# Chapter 1
# A Statistical Perspective on Equating Test Scores

**Alina A. von Davier**

> "The fact that statistical methods of inference play so slight a role... reflect[s] the lack of influence modern statistical methods has so far had on current methods for test equating."
> Rubin (1982, p. 53)

> "The equating problem reduces to one of modeling and statistical theory."
> Morris (1982, p. 170)

## 1.1 Introduction

The comparability of scores across different test forms of a standardized assessment has been a major focus of educational measurement and the testing industry for the past 90 years (see Holland, 2007, for a history of linking). This chapter focuses on the statistical methods available for equating test forms from standardized educational assessments that report scores at the individual level (see also Dorans, Moses, & Eignor, Chapter 2 of this volume). The overview here is given in terms of frameworks[1] that emphasize the statistical perspective with respect to the equating methodologies that have been developed by testing practitioners since the 1920s. The position taken in this paper is that the purpose of the psychometricians' work is to accurately and fairly measure and compare educational skills using multiple test forms from an educational assessment. Therefore, from this measurement perspective, equating of test forms is only one necessary step in the measurement process. Equating is only necessary because a standardized educational assessment uses

[1] *Conceptual frameworks* (theoretical frameworks) are a type of intermediate theory that have the potential to connect to all aspects of inquiry (e.g., problem definition, purpose, literature review, methodology, data collection and analysis). Conceptual frameworks act like maps that give coherence to empirical inquiry ("Conceptual Framework," n.d.).

A.A. von Davier
Educational Testing Service, Rosedale Rd, Princeton, NJ 08541, USA
e-mail: avondavier@ets.org

A.A. von Davier (ed.), *Statistical Models for Test Equating, Scaling, and Linking*,
Statistics for Social and Behavioral Sciences,
DOI 10.1007/978-0-387-98138-3_1, 

numerous test forms that tend to differ in difficulty although they are built to the same specifications ("nominally parallel test forms," Lord & Novick, 1968, p. 180). Hence, equating can be viewed as the process of controlling statistically for the confounding variable "test form" in the measurement process. If the test development process were perfect, then equating would not be necessary. See also Lord's (1980) theorem 13.3.1 in Chapter 13. The term *linking* has slightly different meanings in the field of educational measurement, and it is used here as (a) a general term for denoting a relationship between test forms (at the total score level, at the item parameter level, etc.); (b) as a weaker form of equating; and (c) as a synonym to the process of placing item response theory (IRT) item parameter estimates on the same scale, which sometimes is also called IRT *calibration*. In this chapter I refer to *equating* as a strong form of linking and as a subclass of linking methods (Holland & Dorans, 2006). Test equating can be carried out both using observed-score equating (OSE) and IRT methods, but the word *equating* is most often associated with the raw scores of a test. See Holland and Dorans, Kolen and Brennan (2004), Dorans et al. (Chapter 2 of this volume), and Yen and Fitzpatrick (2006) for an extensive view of categories of linking methods.

The process of measuring and comparing competencies in an educational assessment is described here in ways that integrate various existing approaches. A discussion of equating as a part of the measurement process is given first. Then I introduce the idea of applying a testlet or a bifactor model to measure skills and equate scores. This type of model would capture the test-form effect as a latent variable with a distribution. This variable, the test-form effect, can be (a) monitored over time to inform on the stability of equating, (b) used as feedback for the test developers to improve upon the degree of parallelism of test forms, and (c) used for monitoring the form effect on subgroups. Next, an equating framework for the OSE methods is introduced. I discuss how the search for a theory of OSE led to the development of a framework that provides a map that gives coherence to empirical inquiry. A framework for IRT parameter linking is given by M. von Davier and von Davier (Chapter 14 of this volume), and a practical perspective on equating methods is given by Dorans et al. in Chapter 2. The last section of this chapter and the *Overview* outline the rest of the volume. The chapters are grouped according to the steps of a measurement process that are described in the next section.

## 1.2 The Measurement Model, the Unit of Measurement, and Equating

Parallels between a generic statistical modeling process and an educational measurement process that includes the equating of test forms are presented in this section. Subsequently, a link between the equating methodologies and the unit of measurement is discussed.

### 1.2.1 Statistical Modeling and Assumptions

The measurement process in standardized testing, which includes test form equating, follows the same steps as a typical statistical modeling process. Statistical models are ideal and simplistic representations of a (complex) reality that aid in the description and understanding of a specific process or that explain or predict future outcomes. Statistical modeling is accomplished by first identifying the main variables and their interactions that explain the particular process. Using a simple model to describe a complex reality requires making many assumptions that allow the reality to be simplified. The usual steps in any statistical modeling process are as follows:

1. Statistical modeling starts with a research question and with a set of data.
2. One of the challenges of statistical modeling is the danger of confounding: The inferences one makes about one variable based on a model might be confounded by interactions with other variables that exist in the data and that have not been explicitly modeled. The confounding trap can be addressed by elegant and elaborate sampling procedures, data collection designs, and explicit modeling of the variables.
3. A statistical model is proposed and fitted to the data, and the model parameters are estimated.
4. Assumptions are made about the data generating process. If the model fits the data to an acceptable degree,[2] then inferences are made based on the model.
5. The results are evaluated with respect to (sampling) error and bias. Given that all statistical models are approximations of reality and that they almost never fit the data, statisticians have developed indices that attempt to quantify the degree to which the results are accurate. The bias introduced by the modeling approach is investigated.

The same sequence of events describes the process of measurement in standardized testing (see also Braun & Holland, 1982). The steps in the measurement process are as follows:

1. The measurement process starts with two or more test forms built to the same specifications (nominally parallel test forms), with the research question being how to measure and compare the skills of the test takers regardless of which form they took.
2. The challenge in measuring the skills of test takers, who take different forms of a test, is how to avoid the confounding of differences in form difficulty with the differences in the ability of the test takers. In order to disentangle the test forms differences and ability differences, data are collected in specific ways and assumptions about the data generating process are explicitly incorporated.

[2] "All models are wrong but some are useful" (Box & Draper, 1987, p. 74).

See von Davier, Holland, and Thayer (2004b, Chapter 2) and Dorans et al. (Chapter 2 of this volume) for details on data collection designs.

3. The next step is modeling the data generating process. Data from educational tests are in most cases noisy and models have been proposed to fit them (log-linear models, spline functions, IRT models). These models rely on assumptions. The measurement models that include equating also have underlying assumptions. For example, in OSE, the model-estimated test-score distributions are linked using an equipercentile function. The equipercentile function is a mathematical function composition that requires that the data be continuous, and the test scores usually are not. Hence, the data need to be continuized. Continuization involves an approximation approach commonly employed in probability theory and statistical theory. IRT models make different assumptions from OSE. For example, the estimated item or ability parameters are linked using a linear transformation assuming the IRT model fits the data well for each test form. Or, the method called *IRT true-score equating* assumes that the relationship between the true-scores holds also for the observed-scores.
4. Hence, assumptions are made about the data generating process. If the model fits the data to an acceptable degree, then inferences are made based on the model.
5. Since the parameters of the equating models are sample estimates, the equating results are subject to sample variability. At the end of the equating procedure (after several steps of making assumptions), one will quantify the degree of error cumulated in the process. This is obtained through the use of statistical indices: standard errors of parameters, standard errors of equating (SEE), standard errors of equating differences (SEED), as well as other statistical indices such as the likelihood ratio statistics, Freeman-Tukey residuals, and the Akaike criterion (Bishop, Fienberg, & Holland, 1975; Bozdogan, 1987). In addition, the potential bias in the equating results should be evaluated according to different criteria, such as the historical information available, stability of results over time, consistency checks when multiple equating methods are available, changes in demographics, population invariance, and scale drift. One might employ quality assurance methods or statistical process control methods to monitor the stability of the reported scores over time—such as cumulative sum charts and time series analyses (See Li, Li, & von Davier, Chapter 20 of this volume).

The parallel between a generic statistical process and the educational measurement process is illustrated in Figure 1.1. As already mentioned, no model fits the data perfectly; moreover, many models are very complex and rely on assumptions that are not easily tested. Therefore, a discussion of the merits of different models requires investigation of the assumptions that underlie the models and, more importantly, analysis of the consequences of failure to meet these assumptions.

In very simple data collection equating designs, such as the equivalent-groups design and the single-group design, the OSE methods assume very little. As Braun and Holland (1982) noted, the OSE methods are

> ... completely atheoretical in the sense that they are totally free of any conception (or misconception) of the subject matter of the two tests $X$ and $Y$ ... . We are only

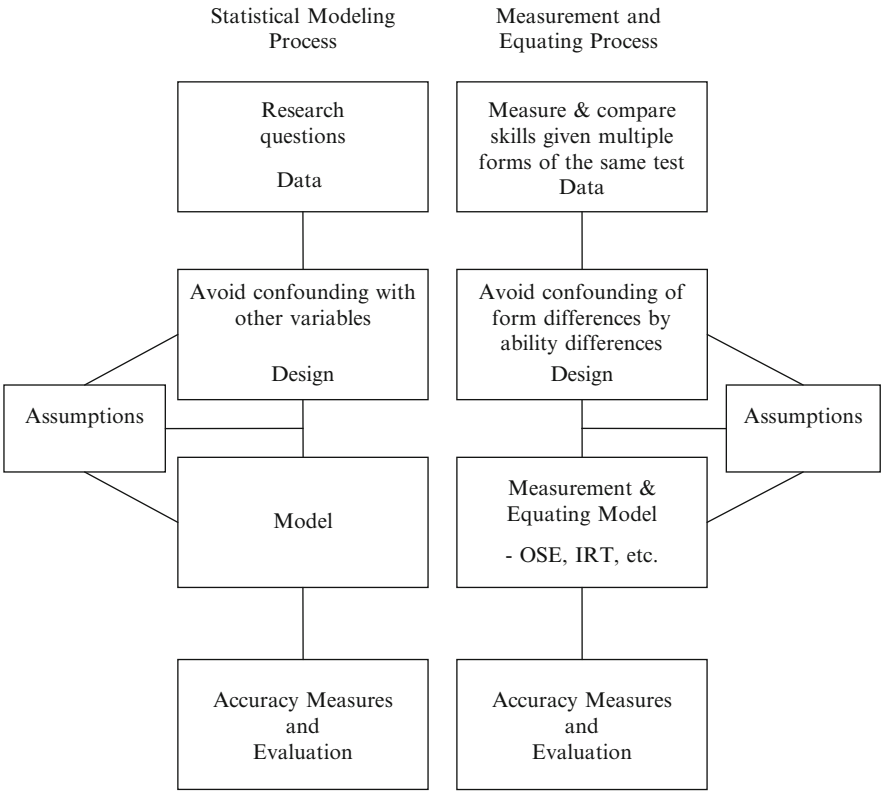


**Fig. 1.1** The parallel between a generic statistical process and the educational measurement process. IRT = item response theory; OSE = observed-score equating

> preventing from equating a verbal test to a mathematical test by common sense. This is an inherent problem with observed-score equating. (p. 16)

On the other hand, with the more complex nonequivalent groups with an anchor test (NEAT) design all OSE methods make more assumptions, some of them untestable (see Sinharay, Holland, & von Davier, Chapter 17 of this volume; also, Braun & Holland, 1982). Due to these untestable assumptions, some OSE models are difficult to evaluate with the data at hand. The IRT model assumptions are equally demanding and difficult to evaluate.

The question with both sets of equating models (OSE and IRT) is whether the model errors necessarily invalidate the procedures or whether the errors are sufficiently limited in their consequences so that the equating approaches are acceptable. This analysis can be difficult to carry out both with IRT and OSE methods when employing complex designs. IRT does provide more possibilities in complex linking situations that are sometimes not feasible with OSE (such as in survey assessments, where the data are collected following a matrix design—where not all test takers take all items). However, a matrix design and a complex IRT

model also involve an increased level of difficulty with respect to verification of assumptions and an increased reliance on strong assumptions that are needed to compensate for missing data. The selection of an equating method mainly matters when the need for equating is strongest (that is, when the forms differ in difficulty) and all methods produce similar results when the forms and populations are identical.

### *1.2.2 Equating and Measurement*

The purpose of this section is to identify the unit of measurement in a measurement process that includes equating of test forms. Then I identify what is to be linked when the equating of scores is desired, when OSE and IRT methods are employed. It is assumed that an appropriate data collection design is available for equating (see Holland & Dorans, 2006). An interesting discussion of similar questions has been given in Braun and Holland (1982), and Morris (1982).

As Lord and Novick (1968) pointed out, any measurement "begins with a procedure for identifying elements of the real world with the elements or constructs of an abstract logical system (a model)" (p. 16). Lord and Novick continued,

> To specify this measurement we must do three things: First we must identify the object being measured, the person, or the experimental unit. Then we must identify the property or behavior being directly measured.... Finally, we must identify the numerical assignment rule by which we assign a number to this property of the unit being measured. (p. 16)

Educational testing programs apply a measurement tool (the test form) to test takers assumed to be randomly sampled from a population. The assessments measure a specific skill that can be "the examinee's responses to the items" (Lord & Novick, 1968, p. 16), a latent skill, or a merely unobserved skill. "Theoretical constructs are often related to the behavioral domain through observable variables by considering the latter as measures or indicants of the former" (Lord & Novick, 1968, p. 19). The idea that a measurement is something true ("the property or behavior" that the instrument is supposed to measure) plus an error of measurement is an old concept developed initially in astronomy and other physical sciences (see Lord & Novick, 1968, p. 31; see Holland, 2007, for a history of testing and psychometrics). The measurement takes place indirectly through a number of carefully developed items that comprise the test form given to a sample of test takers (the random variable with a distribution). The measurement data can be in the form of arrays of direct responses, such as arrays of 0s and 1s representing correct or incorrect responses to multiple-choice items, or in some cases, further aggregated (through adding the number of correct responses) to total scores and distributions. Kolen, Tong, and Brennan (Chapter 3 of this volume) called the unit of measurement "raw score:" "Raw scores can be as simple as a sum of the item scores or be so complicated that they depend on the entire pattern of item responses." Regardless of how the scores are obtained, they are the realizations of the random variable—the testing instrument and form.

In a standardized educational assessment many test forms are built to the same specifications, and each of these test forms is a testing instrument. These nominally parallel test forms (Lord & Novick, 1968, p. 180) usually differ in difficulty, and therefore, the measurement challenge is how to disentangle the unintended differences in difficulty among the test forms from the ability of the test takers. In other words, the role of equating is to insure an accurate measurement of an underlying skill for a test taker, regardless of what test form has been taken by this test taker (see Figure 1.1). The method chosen to equate test forms depends on the model used for measurement.

In assessments where the OSE methods are employed, the item information is aggregated across the test takers, and the test-score distribution is used as the basis for equating the test forms. Test forms are random variables with distributions, and the scores are realizations of these random variables. In (equipercentile) OSE, the cumulative distributions of the random variables test forms are mapped onto each other such that the percentiles on one will match the percentiles on the other. As indicated earlier by quoting Braun and Holland (1982), OSE does not explicitly require a meaning of the score used (i.e., total observed score, number-correct score, weighted number-correct score, formula-score). In conclusion, for the OSE methods, the unit of measurement is the total test score (regardless of how it was obtained), and the equating is accomplished through matching the two test-score distributions (either in terms of percentiles or in terms of their means and standard deviations).

In assessments where IRT-based methods are used for equating, the analysis starts with data as arrays of 0s and 1s representing correct or incorrect responses to multiple-choice items for each person.[3] Then the measurement of the underlying skill is obtained through modeling the interaction between the features of the items and of the persons who take those items. The IRT-based methods rely on a model for the probability of a correct response to a particular item by a particular person. Assuming the model fits the data, the adjustment for differences between the two test forms is accomplished through linking the item (or ability) parameters. In a subsequent step, this linking might be applied to raw test scores, and therefore, achieve equating of scores, or it might be directly applied to scale scores (Yen, 1986). Hence, for IRT-based methods, the unit of measurement is the probability that a person answers an item correctly (item by person's skill) and the adjustment for form differences is done through a linear transformation of the item parameters or of the parameters of the distribution of the underlying skill.

The appeal of the IRT models lies within the psychometric theory: IRT models are mathematical models of a test to infer the ability of a test taker and to classify the test takers according to their ability. Linking the item parameters to adjust for form differences is inherent to the IRT model. In contrast, as Braun and Holland (1982) pointed out, the OSE methods are atheoretical.

[3]There are models for accomplishing the same things with tests using polytomously scored items.

The measurement and equating models that use a total test score as a unit of measurement and match the percentiles of test-score distributions, and models that use the item–person interaction as a unit of measurement and link item or person parameters, do have similarities; sometimes they overlap or build on each other. This volume offers an account of several methods of this sort (see the following chapters: Karabatsos & Walker, Chapter 11; Chen, Livingston, & Holland, Chapter 12; van der Linden, Chapter 13; Glas & Béguin, Chapter 18).

In my opinion, the value of thinking of equating as a part of a complex measurement process lies in the multitude of possibilities that become available to the researcher. These possibilities may include applying existing models from (or developing new models in) other areas of psychology, econometrics, statistics, or even from other parts of psychometrics. That is, borrowing or developing new measurement models in a very different framework than educational measurement that could also achieve equating becomes easier to conceptualize in a broader framework. In the next section I give an example of such a cross-contamination of ideas.

## 1.3 Measurement of Skills and Equating of Test Scores Using a Testlet Model

At least three models have been developed to account for the effects of specific groups of items or testlets that might be included in an assessment. These item bundles may refer to the same passage, or the same test material, and the responses to the items from the testlets might not be independent given the ability, and therefore, the assumption of unidimensionality of the IRT model might be violated. One way to account for the testlet effect is to incorporate specific dimensions in addition to the general underlying dimension of the IRT model. Three such models are the bifactor model (Gibbons & Hedeker, 1992), the second-order factor model (Rijmen, 2009b), and the testlet model (Bradlow, Wainer, & Wang, 1999). The last two models were shown to be formally equivalent in Rijmen, and therefore, I will briefly discuss only the bifactor and second-order model here.

In the bifactor model (Gibbons & Hedeker, 1992), each item measures a general dimension and one of $K$ specific dimensions. Typically, all dimensions are assumed to be independent. Here I will use a less general restriction: These dimensions are assumed to be independent given the general dimension. Figure 1.2 shows a bifactor model with the conditional independence restriction using a directed acyclic graph for four sets of items $y_1$ to $y_4$, the general ability $\theta_g$, and the specific testlets' effects, $\theta_1$ to $\theta_4$.

A second-order model also includes separate testlet effects. Figure 1.3 illustrates a second-order model with the same conditional independence restriction. In a second-order model, each testlet has a separate dimension. As in the bifacor model, the specific testlet effects are assumed to be conditionally independent,

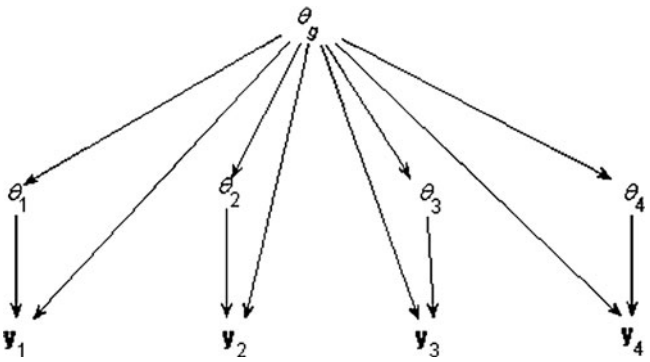


**Fig. 1.2** Directed acyclic graph of the bifactor model. From *Three Multidimensional Models for Testlet Based Tests*, by F. Rijmen, 2009b, Princeton, NJ: ETS, p. 2. Copyright 2009 ETS. Reprinted with permission

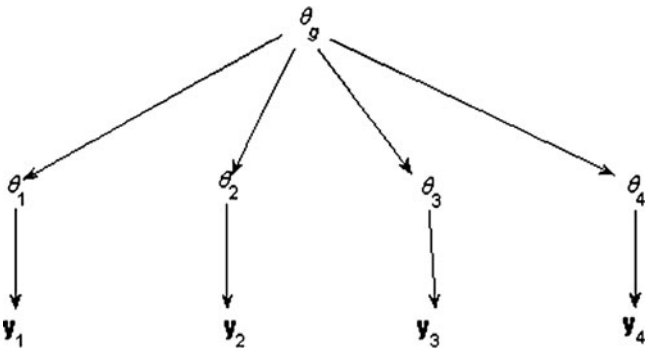


**Fig. 1.3** Directed acyclic graph of the second-order model. From *Three Multidimensional Models for Testlet Based Tests*, by F. Rijmen, 2009b, Princeton, NJ: ETS, p. 5. Copyright 2009 ETS. Reprinted with permission

given the general ability. In this model the general ability is indirectly measured by the items, through the specific testlet factors. In Figure 1.3 this is represented through the absence of directed edges between the general ability $\theta_g$ and the specific testlets' effects, $\theta_1$ to $\theta_4$.

Now assume that each of the $y_1$ to $y_4$ actually denote a test form in Figures 1.2 and 1.3. Assume that these test forms are nominally equivalent forms that need to be equated. Assume for simplicity reasons that each of the four forms represented in Figures 1.2 and 1.3 does not include any testlets. Under these assumptions, each of the three models, the bifactor, the second-order, or the testlet model, can be applied as the measurement model for a single-group data collection design, where the same test takers took all four test forms. Other data collection designs can be eventually considered (see Rijmen, 2009a, where the model was applied to a matrix design from the Progress in International Reading Literacy Study). This assumption is made here only to simplify the parallels between the concurrent unidimensional IRT

calibration and linking, and equating of scores on one side, and the concurrent calibration with a testlet model, and equating of scores on the other side. Once the concurrent calibration of items has been achieved, and the items, test forms, and ability parameters have been estimated using one of the testlet models mentioned here, then equating of scores can be achieved through the general ability $\theta g$ using the method called IRT true-score equating, or using the method called IRT OSE, or using the local equating method (van der Linden, Chapter 13 of this volume).

Obviously, the length of the test forms, the sample size, the specifics of the data collection design, the degree of correlation between the various dimensions, each can be a challenge for fitting successfully a complex model such as any of the testlet models mentioned above. This will be the topic of future research.

In my opinion, the advantage of using a testlet or test-form model for linking and equating lies in the estimate of the test-form effect. As a practitioner, I can see the advantages of monitoring the distribution of the test-form effect over time to support reporting stable equating results and of providing meaningful feedback to the test developers. This feature might be of particular interest for assessments with an almost continuous administration mode. For assessments with numerous administrations one could apply the statistical process control charts to several variables (the means of the general-ability and the form-effect dimensions estimates over time together with a standard deviation band). If differential test-form functioning is of concern, then these specific test-form variables can be monitored for the subgroups of interest. The testlet model applied to test forms also can be extended to incorporate testlets inside each form, as in a hierarchical model.

Another example of a cross-contamination of ideas is presented in Rock (1982). In his paper, "Equating Using Confirmatory Factor Analysis," Rock showed how to use maximum-likelihood factor analysis procedures to estimate the equating parameters, under the assumption that the components of the vector of the test scores have a multivariate normal distribution.

Next, a mathematical framework that includes all OSE methods is described. The OSE framework follows the measurement model described in Figure 1.1 and follows the description of the OSE methods as equating approaches that match the test score distributions.

## 1.4 An OSE Framework

In this section, a framework for the OSE methods is introduced. The advantages of a single framework that includes all OSE methods are (a) a formal level of cohesiveness, (b) a modular structure that leads to one software package for all methods, and (c) the facilitation of development and comparison of new equating models. This framework is referred as the *OSE framework*. This framework follows the line of argument from the previous two sections.

Identifying a framework that connects the methods used in observed-score equating practice is part of the continuous search for a theory of equating (see also Holland & Hoskens, 2003; von Davier, in press). This equating framework together with Dorans and Holland's five requirements of an equating procedure (Dorans & Holland, 2000), is the closest to a theory that is available for observed-score equating.

The OSE framework outlined here consists of the five steps in the OSE process as described in von Davier et al. (2004a) for the kernel equating and includes an explicit description of the relationship between the observed-score equipercentile and linear equating functions. Moreover, the framework described here shows conceptual similarities with the mathematical framework introduced in Braun and Holland (1982). Next, the notation and the OSE framework are presented.

In the following exposition, it is assumed that an appropriate data collection design is available for measuring skills on a standardized educational assessment, where equating of test scores is needed. The two nominally parallel test forms to be equated are assumed to be well constructed and equally reliable. As in Figure 1.1, the research question is how to measure accurately and fairly the educational skills of the test takers who took these two nominally parallel test forms. The two test forms to be equated are denoted here by $X$ and $Y$; the same notation is also used for the test scores as random variables with distributions. Score distributions are usually discrete, so to describe them, both their possible values and the associated probabilities of these possible values are given. The possible values for the random variables $X$ and $Y$ are denoted by $x_j$ (with $j = 1, \ldots, J$) and $y_k$ (with $k = 1, \ldots, K$), respectively. As mentioned earlier, for the OSE methods, the unit of measurement is the test score, and the equating is accomplished by matching the two test score distributions (either in terms of percentiles or in terms of their means and standard deviations). In the simple case of total-number-correct scoring, the possible values for $X$ are consecutive integers, such as $x_1 = 0$, $x_2 = 1$, etc. In other cases, the possible values can be negative or have fractional parts—as it is the case of unrounded formula scores or ability estimates from models that use IRT. We assume in the following that the unit of measurement is the total number correct score.

Most OSE functions (in particular the nonlinear ones) depend on the score probability distributions on a target population, called $T$ here. The vectors of the score probabilities are denoted by $\mathbf{r}$ and $\mathbf{s}$ on $T$:

$$\mathbf{r} = (r_1, \ldots, r_J), \text{ and } \mathbf{s} = (s_1, \ldots, s_K). \tag{1.1}$$

and each $r_j$ and $s_k$ are defined by

$$r_j = P\{X = x_j | T\} \text{ and } s_k = P\{Y = y_k | T\}. \tag{1.2}$$

The score probabilities for $X$ are associated with the $X$ raw scores, $\{x_j\}$, and those for $Y$ are associated with the $Y$ raw scores, $\{y_k\}$. The steps of the OSE

framework describe the equating process and are covered in detail in the following subsections.

### 1.4.1 Step 1: Presmoothing

It is customary to presmooth the data to remove some of the sampling noise if the samples are below 20,000. The score probabilities are either estimated through various procedures such as fitting log-linear models to the observed-score test probabilities or by estimating them using the sample frequencies if the samples are large; either way, they are subsequently collected as part of a row vector, $\hat{\mathbf{u}}$. A description of log-linear model presmoothing is not given here because (a) it is richly documented in the literature (Holland & Thayer, 1987, 1989, 2000; Moses & Holland, 2008); (b) it is an equating step that is already widely followed and understood by practitioners of equating; and (c) in theory (and consistent with the goals of this paper), it can be achieved using other methods and models that easily can be made to match the OSE framework.

### 1.4.2 Step 2: Estimating the Score Probabilities

The estimated marginal score probabilities $\hat{\mathbf{r}}$ and $\hat{\mathbf{s}}$ are actually computed (explicitly or not) using the design function (DF) described below. The estimated equating function can be written to express the influence of the data collection design as

$$\hat{e}_y(x) = e_y[x; \mathrm{DF}(\hat{\mathbf{u}})]. \tag{1.3}$$

Equivalently, it can be written as

$$\hat{e}_y(x) = e_y(x; \hat{\mathbf{r}}, \hat{\mathbf{s}}), \tag{1.4}$$

where $\mathbf{u}$ is generic notation for the data vector that reflects the way the data are collected and $\hat{\mathbf{u}}$ denotes its estimate.

For example, if the data are collected from an equivalent-groups design, then the data are in the form of two univariate distributions; in this case the design function is the identity function and $\mathbf{u} = (\mathbf{r}, \mathbf{s})$. If the data are collected following a single-group design, where the same group of test takers takes both test forms $X$ and $Y$, then $\mathbf{u}$ is the vector whose components are the joint probabilities from the bivariate distribution. In this case, the design function is a linear function that computes the marginal probabilities $\mathbf{r}$ and $\mathbf{s}$ from this bivariate distribution. The design function becomes more complex as the various equating methods for the NEAT design become more complex, but the results of its application to vector $\mathbf{u}$ are always the score probability vectors, $\mathbf{r}$ and $\mathbf{s}$ on $T$.

### 1.4.3 Step 3: Continuization

There are different ways to continuize the discrete score distributions. In the case of kernel equating (Gaussian, uniform, logistic), the kernel functions are the added continuous random variables to the original discrete variable. I am describing the kernel method of continuization because it also includes the linear interpolation. The traditional equipercentile equating function uses a piecewise linear function as the new continuous distribution. This also can be expressed as in Equations 1.5 and 1.6, with $V$ being a uniform kernel (see Holland & Thayer, 1989, and Lee & von Davier, Chapter 10 of this volume).

Consider $X(h_X)$ as a continuous transformation of $X$ such that

$$X(h_X) = a_X(X + h_X V) + (1 - a_X)\mu_{XT}, \tag{1.5}$$

where

$$a_X^2 = \frac{\sigma_{XT}^2}{\sigma_{XT}^2 + \sigma_V^2 h_X^2} \tag{1.6}$$

and $h_X$ is the bandwidth controlling the degree of smoothness. In Equation 1.5, $V$ is a continuous (kernel) distribution with variance $\sigma_V^2$ and mean 0. The mean and the variance of $X$ on $T$ are denoted by $\mu_{XT}$ and $\sigma_{XT}^2$, respectively. The role of $a_X$ in Equation 1.5 is to insure that the first two moments of the transformed random variable $X(h_X)$ are the same as the first two moments of the original discrete variable $X$. When $h_X$ is large, the distribution of $X(h_X)$ approximates the distribution of $V$; when $h_X$ is small, $X(h_X)$ approximates $X$, but as a continuous function. In von Davier et al. (2004a), $V$ follows a standard normal distribution (that is, a Gaussian kernel, with mean 0 and variance 1), which is why the terms *Gaussian kernel equating* and *kernel equating* are sometime used interchangeably. However, Lee and von Davier (2008; also see Chapter 10 of this volume) discussed the use of alternative kernels for equating, and in their approach $V$ is a generic continuous distribution. The $Y$ distribution is continuized in a similar way.

One important property of the OSE framework that was developed for kernel equating functions (Gaussian or other kernels) is that by manipulating the bandwidths for the new distributions one can obtain a family of equating functions that includes linear equating (when the bandwidths are large) and equipercentile equating (when the bandwidths are small) as special cases. The choice of bandwidth balances the closeness of the continuous distribution to the data and the smoothness of the new continuous function. The continuized function $X(h_X)$ can be evaluated or diagnosed by comparing its moments to the moments of the discrete score distribution, in this case, of $X$. Other OSE methods employ different strategies to continuize the distributions (see Haberman, Chapter 8 of this volume; Wang, Chapter 9 of this volume).

### 1.4.4 Step 4: Computing the Equating Function

Once the discrete distribution functions have been transformed into continuous cumulative distribution functions (CDFs), the observed-score equipercentile equating function that equates $X$ to $Y$ is computed as

$$\hat{e}_y(x) = e_y[x; \mathrm{DF}(\hat{\mathbf{u}})] = {G_{Tc}}^{-1}[F_{Tc}(x; \hat{\mathbf{r}}); \hat{\mathbf{s}}], \tag{1.7}$$

where $G_{Tc}$ is the continuized cumulative distribution function of $Y$ on the target population $T$ and $F_{Tc}$ is the continuized cumulative distribution function of $X$ on $T$. The equating function $e_Y$ in Equation 1.7 can have different formulas (linear or nonlinear, for example). In a NEAT design, it can take the form of chained equating, poststratification equating, Levine equating, and so on.

### 1.4.5 Step 5: Evaluating the Equating Results and Computing Accuracy Measures

The equating function can be evaluated by comparing the moments of the equated scores distribution $\hat{e}_y(x)$ to the moments of the targeted discrete-score distribution, in this case, of $Y$. See von Davier et al. (2004b, Chapter 4) for a diagnostic measure, called the percent relative error, that compares the moments of the distributions of the equated scores to the moments of the reference distribution. Other commonly used diagnostic measures involve accuracy measures (see below) and historical information available about the equating results from previous administrations of forms of the assessment. One might employ quality assurance methods or statistical process control methods to monitor the stability of the reported scores over time—such as cumulative sum charts, time series analyses, and so on (see Li et al., Chapter 20 of this volume).

The standard error of equating (SEE) and the standard error of equating difference (SEED) are described next. von Davier et al. (2004b) applied the *delta method* (Kendall & Stuart, 1977; Rao, 1973) to obtain both the SEE and the SEED. The delta method was applied to the function from Equation 1.7 that depends on the parameter vectors $\mathbf{r}$ and $\mathbf{s}$ on $T$. According to the delta method, the analytical expression of the asymptotic variance of the equating function is given by

$$\mathrm{Var}[\hat{e}_y(x)] = \mathrm{Var}\{e_y[x; \mathrm{DF}(\hat{\mathbf{u}})]\} \sim \mathbf{J}_{e_y}\mathbf{J}_{DF}\hat{\Sigma}\mathbf{J}_{DF}^{t}\mathbf{J}_{e_y}^{t}, \tag{1.8}$$

where $\hat{\Sigma}$ is the estimated asymptotic variance of the vectors $\mathbf{r}$ and $\mathbf{s}$ after presmoothing; $\mathbf{J}_{e_y}$ is the Jacobian vector of $e_y$, that is, the vector of the first derivatives of $e_y$ ($x$; $\mathbf{r}$, $\mathbf{s}$) with respect to each component of $\mathbf{r}$ and $\mathbf{s}$; and $\mathbf{J}_{\mathrm{DF}}$ is the Jacobian matrix of DF, that is, the matrix of the first derivatives of the design function with respect to each component of vector $\mathbf{u}$.

The asymptotic SEE for $e_y(x)$ is the square root of the asymptotic variance in Equation 1.8, and it depends on three factors that correspond to the data collection and manipulation steps carried out so far: (a) presmoothing (using a log-linear model, for example) through estimating the **r** and **s** and their estimated covariance matrix $\hat{\Sigma}$; (b) the data collection design through the $\mathbf{J}_{\mathrm{DF}}$, and (c) the combination of continuization and the mathematical form of the equating function from Step 4 (computing the equating function) in the OSE framework.

Moreover, the formula given in Equation 1.8 makes obvious the modular character of the OSE framework (and implicitly, of the software package developed for the OSE framework): If one chooses a different log-linear model, then the only thing that will change in the formula given in Equation 1.8 is $\hat{\Sigma}$. If one changes the data collection design, the only thing that will change in the formula given in Equation 1.8 is $\mathbf{J}_{\mathrm{DF}}$. Finally, if one changes the equating method (linear or nonlinear, chained versus frequency estimation, etc.), the only piece that will change in Equation 1.8 is $\mathbf{J}_{e_y}$.

Hence, the formula of the estimated asymptotic variance of the equating function from Equation 1.8, that is,

$$\text{OSE framework} \sim \mathbf{J}_{e_y}\mathbf{J}_{\mathrm{DF}}\hat{\Sigma}\mathbf{J}_{\mathrm{DF}}^{t}\mathbf{J}_{e_y}^{t}, \tag{1.9}$$

could be seen simplistically as the formal representation of the OSE framework.

In addition to the five steps in the equating process described above that are synthesized in Equation 1.9, the OSE framework includes an explicit description of the relationship between the observed-score equipercentile and linear equating functions, which is described below.

### 1.4.6 The Relation Between Linear and Equipercentile Equating Functions

von Davier et al. (2004a, b) argued that all OSE functions from $X$ to $Y$ on $T$ can be regarded as equipercentile equating functions that have the form shown in Equations 1.7 and 1.10:

$$\mathrm{Equi}_{XY\,T}(x) = G_{Tc}^{-1}[F_{Tc}(x)], \tag{1.10}$$

where $F_{Tc}(x)$ and $G_{Tc}(y)$ are continuous forms of the CDFs of $X$ and $Y$ on $T$, and $y = G_{Tc}^{-1}(p)$ is the inverse function of $p = G_{Tc}(y)$. Different assumptions about $F_{Tc}(x)$ and $G_{Tc}(y)$ lead to different versions of $\mathrm{Equi}_{XY\,T}(x)$, and, therefore, to different OSE functions (e.g., chained equating, frequency estimation, etc.).

Let $\mu_{XT}$, $\mu_{YT}$, $\sigma_{XT}$, and $\sigma_{YT}$ denote the means and standard deviations of $X$ and $Y$ on $T$ that are computed from $F_{Tc}(x)$ and $G_{Tc}(y)$, as in $\mu_{XT} = \int x dF_{Tc}(x)$, and so on.

In general, any linear equating function is formed from the first two moments of $X$ and $Y$ on $T$ as

$$\text{Lin}_{XY\,T}(x) = \mu_{YT} + (\sigma_{YT}/\sigma_{XT})(x - \mu_{XT}). \tag{1.11}$$

The linear equating function in Equation 1.11 that uses the first two moments computed from $F_{Tc}(x)$ and $G_{Tc}(y)$ will be said to be compatible with $\text{Equi}_{XY\,T}(x)$ in Equation 1.10. The compatible version of $\text{Lin}_{XY\,T}(x)$ appears in the theorem below (see von Davier et al. 2004a, for the proof of the theorem). The theorem connects the equipercentile function, $\text{Equi}_{XY\,T}(x)$, in Equation 1.10 to its compatible linear equating function, $\text{Lin}_{XY\,T}(x)$, in Equation 1.11.

**Theorem.** *For any population, T, if $F_{Tc}(x)$ and $G_{Tc}(y)$ are continuous CDFs, and $F_0$ and $G_0$ are the standardized CDFs that determine the shapes of $F_{Tc}(x)$ and $G_{Tc}(y)$, that is, both $F_0$ and $G_0$ have mean 0 and variance 1 and*

$$F_{Tc}(x) = F_0\left(\frac{x - \mu_{XT}}{\sigma_{XT}}\right) \text{ and } G_{Tc}(y) = G_0\left(\frac{y - \mu_{YT}}{\sigma_{YT}}\right), \tag{1.12}$$

*then*

$$\text{Equi}_{XY\,T}(x) = G_{Tc}^{-1}[F_{Tc}(x)] = \text{Lin}_{XY\,T}(x) + R(x), \tag{1.13}$$

$$\text{where the remainder term, } R(x), \text{ is equal to } \sigma_{YT}\, r\left(\frac{x - \mu_{XT}}{\sigma_{XT}}\right), \tag{1.14}$$

*and r(z) is the function*

$$r(z) = G_0^{-1}[F_0(z)] - z. \tag{1.15}$$

*When $F_{Tc}(x)$ and $G_{Tc}(y)$ have the same shape, it follows that $r(z) = 0$ in Equation 1.15 for all z, so that the remainder in Equation 1.13 satisfies $R(x) = 0$, and thus $\text{Equi}_{XY\,T}(x) = \text{Lin}_{XY\,T}(x)$.*

It is important to recognize that, for the various methods used in the NEAT design, it is not always true that the means and standard deviations of $X$ and $Y$ used to compute $\text{Lin}_{XY\,T}(x)$ are the same as those from $F_{Tc}(x)$ and $G_{Tc}(y)$ that are used in Equation 1.8 to form $\text{Equi}_{XY\,T}(x)$. The compatibility of a linear and equipercentile equating function depends on both the equating method employed and how the continuization process for obtaining $F_{Tc}(x)$ and $G_{Tc}(y)$ is carried out. The compatibility of linear and nonlinear equating functions does hold for the kernel equating methods but does not hold for all classes of equating methods, as discussed in von Davier, Fournier-Zajack, and Holland (2007). For example, the traditional method of continuization by linear interpolation (Kolen & Brennan, 2004) does not reproduce the variance of the underlying discrete distribution. The piecewise

linear continuous CDF that the linear interpolation method produces is only guaranteed to reproduce the mean of the discrete distribution that underlies it. The variance of the continuized CDF is larger than that of the underlying discrete distribution by 1/12 (Holland & Thayer, 1989). Moreover, the four moments of $X$ and $Y$ on $T$ that are implicitly used by the chained linear or the Tucker linear method are not necessarily the same, nor are they the same as those of the continuized CDFs of frequency estimation or the chained equipercentile methods.

In conclusion, the OSE framework includes the five steps of the equating practice formally described in Equation 1.9 and incorporates both the linear and nonlinear equating functions together with a description of their relationship. The theorem above, which shows that the linear and equipercentile equating methods are related, emphasizes the generalizability of the framework. It was shown that the OSE framework is a statistical modeling framework as described in Figure 1.1, where the unit of measurement is the test score and the equating of scores is accomplished via distribution matching.

## 1.5 Discussion and Outline of the Book

This chapter reviews the existing measurement and equating models for (one-dimensional) tests that measure the same construct. The intention is to have the reader conceptually anchor the new models and approaches presented in the following chapters of the volume into the frameworks outlined in this introduction.

The measurement model presented in Figure 1.1 is the basis for the structure of this volume. In order to reflect the steps in the measurement model as described in Figure 1.1, the book has three parts: (a) *Research Questions and Data Collection Designs*, (b) *Measurement and Equating Models*, and (c) *Evaluation*. The chapters have been grouped to reflect the match between the research methodologies of their focus and each of the steps in the measurement process. The classification of the chapters in these three parts is, of course, approximate; each of the components of the measurement process is addressed in every paper.

**Author Note:** Many thanks go to my colleagues Paul Holland, Jim Carlson, Shelby Haberman, Dan Eignor, Dianne Henderson-Montero, and Kim Fryer for their detailed reviews and comments on the material that led to this chapter. Any opinions expressed in this chapter are those of the author and not necessarily of Educational Testing Service.

http://www.springer.com/978-0-387-98137-6

Statistical Models for Test Equating, Scaling, and Linking

von Davier, A. (Ed.)
2011, XX, 368 p., Hardcover
ISBN: 978-0-387-98137-6