common items or total scaled scores as the anchor and concluded that using total scaled scores as the anchor is preferable, especially when the internal common items are small.

4.6.8 Multidimensionality and Equating

The call for CR items and subscores on MC tests reflects a shared belief that a total score based on MC items underrepresents the construct of interest. This suggests that more than one dimension may exist in the data.

ETS researchers such as Cook et al. (1985) examined the relationship between violations of the assumption of unidimensionality and the quality of IRT true-score equating. Dorans and Kingston (Dorans and Kingston 1985; Kingston and Dorans 1982) examined the consequences of violations of unidimensionality assumptions on IRT equating and noted that although violations of unidimensionality may have an impact on equating, the effect may not be substantial. Using data from the LSAT, Camilli et al. (1995) examined the effect of multidimensionality on equating and concluded that violations of unidimensionality may not have a substantial impact on estimated item parameters and true-score equating tables. Dorans et al. (2014) did a comparative study where they varied content structure and correlation between underlying dimensions to examine their effect on latent-score and observed-score linking results. They demonstrated analytically and with simulated data that score equating is possible with multidimensional tests, provided the tests are parallel in content structure.

4.6.9 A Caveat on Comparative Studies

Sinharay and Holland (2008, 2010a, b) demonstrated that the equating method with explicit or implicit assumptions most consistent with the model used to generate the data performs best with those simulated data. When they compared three equating methods—the FE equipercentile equating method, the chained equipercentile equating method, and the IRT observed-score equating method—each one worked best in data consistent with its assumptions. The chained equipercentile equating method was never the worst performer. These studies by Sinharay and Holland provide a valuable lens from which to view the simulation studies summarized in Sect. 4.6 whether they used data simulated from a model or real test data to construct simulated scenarios: The results of the simulation follow from the design of the simulation. As Dorans (2014) noted, simulation studies may be helpful in studying the strengths and weakness of methods but cannot be used as a substitute for analysis of real data.

4.7 The Ebb and Flow of Equating Research at ETS

In this section, we provide a high-level summary of the ebb and flow of equating research reported in Sects. 4.2, 4.3, 4.5, and 4.6. We divide the period from 1947, the birth of ETS, through 2015 into four periods: (a) before 1970, (b) 1970s to mid-1980s, (c) mid-1980s to 2000, and (d) 2001–2015.

4.7.1 Prior to 1970

As might be expected, much of the early research on equating was procedural as many methods were introduced, including those named after Tucker and Levine (Sect. 4.4.1). Lord attended to the SEE (Sect. 4.5.1.1). There were early efforts to smooth data from small samples (Sect. 4.3.2.3). With the exception of work done by Lord in 1964, distinctions between equating and other forms of what is now called score linking did not seem to be made (Sect. 4.2.1).

4.7.2 The Year 1970 to the Mid-1980s

Equating research took on new importance in the late 1970s and early 1980s as test disclosure legislation led to the creation of many more test forms in a testing program than had been needed in the predisclosure period. This required novel data collection designs and led to the investigation of preequating approaches. Lord introduced his equating requirements (Sect. 4.2.1) and concurrently introduced IRT score linking methods, which became the subject of much research (Sects. 4.4.2 and 4.6.4). Lord estimated the SEE for IRT (Sect. 4.5.1.1). IRT preequating research was prevalent and generally discouraging (Sect. 4.6.5). Holland and his colleagues introduced section preequating (section 4.4.4) as another preequating solution to the problems posed by the test disclosure legislation.

4.7.3 The Mid-1980s to 2000

Equating research was more dormant in this period, as first differential item functioning and then computer-adaptive testing garnered much of the research funding at ETS. While some work was motivated by practice, such as matched-sample equating research (Sect. 4.6.3) and continued investigations of IRT score linking (Sect. 4.6.4), there were developments of theoretical import. Most notable among these were the development of kernel equating by Holland and his colleagues (Sects. 4.4.3 and 4.6.1), which led to much research about its use in estimating standard errors (Sect. 4.5.1.1). Claims made by some that scores from a variety of sources could be used interchangeably led to the development of cogent frameworks for distinguishing between different kinds of score linkings (Sect. 4.2.1). The role of dimensionality in equating was studied (Sect. 4.6.8).

4.7.4 The Years 2002–2015

The twenty-first century witnessed a surge of equating research. The kernel equating method and its use in estimating standard errors was studied extensively (Sects. 4.4.3, 4.5.1, 4.5.2, and 4.6.1). A new equating method was proposed by Haberman (Sect. 4.4.3).

Data collection and preparation received renewed interest in the areas of sample selection (Sect. 4.3.2.1) and weighting of samples (Sect. 4.3.2.2). A considerable amount of work was done on smoothing (Sect. 4.3.2.3), mostly by Moses and Holland and their colleagues. Livingston and Puhan and their colleagues devoted much attention to developing small-sample equating methods (Sect. 4.4.5).

CE was the focus of many comparative investigations (Sect. 4.6.1). The anchor continued to receive attention (Sect. 4.6.2). Equating subscores became an important issue as there were more and more calls to extract information from less and less (Sect. 4.6.7). The comparability problems faced by reliance on subjectively scored CR items began to be addressed (Sect. 4.6.6). The role of dimensionality in equating was examined again (Sect. 4.6.8).

Holland and Dorans provided a detailed framework for classes of linking (Sect. 4.2.1) as a further response to calls for linkages among scores from a variety of sources. Central to that framework was the litmus test of population invariance, which led to an area of research that uses equating to assess the fairness of test scores across subgroups (Sect. 4.5.2).

4.8 Books and Chapters

Books and chapters can be viewed as evidence that the authors are perceived as possessing expertise that is worth sharing with the profession. We conclude this chapter by citing the various books and chapters that have been authored by ETS staff in the area of score linking, and then we allude to work in related fields and forecast our expectation that ETS will continue to work the issues in this area.

An early treatment of score equating appeared in Gulliksen (1950), who described, among other things, Ledyard R Tucker's proposed use of an anchor test to adjust for differences in the abilities of samples. Tucker proposed this approach to deal with score equating problems with the SAT that occurred when the SAT started to be administered more than once a year to test takers applying to college. Books that dealt exclusively with score equating did not appear for more than 30 years, until the volume edited by ETS researchers Holland and Rubin (1982) was published. The 1980s was the first decade in which much progress was made in score equating research, spearheaded in large part by Paul Holland and his colleagues.

During the 1990s, ETS turned its attention first toward differential item functioning (Dorans, Chap. 7, this volume) and then toward CR and computer-adaptive testing. The latter two directions posed particular challenges to ensuring comparability of measurements, leaning more on strong assumptions than on an empirical basis. After a relatively dormant period in the 1990s, score equating research blossomed in the twenty-first century. Holland and his colleagues played major roles in this rebirth. The Dorans and Holland (2000a, b) article on the population sensitivity of score linking functions marked the beginning of a renaissance of effort on score equating research at ETS.

With the exception of early chapters by Angoff (1967, 1971), most chapters on equating prior to 2000 appeared between 1981 and 1990. Several appeared in the aforementioned Holland and Rubin (1982). Angoff (1981) provided a summary of procedures in use at ETS up until that time. Braun and Holland (1982) provided a formal mathematical framework to examine several observed-score equating procedures used at ETS at that time. Cowell (1982) presented an early application of IRT true-score linking, which was also described in a chapter by Lord (1982a). Holland and Wightman (1982) described a preliminary investigation of a linear section pre-equating procedure. Petersen et al. (1982) summarized the linear equating portion of a massive simulation study that examined linear and curvilinear methods of anchor test equating, ranging from widely used methods to rather obscure methods. Some anchors were external (did not count toward the score), whereas others were internal. They examined different types of content for the internal anchor. Anchors varied in difficulty. In addition, equating samples were randomly equivalent, similar, or dissimilar in ability. Rock (1982) explored how equating could be represented from the perspective of confirmatory factor analysis. Rubin (1982) commented on the chapter by Braun and Holland, whereas Rubin and Szatrowski (1982) critiqued the preequating chapter.

ETS researchers contributed chapters related to equating and linking in edited volumes other than Holland and Rubin's (1982). Angoff (1981) discussed equating and equity in a volume on new directions in testing and measurement circa 1980. Marco (1981) discussed the efforts of test disclosure on score equating in a volume on coaching, disclosure, and ethnic bias. Marco et al. (1983b) published the curvilinear equating analogue to their linear equating chapter that appeared in Holland and Rubin (1982) in a volume on latent trait theory and computer-adaptive testing. Cook and Eignor (1983) addressed the practical considerations associated with using IRT to equate or link test scores in a volume on IRT. Dorans (1990b) produced a chapter on scaling and equating in a volume on computer-adaptive testing edited by Wainer et al. (1990). Angoff and Cook (1988) linked scores across languages by relating the SAT to the College Board *PAA*™ test in a chapter on access and assessment for Hispanic students.

Since 2000, ETS authors have produced several books on the topics of score equating and score linking, including two quite different books, the theory-oriented unified statistical treatment of score equating by A.A. von Davier et al. (2004b) and an introduction to the basic concepts of equating by Livingston (2004). A.A. von Davier et al. (2004b) focused on a single method of test equating (i.e., kernel equating) in a unifying way that introduces several new ideas of general use in test equating. Livingston (2004) is a lively and straightforward account of many of the major issues and techniques. Livingston (2014b) is an updated version of his 2004 publication.

In addition to these two equating books were two edited volumes, one by Dorans et al. (2007) and one by A.A. von Davier (2011c). ETS authors contributed several chapters to both of these volumes.

There were six integrated parts to the volume *Linking and Aligning Scores and Scales* by Dorans et al. (2007). The first part set the stage for the remainder of the volume. Holland (2007) noted that linking scores or scales from different tests has a history about as long as the field of psychometrics itself. His chapter included a typology of linking methods that distinguishes among predicting, scaling, and equating. In the second part of the book, Cook (2007) considered some of the daunting challenges facing practitioners and discussed three major stumbling blocks encountered when attempting to equate scores on tests under difficult conditions: characteristics of the tests to be equated, characteristics of the groups used for equating, and characteristics of the anchor tests. A. A. von Davier (2007) addressed potential future directions for improving equating practices and included a brief introduction to kernel equating and issues surrounding assessment of the population sensitivity of equating functions. Educational testing programs in a state of transition were considered in the third part of the volume. J. Liu and Walker (2007) addressed score linking issues associated with content changes to a test. Eignor (2007) discussed linkings between test scores obtained under different modes of administration, noting why scores from computer-adaptive tests and paper-and-pencil tests cannot be considered equated. Concordances between tests built for a common purpose but in different ways were discussed by Dorans and Walker (2007) in a whimsical chapter that was part of the fourth part of the volume, which dealt with concordances. Yen (2007) examined the role of vertical scaling in the pre–No Child Left Behind (NCLB) era and the NCLB era in the fifth part, which was dedicated to vertical scaling. The sixth part dealt with relating the results obtained by surveys of educational achievement that provide aggregate results to tests designed to assess individual test takers. Braun and Qian (2007) modified and evaluated a procedure developed to link state standards to the National Assessment of Educational Progress scale and illustrated its use. In the book's postscript, Dorans et al. (2007) peered into the future and speculated about the likelihood that more and more linkages of dubious merit would be sought.

The A.A. von Davier (2011c) volume titled *Statistical Models for Test Equating, Scaling and Linking*, which received the American Educational Research Association 2013 best publication award, covered a wide domain of topics. Several chapters in the book addressed score linking and equating issues. In the introductory chapter of the book, A.A. von Davier (2011a) described the equating process as a feature of complex statistical models used for measuring abilities in standardized assessments and proposed a framework for observed-score equating methods. Dorans et al. (2011) emphasized the practical aspects of the equating process, the need for a solid data collection design for equating, and the challenges involved in applying specific equating procedures. Carlson (2011) addressed how to link vertically the results of tests that are constructed to intentionally differ in difficulty and content and that are taken by groups of test takers who differ in ability. Holland and Strawderman (2011) described a procedure that might be considered for averaging equating conversions that come from linkings to multiple old forms. Livingston and Kim (2011) addressed different approaches to dealing with the problems associated with equating test scores in small samples. Haberman (2011b) described the use of exponential families for continuizing test score distributions. Lee and von Davier (2011) discussed how various continuous variables with distributions (normal, logistic, and uniform) can be used as kernels to continuize test score distributions. Chen et al. (2011) described new hybrid models within the kernel equating framework, including a nonlinear version of Levine linear equating. Sinharay et al. (2011a) presented a detailed investigation of the untestable assumptions behind two popular nonlinear equating methods used with a nonequivalent-groups design. Rijmen et al. (2011) applied the SEE difference developed by A.A. von Davier et al. (2004b) to the full vector of equated raw scores and constructed a test for testing linear hypotheses about the equating results. D. Li et al. (2011) proposed the use of time series methods for monitoring the stability of reported scores over a long sequence of administrations.

ETS researchers contributed chapters related to equating and linking in edited volumes other than Dorans et al. (2007) and A. A. von Davier (2011c). Dorans (2000) produced a chapter on scaling and equating in a volume on computer-adaptive testing edited by Wainer et al. (2000). In a chapter in a volume dedicated to examining the adaptation of tests from one language to another, Cook and Schmitt-Cascallar (2005) reviewed different approaches to establishing score linkages on tests that are administered in different languages to different populations and critiqued three attempts to link the English-language SAT to the Spanish-language PAA over a 25-year period, including Angoff and Cook (1988) and Cascallar and Dorans (2005). In volume 26 of the *Handbook of Statistics*, dedicated to psychometrics and edited by Rao and Sinharay (2007), Holland et al. (2007) provided an introduction to test score equating, its data collection procedures, and methods used for equating. They also presented sound practices in the choice and evaluation of equating designs and functions and discussed challenges often encountered in practice.

Dorans and Sinharay (2011) edited a volume dedicated to feting the career of Paul Holland, titled *Looking Back*, in which the introductory chapter by Haberman (2011a) listed score equating as but one of Holland's many contributions. Three chapters on score equating were included in that volume. These three authors joined Holland and other ETS researchers in promoting the rebirth of equating research at ETS. Moses (2011) focused on one of Holland's far-reaching applications: his application of loglinear models as a smoothing method for equipercentile equating. Sinharay (2011) discussed the results of several studies that compared the performances of the poststratification equipercentile and chained equipercentile equating methods. Holland was involved in several of these studies. In a book chapter, A. A. von Davier (2011b) focused on the statistical methods available for equating test forms from standardized educational assessments that report scores at the individual level.

4.9 Concluding Comment

Lord (1980) stated that score equating is either not needed or impossible. Scores will be compared, however. As noted by Dorans and Holland (2000a),

> The comparability of measurements made in differing circumstances by different methods and investigators is a fundamental pre-condition for all of science. Psychological and educational measurement is no exception to this rule. Test equating techniques are those statistical and psychometric methods used to adjust scores obtained on different tests measuring the same construct so that they are comparable. (p. 281)

Procedures will attempt to facilitate these comparisons.

As in any scientific endeavor, instrument preparation and data collection are critical. With large equivalent groups of motivated test takers taking essentially parallel forms, the ideal of "no need to equate" is within reach. Score equating methods converge. As samples get small or contain unmotivated test takers or test takers with preknowledge of the test material, or as test takers take un-pretested tests that differ in content and difficulty, equating will be elusive. Researchers in the past have suggested solutions for suboptimal conditions. They will continue to do so in the future. We hope this compilation of studies will be valuable for future researchers who grapple with the inevitable less-than-ideal circumstances they will face when linking score scales or attempting to produce interchangeable scores via score equating.

References

Allen, N. L., Holland, P. W., & Thayer, D. T. (1993). *The optional essay problem and the hypothesis of equal difficulty* (Research Report No. RR-93-40). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1993.tb01551.x

Angoff, W. H. (1953). *Equating of the ACE psychological examinations for high school students* (Research Bulletin No. RB-53-03). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1953.tb00887.x

Angoff, W. H. (1967). Technical problems of obtaining equivalent scores on tests. In W. A. Mehrens & R. L. Ebel (Eds.), *Principles of educational and psychological measurement: A book of selected readings* (pp. 84–86). Chicago: Rand McNally.

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.

Angoff, W. H. (1981). Equating and equity. *New Directions for Testing and Measurement, 9*, 15–20.

Angoff, W. H., & Cook, L. L. (1988). *Equating the scores on the "Prueba de Apitud Academica" and the "Scholastic Aptitude Test"* (College Board Report No. 88-2). New York: College Board. http://dx.doi.org/10.1002/j.2330-8516.1988.tb00259.x

Angoff, W. H., & Cowell, W. R. (1986). An examination of the assumption that the equating of parallel forms is population-independent. *Journal of Educational Measurement, 23*, 327–345. https://doi.org/10.1111/j.1745-3984.1986.tb00253.x

Bejar, I. I., & Wingersky, M. S. (1981). *An application of item response theory to equating the Test of Standard Written English* (Research Report No. RR-81-35). Princeton: Educational Testing Service.

Boldt, R. F. (1972). *Anchored scaling and equating: Old conceptual problems and new methods* (Research Bulletin No. RB-72-28). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1972.tb01025.x

Boldt, R. F. (1993). *Simulated equating using several item response curves* (Research Report No. RR-93-57). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1993.tb01568.x

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: Academic Press.

Braun, H. I., & Qian, J. (2007). An enhanced method for mapping state standards onto the NAEP scale. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 313–338). New York: Springer. https://doi.org/10.1007/978-0-387-49771-6_17

Camilli, G., Wang, M.-M., & Fesq, J. (1995). The effects of dimensionality on equating the Law School Admission Test. *Journal of Educational Measurement, 32*, 79–96. https://doi.org/10.1111/j.1745-3984.1995.tb00457.x

Carlson, J. (2011). Statistical models for vertical linking. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 59–70). New York: Springer. https://doi.org/10.1007/978-0-387-98138-3_4

Cascallar, A. S., & Dorans, N. J. (2005). Linking scores from tests of similar content given in different languages: An illustration involving methodological alternatives. *International Journal of Testing, 5*, 337–356. https://doi.org/10.1207/s15327574ijt0504_1

Chen, H. (2012). A comparison between linear IRT observed score equating and Levine observed score equating under the generalized kernel equating framework. *Journal of Educational Measurement, 49*, 269–284. https://doi.org/10.1111/j.1745-3984.2012.00175.x

Chen, H., & Holland, P. (2010). New equating methods and their relationships with Levine observed score linear equating under the kernel equating framework. *Psychometrika, 75*, 542–557. https://doi.org/10.1007/s11336-010-9171-7

Chen, H., & Livingston, S. A. (2013). *Poststratification equating based on true anchor scores and its relationship to Levine observed score equating* (Research Report No. RR-13-11). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2013.tb02318.x

Chen, H., Livingston, S. A., & Holland, P. W. (2011). Generalized equating functions for NEAT designs. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 89–107). New York: Springer. https://doi.org/10.1007/978-0-387-98138-3_12

Cook, L. L. (1988). *Achievement test scaling* (Research Report No. RR-88-34). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1988.tb00290.x

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: Springer. https://doi.org/10.1007/978-0-387-49771-6_5

Cook, L. L., & Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 175–195). Vancouver: Educational Research Institute of British Columbia.

Cook, L. L., & Eignor, D. R. (1985). *An investigation of the feasibility of applying item response theory to equate achievement tests* (Research Report No. RR-85-31). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1985.tb00116.x

Cook, L. L., & Eignor, D. R. (1989). Using item response theory in test score equating. *International Journal of Educational Research, 13*, 161–173. https://doi.org/10.1016/0883-0355(89)90004-9

Cook, L. L., & Eignor, D. R. (1991). IRT equating methods. *Educational Measurement: Issues and Practice, 10*, 37–45. https://doi.org/10.1111/j.1745-3992.1991.tb00207.x

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11*, 225–244. https://doi.org/10.1177/014662168701100302

Cook, L. L., & Schmitt-Cascallar, A. P. (2005). Establishing score comparability for tests given in different languages. In R. Hambleton, P. F. Meranda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 171–192). Mahwah: Erlbaum.

Cook, L. L., Dorans, N. J., Eignor, D. R., & Petersen, N. S. (1985). *An assessment of the relationship between the assumption of unidimensionality and the quality of IRT true-score equating* (Research Report No. RR-85-30). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1985.tb00115.x

Cook, L. L., Eignor, D. R., & Schmitt, A. P. (1988). *The effects on IRT and conventional achievement test equating results of using equating samples matched on ability* (Research Report No. RR-88-52). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1988.tb00308.x

Cowell, W. R. (1972). *A technique for equating essay question scores* (Statistical Report No. SR-72-70). Princeton: Educational Testing Service.

Cowell, W. R. (1982). Item-response-theory pre-equating in the TOEFL testing program. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 149–161). New York: Academic Press.

Davey, T., & Lee, Y.-H. (2011). *Potential impact of context effects on the scoring and equating of the multistage GRE revised General Test* (Research Report No. RR-11-26). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02262.x

DeMauro, G. E. (1992). *An investigation of the appropriateness of the TOEFL test as a matching variable to equate TWE topics* (Research Report No. RR-92-26). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1992.tb01457.x

Dorans, N. J. (1990a). The equating methods and sampling designs. *Applied Measurement in Education, 3*, 3–17. https://doi.org/10.1207/s15324818ame0301_2

Dorans, N. J. (1990b). Scaling and equating. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 137–160). Hillsdale: Erlbaum.

Dorans, N. J. (Ed.). (1990c). Selecting samples for equating: To match or not to match [Special issue]. *Applied Measurement in Education, 3*(1).

Dorans, N. J. (1999). *Correspondences between ACT and SAT I scores* (College Board Report No. 99-1). New York: College Board. https://doi.org/10.1002/j.2333-8504.1999.tb01800.x

Dorans, N. J. (2000). Scaling and equating. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (2nd ed., pp. 135–158). Hillsdale: Erlbaum.

Dorans, N. J. (2002a). *The recentering of SAT scales and its effects on score distributions and score interpretations* (College Board Research Report No. 2002-11). New York: College Board. https://doi.org/10.1002/j.2333-8504.2002.tb01871.x

Dorans, N. J. (2002b). Recentering the SAT score distributions: How and why. *Journal of Educational Measurement, 39*(1), 59–84. https://doi.org/10.1111/j.1745-3984.2002.tb01135.x

Dorans, N. J. (Ed.). (2004a). Assessing the population sensitivity of equating functions. [Special issue]. *Journal of Educational Measurement, 41*(1).

Dorans, N. J. (2004b). Equating, concordance and expectation. *Applied Psychological Measurement, 28*, 227–246. https://doi.org/10.1177/0146621604265031

Dorans, N. J. (2004c). Using population invariance to assess test score equity. *Journal of Educational Measurement, 41*, 43–68. https://doi.org/10.1111/j.1745-3984.2004.tb01158.x

Dorans, N. J. (2007). Linking scores from multiple health outcome instruments. *Quality of Life Research, 16*(S1), S85–S94. https://doi.org/10.1007/s11136-006-9155-3

Dorans, N. J. (2013). On attempting to do what Lord said was impossible: Commentary on van der Linden's conceptual issues in observed-score equating. *Journal of Educational Measurement, 50*, 304–314. https://doi.org/10.1111/jedm.12017

Dorans, N. J. (2014). *Simulate to understand models, not nature* (Research Report No. RR-14-16). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12013

Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and *PSAT/NMSQT*®. In I. M. Lawrence, N. J. Dorans, M. D. Feigenbaum, N. J. Feryok, A. P. Schmitt, & N. K. Wright (Eds.), *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (Research Memorandum No. RM-94-10). Princeton: Educational Testing Service.

Dorans, N. J., & Holland, P. W. (2000a). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*, 281–306. https://doi.org/10.1111/j.1745-3984.2000.tb01088.x

Dorans, N. J., & Holland, P. W. (2000b). *Population invariance and the equatability of tests: Basic theory and the linear case* (Research Report No. RR-00-19). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2000.tb01842.x

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement, 22*, 249–262. https://doi.org/10.1111/j.1745-3984.1985.tb01062.x

Dorans, N. J., & Lawrence, I. M. (1988). *Checking the equivalence of nearly identical test forms* (Research Report No. RR-88-06). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1988.tb00262.x

Dorans, N. J., & Lawrence, I. M. (1990). Checking the statistical equivalence of nearly identical test editions. *Applied Measurement in Education, 3*, 245–254. https://doi.org/10.1207/s15324818ame0303_3

Dorans, N. J., & Liu, J. (2009). *Score equity assessment: Development of a prototype analysis using SAT Mathematics test data across several administrations* (Research Report No. RR-09-08). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02165.x

Dorans, N. J., & Middleton, K. (2012). Addressing the extreme assumptions of presumed linkings. *Journal of Educational Measurement, 49*, 1–18. https://doi.org/10.1111/j.1745-3984.2011.00157.x

Dorans, N. J., & Sinharay, S. (Eds.). (2011). *Looking back: Proceedings of a conference in honor of Paul W. Holland*. New York: Springer. https://doi.org/10.1007/978-1-4419-9389-2

Dorans, N. J., & Walker, M. E. (2007). Sizing up linkages. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 179–198). New York: Springer. https://doi.org/10.1007/978-0-387-49771-6_10

Dorans, N. J., & Wright, N. K. (1993). *Using the selection variable for matching or equating* (Research Report No. RR-93-04). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1993.tb01515.x

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

Dorans, N. J., Liu, J., & Hammond, S. (2008). Anchor test type and population invariance: An exploration across subpopulations and test administrations. *Applied Psychological Measurement, 32*, 81–97. https://doi.org/10.1177/0146621607311580

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010a). *Principles and practices of test score equating* (Research Report No. RR-10-29). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2010.tb02236.x

Dorans, N. J., Liang, L., & Puhan, G. (2010b). *Aligning scales of certification tests* (Research Report No. RR-10-07). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2010.tb02214.x

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2011). Equating test scores: Towards best practices. In A. A. von Davier (Ed.), *Statistical models for scaling, equating and linking* (pp. 21–42). New York: Springer. https://doi.org/10.1007/978-0-387-98138-3_2

Dorans, N. J., Lin, P., Wang, W., & Yao, L. (2014). *The invariance of latent and observed linking functions in the presence of multiple latent test-taker dimensions* (Research Report No. RR-14-41). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12041

Douglass, J. B., Marco, G. L., & Wingersky, M. S. (1985). *An evaluation of three approximate item response theory models for equating test scores* (Research Report No. RR-85-46). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1985.tb00131.x

Duong, M., & von Davier, A. A. (2012). Observed-score equating with a heterogeneous target population. *International Journal of Testing, 12*, 224–251. https://doi.org/10.1080/15305058.2011.620725

Echternacht, G. (1971). *Alternate methods of equating GRE advanced tests* (GRE Board Professional Report No. GREB No. 69-2P). Princeton: Educational Testing Service.

Eignor, D. R. (1985). *An investigation of the feasibility and practical outcomes of the pre-equating of the SAT Verbal and Mathematical sections* (Research Report No. RR-85-10). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1985.tb00095.x

Eignor, D. R. (2007). Linking scores derived under different modes of test administration. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 135–159). New York: Springer. https://doi.org/10.1007/978-0-387-49771-6_8

Eignor, D. R., & Stocking, M. L. (1986). *An investigation of the possible causes of the inadequacy of IRT pre-equating* (Research Report No. RR-86-14). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1986.tb00169.x

Eignor, D. R., Stocking, M. L., & Cook, L. L. (1990a). *The effect on observed- and true-score equating procedures of matching on a fallible criterion: A simulation with test variation* (Research Report No. RR-90-25). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1990.tb01361.x

Eignor, D. R., Stocking, M. L., & Cook, L. L. (1990b). Simulation results of the effects on linear and curvilinear observed- and true-score equating procedures of matching on a fallible criterion. *Applied Measurement in Education, 3*, 37–55. https://doi.org/10.1207/s15324818ame0301_4

Fan, C. T., & Swineford, F. (1954). *A method of score conversion through item statistics* (Research Bulletin No. RB-54-16). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1954.tb00243.x

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

Gao, R., He, W., & Ruan, C. (2012). *Does preequating work? An investigation into a preequated testlet-based college placement exam using postadministration data* (Research Report No. RR-12-12). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2012.tb02294.x

Gonzalez, J., & von Davier, M. (2013). Statistical models and inference for the true equating transformation in the context of local equating. *Journal of Educational Measurement, 50*, 315–320. https://doi.org/10.1111/jedm.12018

Grant, M. C. (2011). *The single group with nearly equivalent tests (SiGNET) design for equating very small volume multiple-choice tests* (Research Report No. RR-11-31). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02267.x

Grant, M. C., Zhang, Y., & Damiano, M. (2009). *An evaluation of kernel equating: Parallel equating with classical methods in the SAT Subject tests program* (Research Report No. RR-09-06). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02163.x

Gulliksen, H. (1950). *Theory of mental test scores*. New York: Wiley. https://doi.org/10.1037/13240-000

Gulliksen, H. (1968). Methods for determining equivalence of measures. *Psychological Bulletin, 70*, 534–544. https://doi.org/10.1037/h0026721

Guo, H. (2010). Accumulative equating error after a chain of linear equatings. *Psychometrika, 75*, 438–453. https://doi.org/10.1007/s11336-010-9160-x

Guo, H., & Oh, H.-J. (2009). *A study of frequency estimation equipercentile equating when there are large ability differences* (Research Report No. RR-09-45). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02202.x

Guo, H., & Puhan, G. (2014). Section pre-equating under the equivalent groups design without IRT. *Journal of Educational Measurement, 51*, 301–317. https://doi.org/10.1111/jedm.12049

Guo, H., Liu, J., Dorans, N., & Feigenbaum, M. (2011). *Multiple linking in equating and random scale drift* (Research Report No. RR-11-46). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02282.x

Guo, H., Liu, J., Curley, E., Dorans, N., & Feigenbaum, M. (2012). *The stability of the score scale for the SAT Reasoning Test from 2005–2012* (Research Report No. RR-12-15). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2012.tb02297.x

Guo, H., Puhan, G., & Walker, M. E. (2013). *A criterion to evaluate the individual raw-to-scale equating conversions* (Research Report No. RR-13-05). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2013.tb02312.x

Haberman, S. J. (2008a). *Continuous exponential families: An equating tool* (Research Report No. RR-08-05). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02091.x

Haberman, S. J. (2008b). *Linking with continuous exponential families: Single-group designs* (Research Report No. RR-08-61). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02147.x

Haberman, S. (2010). *Limits on the accuracy of linking* (Research Report No. RR-10-22). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2010.tb02229.x

Haberman, S. J. (2011a). The contributions of Paul Holland. In N. J. Dorans & S. Sinharay (Eds.), *Looking back: Proceedings of a conference in honor of Paul W. Holland* (pp. 3–17). New York: Springer. https://doi.org/10.1007/978-1-4419-9389-2_1

Haberman, S. J. (2011b). Using exponential families for equating. In A. A. von Davier (Ed.), *Statistical models for scaling, equating and linking* (pp. 125–140). New York: Springer.

Haberman, S. J. (2015). Pseudo-equivalent groups and linking. *Journal of Educational and Behavioral Statistics, 40*, 254–273. https://doi.org/10.3102/1076998615574772

Haberman, S. J., & Dorans, N. J. (2011). *Sources of scale inconsistency* (Research Report No. RR-11-10). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02246.x

Haberman, S. J., & Yan, D. (2011). *Use of continuous exponential families to link forms via anchor tests* (Research Report No. RR-11-11), Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02247.x

Haberman, S. J., Lee, Y.-H., & Qian, J. (2009). *Jackknifing techniques for evaluation of equating accuracy* (Research Report No. RR-09-39). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02196.x

Hicks, M. M. (1983). True score equating by fixed b's scaling: A flexible and stable equating alternative. *Applied Psychological Measurement, 7*, 255–266. https://doi.org/10.1177/014662168300700302

Hicks, M. M. (1984). *A comparative study of methods of equating TOEFL test scores* (Research Report No. RR-84-20). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1984.tb00060.x

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: Springer. https://doi.org/10.1007/978-0-387-49771-6_2

Holland, P. W. (2013). Comments on van der Linden's critique and proposal for equating. *Journal of Educational Measurement, 50*, 286–294. https://doi.org/10.1111/jedm.12015

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

Holland, P. W., & Rubin, D. B. (Eds.). (1982). *Test equating*. New York: Academic Press.

Holland, P. W., & Strawderman, W. (2011). How to average equating functions, if you must. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 89–107). New York: Springer.

Holland, P. W., & Thayer, D. T. (1981). *Section pre-equating the graduate record examination* (Program Statistics Research Technical Report No. 81-51). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1981.tb01278.x

Holland, P. W., & Thayer, D. T. (1984). *Section pre-equating in the presence of practice effects* (Research Report No. RR-84-07). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1984.tb00047.x

Holland, P. W., & Thayer, D. T. (1985). Section pre-equating in the presence of practice effects. *Journal of Educational Statistics, 10*, 109–120. https://doi.org/10.2307/1164838

Holland, P. W., & Thayer, D. T. (1987). *Notes on the use of log-linear model for fitting discrete probability distribution* (Research Report No. RR-87-31). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1987.tb00235.x

Holland, P. W., & Thayer, D. T. (1989). *The kernel method of equating score distributions* (Program Statistics Research Technical Report No. 89-84).Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1981.tb01278.x

Holland, P. W., & Wightman, L. E. (1982). Section pre-equating. A preliminary investigation. In P. W. Holland & D. B. Rubin (Eds.), *Testing equating* (pp. 217–297). New York: Academic Press.

Holland, P. W., King, B. F., & Thayer, D. T. (1989). *The standard error of equating for the kernel method of equating score distributions* (Research Report No. RR-89-06). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1989.tb00332.x

Holland, P. W., von Davier, A. A., Sinharay, S., & Han, N. (2006). *Testing the untestable assumptions of the chain and poststratification equating methods for the NEAT design* (Research Report No. RR-06-17). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2006.tb02023.x

Holland, P. W., Dorans, N. J., & Petersen, N. S. (2007). Equating test scores. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26, Psychometrics* (pp. 169–203). Amsterdam: Elsevier.

Holland, P. W., Sinharay, S., von Davier, A. A., & Han, N. (2008). An approach to evaluating the missing data assumptions of the chain and post-stratification equating methods for the NEAT design. *Journal of Educational Measurement, 45*(1), 17–43. https://doi.org/10.1111/j.1745-3984.2007.00050.x

Karon, B. P. (1956). The stability of equated test scores. *Journal of Experimental Education, 24*, 181–195. https://doi.org/10.1080/00220973.1956.11010539

Karon, B. P., & Cliff, R. H. (1957). *The Cureton–Tukey method of equating test scores* (Research Bulletin No. RB-57-06). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1957.tb00072.x

Kim, S., & Livingston, S. A. (2009). *Methods of linking with small samples in a common-item design: An empirical comparison* (Research Report No. RR-09-38). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02195.x

Kim, S., & Livingston, S. A. (2010). Comparisons among small sample equating methods in a common-item design. *Journal of Educational Measurement, 47*, 286–298. https://doi.org/10.1111/j.1745-3984.2010.00114.x

Kim, S., & Moses, T. P. (2013). Determining when single scoring for constructed-response items is as effective as double scoring in mixed-format licensure tests. *International Journal of Testing, 13*, 314–328. https://doi.org/10.1080/15305058.2013.776050

Kim, S., & Walker, M. E. (2009a). *Effect of repeaters on score equating in a large scale licensure test* (Research Report No. RR-09-27). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02184.x

Kim, S., & Walker, M. E. (2009b). *Evaluating subpopulation invariance of linking functions to determine the anchor composition for a mixed-format test* (Research Report No. RR-09-36). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02193.x

Kim, S., & Walker, M. E. (2012a). Determining the anchor composition for a mixed-format test: Evaluation of subpopulation invariance of linking functions. *Applied Measurement in Education, 25*, 178–195. https://doi.org/10.1080/08957347.2010.524720

Kim, S., & Walker, M. E. (2012b). Examining repeater effects on chained equipercentile equating with common anchor items. *Applied Measurement in Education, 25*, 41–57. https://doi.org/10.1080/08957347.2012.635481

Kim, S., von Davier, A. A., & Haberman, S. J. (2006). *An alternative to equating with small samples in the non-equivalent groups anchor test design* (Research Report No. RR-06-27). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2006.tb02033.x

Kim, S., von Davier, A. A., & Haberman, S. (2007). *Investigating the effectiveness of a synthetic linking function on small sample equating* (Research Report No. RR-07-37). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2007.tb02079.x

Kim, S., Walker, M. E., & McHale, F. (2008a). *Comparisons among designs for equating constructed response tests* (Research Report No. RR-08-53). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02139.x

Kim, S., Walker, M. E., & McHale, F. (2008b). *Equating of mixed-format tests in large-scale assessments* (Research Report No. RR-08-26). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02112.x

Kim, S., von Davier, A. A., & Haberman, S. (2008c). Small-sample equating using a synthetic linking function. *Journal of Educational Measurement, 45*, 325–342. https://doi.org/10.1111/j.1745-3984.2008.00068.x

Kim, S., Livingston, S. A., & Lewis, C. (2008d). *Investigating the effectiveness of collateral information on small-sample equating.* (Research Report No. RR-08-52). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02138.x

Kim, S., Livingston, S. A., & Lewis, C. (2009). *Evaluating sources of collateral information on small-sample equating* (Research Report No. RR-09-14). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02171.x

Kim, S., Walker, M. E., & McHale, F. (2010a). Comparisons among designs for equating mixed-format tests in large-scale assessments. *Journal of Educational Measurement, 47*, 36–53. https://doi.org/10.1111/j.1745-3984.2009.00098.x

Kim, S., Walker, M. E., & McHale, F. (2010b). Investigation the effectiveness of equating designs for constructed response tests in large scale assessment. *Journal of Educational Measurement, 47*, 186–201. https://doi.org/10.1111/j.1745-3984.2010.00108.x

Kim, S., von Davier, A. A., & Haberman, S. (2011). Practical application of a synthetic linking function on small sample equating. *Applied Measurement in Education, 24*, 95–114. http://dx.doi.org/10.1080/08957347.2011.554601

Kim, S., Walker, M. E., & Larkin, K. (2012). Examining possible construct changes to a licensure test by evaluating equating requirements. *International Journal of Testing, 12*, 365–381. https://doi.org/10.1080/15305058.2011.645974

Kingston, N. M., & Dorans, N. J. (1982). *The feasibility of using item response theory as a psychometric model for the GRE aptitude test* (Research Report No. RR-82-12). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1982.tb01298.x

Kingston, N. M., & Holland, P. W. (1986). *Alternative methods for equating the GRE general test* (Research Report No. RR-86-16). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1986.tb00171.x

Kingston, N. M., Leary, L. F., & Wightman, L. E. (1985). *An exploratory study of the applicability of item response theory methods to the Graduate Management Admission Test* (Research Report No. RR-85-34). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1985.tb00119.x

Koutsopoulos, C. J. (1961). *A linear practice effect solution for the counterbalanced case of equating* (Research Bulletin No. RB-61-19). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1961.tb00287.x

Lawrence, I. M., & Dorans, N. J. (1988). *A comparison of observed score and true score equating methods for representative samples and samples matched on an anchor test* (Research Report No. RR-88-23). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1988.tb00279.x

Lawrence, I. M., & Dorans, N. J. (1990). Effect on equating results of matching samples on an anchor test. *Applied Measurement in Education, 3*, 19–36. https://doi.org/10.1207/s15324818ame0301_3

Lee, Y.-H., & Haberman, S. H. (2013). Harmonic regression and scale stability. *Psychometrika, 78*, 815–829. https://doi.org/10.1007/s11336-013-9337-1

Lee, Y.-H., & von Davier, A. A. (2008). *Comparing alternative kernels for the kernel method of test equating: Gaussian, logistic, and uniform kernels* (Research Report No. RR-08-12). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02098.x

Lee, Y.-H., & von Davier, A. A. (2011). Equating through alternative kernels. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 159–173). New York: Springer. https://doi.org/10.1007/978-0-387-98138-3_10

Lee, Y.-H., & von Davier, A. A. (2013). Monitoring scale scores over time via quality control charts, model-based approaches, and time series techniques. *Psychometrika, 78*, 557–575. https://doi.org/10.1007/s11336-013-9317-5

Levine, R. (1955). *Equating the score scales of alternate forms administered to samples of different ability* (Research Bulletin No. RB-55-23). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1955.tb00266.x

Li, D., Li, S., & von Davier, A. A. (2011). Applying time-series analysis to detect scale drift. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 327–346). New York: Springer. https://doi.org/10.1007/978-0-387-98138-3_20

Li, D., Jiang, Y., & von Davier, A. A. (2012). The accuracy and consistency of a series of IRT true score equatings. *Journal of Educational Measurement, 49*, 167–189. https://doi.org/10.1111/j.1745-3984.2012.00167.x

Li, Y. (2012). *Examining the impact of drifted polytomous anchor items on test characteristic curve (TCC) linking and IRT true score equating* (Research Report No. RR-12-09). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2012.tb02291.x

Li, Y., & Brown, T. (2013). *A trend-scoring study for the TOEFL iBT Speaking and Writing Sections* (Research Memorandum No. RM-13-05). Princeton: Educational Testing Service.

Liang, L., Dorans, N. J., & Sinharay, S. (2009). *First language of examinees and its relationship to equating* (Research Report No. RR-09-05). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02162.x

Liao, C. (2013). *An evaluation of differential speededness and its impact on the common item equating of a reading test* (Research Memorandum No. RM-13-02). Princeton: Educational Testing Service.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6*, 83–102. https://doi.org/10.1207/s15324818ame0601_5

Liou, M., & Cheng, P. E. (1995). Asymptotic standard error of equipercentile equating. *Journal of Educational and Behavioral Statistics, 20*, 259–286. https://doi.org/10.3102/10769986020003259

Liou, M., Cheng, P. E., & Johnson, E. G. (1996). *Standard errors of the kernel equating methods under the common-item design* (Research Report No. RR-96-11). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1996.tb01689.x

Liou, M., Cheng, P. E., & Johnson, E. G. (1997). Standard errors of the kernel equating methods under the common-item design. *Applied Psychological Measurement, 21*, 349–369. https://doi.org/10.1177/01466216970214005

Liu, J., & Dorans, N. J. (2013). Assessing a critical aspect of construct continuity when test specifications change or test forms deviate from specifications. *Educational Measurement: Issues and Practice, 32*, 15–22. https://doi.org/10.1111/emip.12001

Liu, M., & Holland, P. W. (2008). Exploring population sensitivity of linking functions across three law school admission test administrations. *Applied Psychological Measurement, 32*, 27–44. https://doi.org/10.1177/0146621607311576.

Liu, J., & Low, A. C. (2007). *An exploration of kernel equating using SAT data: Equating to a similar population and to a distant population* (Research Report No. RR-07-17). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2007.tb02059.x

Liu, J., & Low, A. (2008). A comparison of the kernel equating method with the traditional equating methods using SAT data. *Journal of Educational Measurement, 45*, 309–323. https://doi.org/10.1111/j.1745-3984.2008.00067.x

Liu, J., & Walker, M. E. (2007). Score linking issues related to test content changes. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 109–134). New York: Springer. https://doi.org/10.1007/978-0-387-49771-6_7

Liu, J., Sinharay, S., Holland, P., Curley, E., & Feigenbaum, M. (2009a). *The effect of different types of anchor test on observed score equating* (Research Report No. RR-09-41). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02198.x

Liu, J., Moses, T. P., & Low, A. C. (2009b). *Evaluation of the effects of loglinear smoothing models on equating functions in the presence of structured data irregularities* (Research Report No. RR-09-22). Princeton*:* Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02179.x

Liu, J., Sinharay, S., Holland, P. W., Curley, E., & Feigenbaum, M. (2011a). Observed score equating using a mini-version anchor and an anchor with less spread of difficulty: A comparison study. *Educational and Psychological Measurement, 71*, 346–361. https://doi.org/10.1177/0013164410375571

Liu, J., Sinharay, S., Holland, P., Curley, E., & Feigenbaum, M. (2011b). Test score equating using a mini-version anchor and a midi anchor: A case study using SAT data. *Journal of Educational Measurement, 48*, 361–379. https://doi.org/10.1111/j.1745-3984.2011.00150.x

Liu, J., Guo, H., & Dorans, N. J. (2014). *A comparison of raw-to-scale conversion consistency between single- and multiple-linking using a nonequivalent groups anchor test design*

(Research Report No. RR-14-13). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12014

Liu, M., & Holland, P. W. (2008). Exploring population sensitivity of linking functions across three law school admission test administrations. *Applied Psychological Measurement, 32*, 27–44. https://doi.org/10.1177/0146621607311576.

Liu, Y., Shultz, E. M., & Yu, L. (2008). Standard error estimation of 3PL IRT true score equating with an MCMC method. *Journal of Educational and Behavioral Statistics, 33*, 257–278. https://doi.org/10.3102/1076998607306076.

Livingston, S. A. (1988). *Adjusting scores on examinations offering a choice of essay questions* (Research Report No. RR-88-64). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1988.tb00320.x

Livingston, S. A. (1993a). *An empirical tryout of kernel equating* (Research Report No. RR-93-33). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1993.tb01544.x

Livingston, S. A. (1993b). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement, 30*, 23–39. https://doi.org/10.1111/j.1745-3984.1993.tb00420.x.

Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton: Educational Testing Service.

Livingston, S. A. (2014a). *Demographically adjusted groups for equating test scores* (Research Report No. RR-14-30). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12030

Livingston, S. A. (2014b). *Equating test scores (without IRT)* (2nd ed.). Princeton: Educational Testing Service.

Livingston, S. A., & Feryok, N. J. (1987). *Univariate versus bivariate smoothing in frequency estimation equating* (Research Report No. RR-87-36). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1987.tb00240.x

Livingston, S. A., & Kim, S. (2008). *Small sample equating by the circle-arc method* (Research Report No. RR-08-39). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02125.x

Livingston, S. A., & Kim, S. (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement, 46*, 330–343. https://doi.org/10.1111/j.1745-3984.2009.00084.x

Livingston, S. A., & Kim, S. (2010a). *An empirical comparison of methods for equating with randomly equivalent groups of 50 to 400 test takers* (Research Report No. RR-10-05). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2010.tb02212.x

Livingston, S. A., & Kim, S. (2010b). Random-groups equating with samples of 50 to 400 test takers. *Journal of Educational Measurement, 47*, 175–185. https://doi.org/10.1111/j.1745-3984.2010.00107.x

Livingston, S. A., & Kim, S. (2011). New approaches to equating with small samples. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 109–122). New York: Springer. https://doi.org/10.1007/978-0-387-98138-3_7

Livingston, S. A., & Lewis, C. (2009). *Small-sample equating with prior information* (Research Report No. RR-09-25). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02182.x

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education, 3*, 73–95. https://doi.org/10.1207/s15324818ame0301_6

Lord, F. M. (1950). *Notes on comparable scales for test scores* (Research Bulletin No. RB-50-48). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1950.tb00673.x

Lord, F. M. (1954). *Equating test scores: The maximum likelihood solution for a common item equating problem* (Research Bulletin No. RB-54-01). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1954.tb00040.x

Lord, F. M. (1955). Equating test scores: A maximum likelihood solution. *Psychometrika, 20*, 193–200. https://doi.org/10.1007/BF02289016

Lord, F. M. (1964a). Nominally and rigorously parallel test forms. *Psychometrika, 29*, 335–345. https://doi.org/10.1007/BF02289600

Lord, F. M. (1964b). *Rigorously and nonrigorously parallel test forms* (Research Bulletin No. RB-64-14). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1964.tb00323.x

Lord, F. M. (1975). *A survey of equating methods based on item characteristic curve theory* (Research Bulletin No. RB-75-13). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1975.tb01052.x

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

Lord, F. M. (1981). *Standard error of an equating by item response theory* (Research Report No. RR-81-49). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1981.tb01276.x

Lord, F. M. (1982a). Item response theory and equating: A technical summary. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 141–148). New York: Academic Press. https://doi.org/10.2307/1164642

Lord, F. M. (1982b). The standard error of equipercentile equating. *Journal of Educational Statistics, 7*, 165–174. https://doi.org/10.2307/1164642

Lord, F. M., & Wingersky, M. S. (1983). *Comparison of IRT observed-score and true-score equatings* (Research Report No. RR-83-26). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1983.tb00026.x

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed score equatings. *Applied Psychological Measurement, 8*, 453–461. https://doi.org/10.1177/014662168400800409

Mao, X., von Davier, A. A., & Rupp, S. (2006). *Comparisons of the kernel equating method with the traditional equating methods on PRAXIS data* (Research Report No. RR-06-30). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2006.tb02036.x

Marco, G. L. (1981). Equating tests in an era of test disclosure. In B. F. Green (Ed.), *New directions for testing and measurement: Issues in testing—coaching, disclosure, and ethnic bias* (pp. 105–122). San Francisco: Jossey-Bass.

Marco, G. L., Petersen, N. S., & Stewart, E. E. (1983a). *A large-scale evaluation of linear and curvilinear score equating models* (Research Memorandum No. RM-83-02). Princeton: Educational Testing Service.

Marco, G. L., Stewart, E. E., & Petersen, N. S. (1983b). A test of the adequacy of curvilinear score equating models. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 147–177). New York: Academic Press. https://doi.org/10.1016/B978-0-12-742780-5.50018-4

McHale, F. J., & Ninneman, A. M. (1994). *The stability of the score scale for the Scholastic Aptitude Test from 1973 to 1984* (Statistical Report No. SR-94-27). Princeton: Educational Testing Service.

McKinley, R. L., & Kingston, N. M. (1987). *Exploring the use of IRT equating for the GRE Subject Test in Mathematics* (Research Report No. RR-87-21). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1987.tb00225.x

McKinley, R. L., & Schaefer, G. (1989). *Reducing test form overlap of the GRE Subject Test in Mathematics using IRT triple-part equating* (Research Report No. RR-89-08). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1989.tb00334.x

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects* (Policy Information Report). Princeton: Educational Testing Service.

Modu, C. C., & Stern, J. (1975). *The stability of the SAT-Verbal score scale* (Research Bulletin No. RB-75-09). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1975.tb01048.x

Moses, T. P. (2006). *Using the kernel method of test equating for estimating the standard errors of population invariance measures* (Research Report No. RR-06-20). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2006.tb02026.x

Moses, T. P. (2008a). *An evaluation of statistical strategies for making equating function selections* (Research Report No. RR-08-60). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02146.x

Moses, T. P. (2008b). Using the kernel method of test equating for estimating the standard errors of population invariance measures. *Journal of Educational and Behavioral Statistics, 33*, 137–157. https://doi.org/10.3102/1076998607302634

Moses, T. P. (2009). A comparison of statistical significance tests for selecting equating functions. *Applied Psychological Measurement, 33*, 285–306. https://doi.org/10.1177/0146621608321757

Moses, T. P. (2011). Log-linear models as smooth operators: Holland's statistical applications and their practical uses. In N. J. Dorans & S. Sinharay (Eds.), *Looking back: Proceedings of a conference in honor of Paul W. Holland* (pp. 185–202). New York: Springer. https://doi.org/10.1007/978-1-4419-9389-2_10

Moses, T. P. (2014). Alternative smoothing and scaling strategies for weighted composite scores. *Educational and Psychological Measurement, 74*, 516–536. https://doi.org/10.1177/0013164413507725

Moses, T. P., & Holland, P. (2007). *Kernel and traditional equipercentile equating with degrees of presmoothing* (Research Report No. RR-07-15). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2007.tb02057.x

Moses, T. P., & Holland, P. W. (2008). *The influence of strategies for selecting loglinear smoothing models on equating functions* (Research Report No. RR-08-25). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02111.x

Moses, T. P., & Holland, P. W. (2009a). *Alternative loglinear smoothing models and their effect on equating function accuracy* (Research Report No. RR-09-48). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02205.x

Moses, T. P., & Holland, P. W. (2009b). *Selection strategies for bivariate loglinear smoothing models and their effects on NEAT equating functions* (Research Report No. RR-09-04). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02161.x

Moses, T. P., & Holland, P. W. (2009c). Selection strategies for univariate loglinear smoothing models and their effects on equating function accuracy. *Journal of Educational Measurement, 46*, 159–176. https://doi.org/10.1111/j.1745-3984.2009.00075.x

Moses, T. P., & Holland, P. W. (2010a). A comparison of statistical selection strategies for univariate and bivariate loglinear smoothing models. *British Journal of Mathematical and Statistical Psychology, 63*, 557–574. https://doi.org/10.1348/000711009X478580

Moses, T. P., & Holland, P. W. (2010b). The effects of selection strategies for bivariate loglinear smoothing models on NEAT equating functions. *Journal of Educational Measurement, 47*(1), 76–91. https://doi.org/10.1111/j.1745-3984.2009.00100.x

Moses, T. P., & Kim, S. (2007). *Reliability and the nonequivalent groups with anchor test design* (Research Report No. RR-07-16). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2007.tb02058.x

Moses, T. P., & Oh, H. (2009). *Pseudo Bayes estimates for test score distributions and chained equipercentile equating* (Research Report No. RR-09-47). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02204.x

Moses, T. P., & von Davier, A. A. (2006). *A SAS macro for loglinear smoothing: Applications and implications* (Research Report No. RR-06-05). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2006.tb02011.x

Moses, T. P., & von Davier, A. A. (2013). A SAS IML macro for loglinear smoothing. *Applied Psychological Measurement, 35*, 250–251. https://doi.org/10.1177/0146621610369909

Moses, T. P., & Zhang, W. (2010). *Research on standard errors of equating differences* (Research Report No. RR-10-25). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2010.tb02232.x

Moses, T. P., & Zhang, W. (2011). Standard errors of equating differences: Prior developments, extensions, and simulations. *Journal of Educational and Behavioral Statistics, 36*, 779–803. https://doi.org/10.3102/1076998610396892

Moses, T. P., von Davier, A. A., & Casabianca, J. (2004). *Loglinear smoothing: An alternative numerical approach using SAS* (Research Report No. RR-04-27). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2004.tb01954.x

Moses, T. P., Liu, J., & Dorans, N. J. (2009). *Systematized score equity assessment in SAS* (Research Memorandum No. RM-09-08). Princeton: Educational Testing Service.

Moses, T. P., Deng, W., & Zhang, Y.-L. (2010a). *The use of two anchors in nonequivalent groups with anchor test (NEAT) equating* (Research Report No. RR-10-23). Princeton*:* Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2010.tb02230.x

Moses, T. P., Liu, J., & Dorans, N. J. (2010b). Systemized SEA in SAS. *Applied Psychological Measurement, 34*, 552–553. https://doi.org/10.1177/0146621610369909

Moses, T. P., Deng, W., & Zhang, Y.-L. (2011). Two approaches for using multiple anchors in NEAT equating: A description and demonstration. *Applied Psychological Measurement, 35*, 362–379. https://doi.org/10.1177/0146621611405510

Muraki, E., Hombo, C. M., & Lee, Y.-W. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement, 24*, 325–337. https://doi.org/10.1177/01466210022031787

Myford, C., Marr, D. B., & Linacre, J. M. (1995). *Reader calibration and its potential role in equating for the Test of Written English* (Research Report No. RR-95-40). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1995.tb01674.x

Oh, H., & Moses, T. P. (2012). Comparison of the one- and bi-direction chained equipercentile equating. *Journal of Educational Measurement, 49*, 399–418. https://doi.org/10.1111/j.1745-3984.2012.00183.x

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating models. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 71–135). New York: Academic Press.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8*, 137–156. https://doi.org/10.2307/1164922

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: Macmillan.

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

Puhan, G. (2007). *Scale drift in equating on a test that employs cut scores* (Research Report No. RR-07-34). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2007.tb02076.x

Puhan, G. (2009a). Detecting and correcting scale drift in test equating: An illustration from a large scale testing program. *Applied Measurement in Education, 22*, 79–103. https://doi.org/10.1080/08957340802558391

Puhan, G. (2009b). *What effect does the inclusion or exclusion of repeaters have on test equating?* (Research Report No. RR-09-19). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02176.x

Puhan, G. (2010a). *Chained versus post stratification equating: An evaluation using empirical data* (Research Report No. RR-10-06). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2010.tb02213.x

Puhan, G. (2010b). A comparison of chained linear and poststratification linear equating under different testing conditions. *Journal of Educational Measurement, 47*, 54–75. https://doi.org/10.1111/j.1745-3984.2009.00099.x

Puhan, G. (2011a). *Can smoothing help when equating with unrepresentative small samples* (Research Report No. RR-11-09). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02245.x

Puhan, G. (2011b). Futility of log linear smoothing when equating with unrepresentative small samples. *Journal of Educational Measurement, 48*, 274–292. https://doi.org/10.1111/j.1745-3984.2011.00147.x

Puhan, G. (2011c). Impact of inclusion or exclusion of repeaters on test equating. *International Journal of Testing, 11*, 215–230. https://doi.org/10.1080/15305058.2011.555575

Puhan, G. (2012). Tucker versus chained linear equating in two equating situations—Rater comparability scoring and randomly equivalent groups with an anchor. *Journal of Educational Measurement, 49*, 313–330. https://doi.org/10.1111/j.1745-3984.2012.00177.x.

Puhan, G. (2013a). *Choice of target population weights in rater comparability scoring and equating* (Research Report No. RR-13-03). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2013.tb02310.x

Puhan, G. (2013b). Rater comparability scoring and equating: Does choice of target population weights matter in this context? *Journal of Educational Measurement, 50*, 374–380. https://doi.org/10.1111/jedm.12023

Puhan, G., & Liang, L. (2011a). Equating subscores under the non-equivalent anchor test (NEAT) design. *Educational Measurement: Issues and Practice, 30*(1), 23–35. https://doi.org/10.1111/j.1745-3992.2010.00197.x

Puhan, G., & Liang, L. (2011b). *Equating subscores using total scaled scores as the anchor test* (Research Report No. RR-11-07). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02243.x

Puhan, G., Moses, T. P., Grant, M., & McHale, F. (2008a). *An alternative data collection design for equating with very small samples* (Research Report No. RR-08-11). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02097.x

Puhan, G., von Davier, A. A., & Gupta, S. (2008b). *Impossible scores resulting in zero frequencies in the anchor test: Impact on smoothing and equating* (Research Report No. RR-08-10). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02096.x

Puhan, G., Moses, T. P., Grant, M., & McHale, F. (2009). Small sample equating using a single group nearly equivalent test (SiGNET) design. *Journal of Educational Measurement, 46*, 344–362. https://doi.org/10.1111/j.1745-3984.2009.00085.x

Puhan, G., von Davier, A. A., & Gupta, S. (2010). A brief report on how impossible scores impact smoothing and equating. *Educational and Psychological Measurement, 70*, 953–960. https://doi.org/10.1177/0013164410382731

Qian, J., von Davier, A. A., & Jiang, Y. (2013). *Weighting test samples in IRT linking and equating: Toward an improved sampling design for complex equating* (Research Report No. RR-13-39). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2013.tb02346.x

Rao, C. R., & Sinharay, S. (Eds.). (2007). *Psychometrics* (Handbook of statistics, Vol. 26). Amsterdam: Elsevier.

Ricker, K. L., & von Davier, A. A. (2007). *The impact of anchor test lengths on equating results in a nonequivalent groups design* (Research Report No. RR-07-44). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2007.tb02086.x

Rijmen, F., Manalo, J., & von Davier, A. A. (2009). Asymptotic and sampling-based standard errors for two population invariance measures in the linear equating case. *Applied Psychological Measurement, 33*, 222–237. https://doi.org/10.1177/0146621608323927

Rijmen, F., Qu, Y., & von Davier, A. A. (2011). Hypothesis testing of equating differences in the kernel equating framework. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 317–326). New York: Springer. https://doi.org/10.1007/978-0-387-98138-3_19

Rock, D. R. (1982). Equating using the confirmatory factor analysis model. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 247–257). New York: Academic Press.

Rosenbaum, P. R. (1985). *A generalization of adjustment, with an application to the scaling of essay scores* (Research Report No. RR-85-02). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1985.tb00087.x

Rosenbaum, P. R., & Thayer, D. T. (1987). Smoothing the joint and marginal distributions of scored two-way contingency tables in test equating. *British Journal of Mathematical and Statistical Psychology, 40*, 43–49. https://doi.org/10.1111/j.2044-8317.1987.tb00866.x

Rubin, D. B. (1982). Discussion of "Partial orders and partial exchangeability in test theory." In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 339–341). New York: Academic Press.

Rubin, D. B., & Szatrowski, T. (1982). Discussion of "Section pre-equating: A preliminary investigation." In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 301–306). New York: Academic Press.

Schmitt, A. P., Cook, L. L., Dorans, N. J., & Eignor, D. R. (1990). Sensitivity of equating results to different sampling strategies. *Applied Measurement in Education, 3*, 53–71. https://doi.org/10.1207/s15324818ame0301_5

Schultz, D. G., & Wilks, S. S. (1950). *A method for adjusting for lack of equivalence in groups* (Research Bulletin No. RB-50-59). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1950.tb00682.x

Sinharay, S. (2011). Chain equipercentile equating and frequency estimation equipercentile equating: Comparisons based on real and stimulated data. In N. J. Dorans & S. Sinharay (Eds.), *Looking back: Proceedings of a conference in honor of Paul W. Holland* (pp. 203–219). New York: Springer. https://doi.org/10.1007/978-1-4419-9389-2_11

Sinharay, S., & Haberman, S. J. (2011a). Equating of augmented subscores. *Journal of Educational Measurement, 48*, 122–145. https://doi.org/10.1111/j.1745-3984.2011.00137.x

Sinharay, S., & Haberman, S. J. (2011b). *Equating of subscores and weighted averages under the NEAT design* (Research Report No. RR-11-01). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02237.x

Sinharay, S., & Holland, P. W. (2006a). *Choice of anchor test in equating* (Research Report No. RR-06-35). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2006.tb02040.x

Sinharay, S., & Holland, P. W. (2006b). *The correlation between the scores of a test and an anchor test* (Research Report No. RR-06-04). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2006.tb02010.x

Sinharay, S., & Holland, P. W. (2008). *The missing data assumption of the NEAT design and their implications for test equating* (Research Report No. RR-09-16). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02173.x

Sinharay, S., & Holland, P. W. (2010a). A new approach to comparing several equating methods in the context of the NEAT design. *Journal of Educational Measurement, 47*, 261–285. https://doi.org/10.1111/j.1745-3984.2010.00113.x

Sinharay, S., & Holland, P. W. (2010b). The missing data assumption of the NEAT design and their implications for test equating. *Psychometrika, 75*, 309–327. https://doi.org/10.1007/s11336-010-9156-6

Sinharay, S., Holland, P. W., & von Davier, A. A. (2011a). Evaluating the missing data assumptions of the chain and poststratification equating methods. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 281–296). New York: Springer. https://doi.org/10.1007/978-0-387-98138-3_17

Sinharay, S., Dorans, N. J., & Liang, L. (2011b). First language of examinees and fairness assessment procedures. *Educational Measurement: Issues and Practice, 30*, 25–35. https://doi.org/10.1111/j.1745-3992.2011.00202.x

Sinharay, S., Haberman, S., Holland, P. W., & Lewis, C. (2012). *A note on the choice of an anchor test in equating* (Research Report No. RR-12-14). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2012.tb02296.x

Stocking, M. L., & Eignor, D. R. (1986). *The impact of different ability distributions on IRT preequating* (Research Report No. RR-86-49). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1986.tb00204.x

Stocking, M. L., Eignor, D. R., & Cook, L. L. (1988). *Factors affecting the sample invariant properties of linear and curvilinear observed- and true-score equating procedures* (Research Report No. RR-88-41). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1988.tb00297.x

Swineford, F., & Fan, C. T. (1957). A method of score conversion through item statistics. *Psychometrika, 22*, 185–188. https://doi.org/10.1007/BF02289053

Tan, X., Ricker, K., & Puhan, G. (2010). *Single versus double scoring of trend responses in trend score equating with constructed response tests* (Research Report No. RR-10-12). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2010.tb02219.x

Tang, L. K., Way, W. D., & Carey, P. A. (1993). *The effect of small calibration sample sizes on TOEFL IRT-based equating* (Research Report No. RR-93-59). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1993.tb01570.x

Thayer, D. T. (1983). Maximum likelihood estimation of the joint covariance matrix for sections of tests given to distinct samples with application to test equating. *Psychometrika, 48*, 293–297. https://doi.org/10.1007/BF02294023

Tucker, L. (1951). *Notes on the nature of gamble in test score scaling* (Research Bulletin No. RB-51-27). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1951.tb00226.x

von Davier, A. A. (2003). *Notes on linear equating methods for the non-equivalent-groups design* (Research Report No. RR-03-24). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2003.tb01916.x

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: Springer. https://doi.org/10.1007/978-0-387-49771-6_6

von Davier, A. A. (2008). New results on the linear equating methods for the non-equivalent-groups design. *Journal of Educational and Behavioral Statistics, 33*, 186–203. https://doi.org/10.3102/1076998607302633

von Davier, A. A. (2011a). A statistical perspective on equating test scores. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling and linking* (pp. 1–17). New York: Springer. https://doi.org/10.1007/978-0-387-98138-3

von Davier, A. A. (2011b). An observed-score equating framework. In N. J. Dorans & S. Sinharay (Eds.), *A festschrift for Paul W. Holland* (pp. 221–237). New York: Springer. https://doi.org/10.1007/978-1-4419-9389-2_12

von Davier, A. A. (Ed.). (2011c). *Statistical models for test equating, scaling and linking*. New York: Springer. https://doi.org/10.1007/978-0-387-98138-3

von Davier, A. A. (2013). Observed score equating: An overview. *Psychometrika, 78*, 605–623. https://doi.org/10.1007/s11336-013-9319-3

von Davier, A. A., & Kong, N. (2005). A unified approach to linear equating for the non-equivalent groups design. *Journal of Educational and Behavioral Statistics, 30*, 313–334. https://doi.org/10.3102/10769986030003313

von Davier, A. A., & Liu, M. (Eds.). (2007). Population invariance. [Special issue]. *Applied Psychological Measurement, 32*(1).

von Davier, A. A., & Wilson, C. (2007). IRT true-score test equating: A guide through assumptions and applications. *Educational and Psychological Measurement, 67*, 940–957. https://doi.org/10.1177/0013164407301543

von Davier, A. A., & Wilson, C. (2008). Investigation the population sensitivity assumption of item response theory true-score equating across two subgroups of examinees and two test formats. *Applied Psychological Measurement, 32*, 11–26. https://doi.org/10.1177/0146621607311560

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004a). The chain and post-stratification methods of observed-score equating: Their relationship to population invariance. *Journal of Educational Measurement, 41*, 15–32. https://doi.org/10.1111/j.1745-3984.2004.tb01156.x von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004b). *The kernel method of test equating*. New York: Springer. https://doi.org/10.1007/b97446

von Davier, A. A., Holland, P. W., Livingston, S. A., Casabianca, J., Grant, M. C., & Martin, K. (2006). *An evaluation of the kernel equating method: A special study with pseudo-tests constructed from real test data* (Research Report No. RR-06-02). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2006.tb02008.x

von Davier, A.A., Fournier-Zajac, S., & Holland, P. W. (2007). *An equipercentile version of the Levine linear observed score equating function using the methods of kernel equating* (Research Report No. RR-07-14). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2007.tb02056.x

von Davier, M., & von Davier, A. A. (2007). A unified approach to IRT scale linkage and scale transformations. *Methodology, 3*, 115–124. https://doi.org/10.1027/1614-2241.3.3.115.

von Davier, M., Gonzalez, J., & von Davier, A. A. (2013). Local equating using the Rasch model, the OPLM, and the 2PL IRT model—or—What is it anyway if the model captures everything there is to know about the test takers? *Journal of Educational Measurement, 50*, 295–303. https://doi.org/10.1111/jedm.12016

Wainer, H., & Thissen, D. (1994). On examinee choice in educational testing. *Review of Educational Research, 64*, 159–195. https://doi.org/10.3102/00346543064001159

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R., Sternberg, L., & Thissen, D. (1990). *Computerized adaptive testing: A primer*. Hillsdale: Lawrence Erlbaum Associates.

Wainer, H., Wang, X.-B., & Thissen, D. (1991). *How well can we equate test forms that are constructed by the examinees* (Research Report No. RR-91-57). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1991.tb01424.x

Wainer, H., Wang, X.-B., & Thissen, D. (1994). How well can we compare scores on test forms that are constructed by examinee choice? *Journal of Educational Measurement, 31*, 183–199. https://doi.org/10.1111/j.1745-3984.1994.tb00442.x

Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R., Sternberg, L., & Thissen, D. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Hillsdale: Erlbaum.

Walker, M. E., & Kim, S. (2010). *Examining two strategies to link mixed-format tests using multiple-choice anchors* (Research Report No. RR-10-18). Princeton: Educational Testing Service.. https://doi.org/10.1002/j.2333-8504.2010.tb02225.x

Wang, X.-B., Wainer, H., & Thissen, D. (1995). On the viability of some untestable assumptions in equating exams that allow examinee choice. *Applied Measurement in Education, 8*, 211–225. https://doi.org/10.1207/s15324818ame0803_2

Wiberg, M., van der Linden, W., & von Davier, A. A. (2014). Local observed-score kernel equating. *Journal of Educational Measurement, 51*, 57–74. https://doi.org/10.1111/jedm.12034

Wightman, L. E., & Wightman, L. F. (1988). *An empirical investigation of one variable section pre-equating* (Research Report No. RR-88-37). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1988.tb00293.x

Yang, W.-L. (2004). Sensitivity of linkings between *AP*® multiple-choice scores and composite scores to geographical region: An illustration of checking for population invariance. *Journal of Educational Measurement, 41*, 33–41. https://doi.org/10.1111/j.1745-3984.2004.tb01157.x

Yang, W.-L., & Gao, R. (2008). Invariance of score linkings across gender groups for forms of a testlet-based college-level examination program examination. *Applied Psychological Measurement, 32*, 45–61. https://doi.org/10.1177/0146621607311577

Yang, W.-L., Bontya, A. M., & Moses, T. P. (2011). *Repeater effects on score equating for a graduate admissions exam* (Research Report No. RR-11-17). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02253.x

Yen, W. (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: Springer. https://doi.org/10.1007/978-0-387-49771-6_15

Zu, J., & Liu, J. (2009). *Comparison of the effects of discrete anchor items and passage-based anchor items on observed-score equating results* (Research Report No. RR-09-44). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02201.x

Zu, J., & Liu, J. (2010). Observed score equating using discrete and passage-based anchor items. *Journal of Educational Measurement, 47*, 395–412. https://doi.org/10.1111/j.1745-3984.2010.00120.x

Zu, J., & Puhan, G. (2014). Pre-equating with empirical item characteristic curves: An observed-score pre-equating method. *Journal of Educational Measurement, 51*, 281–300. https://doi.org/10.1111/jedm.12047

Zu, J., & Yuan, K. (2012). Standard error of linear observed-score equating for the neat design with nonnormally distributed data. *Journal of Educational Measurement, 49*, 190–213. https://doi.org/10.1111/j.1745-3984.2012.00168.x