# Chapter 13
# Local Observed-Score Equating

**Wim J. van der Linden**

## 13.1 Introduction

One of the highlights in the observed-score equating literature is a theorem by Lord in his 1980 monograph, *Applications of Item Response Theory to Practical Testing Problems.* The theorem states that observed scores on two different tests cannot be equated unless the scores are perfectly reliable or the forms are strictly parallel (Lord, 1980, Chapter 13, Theorem 13.3.1). Because the first condition is impossible and equating under the second condition is unnecessary, the theorem is rather sobering.

My research on local equating was deeply motivated by Lord's theorem and its related notion of *equity of equating* introduced in the same chapter to explain the "cannot be equated" part of the theorem. Before discussing the principles of local equating, we therefore review the chapter in which the theorem was introduced.

It is quite instructive to see how cautiously Lord (1980) proceeded in the chapter: He began by introducing the problem of observed-score equating under the ideal condition of no measurement error ("case of infallible measures") and used the equipercentile transformation—one of the historic achievements of observed-score equating research—for this case. His next step was the introduction of measurement error ("case of fallible measures"). For this case he gave his famous theorem to show that the use of the equipercentile transformation either does not hold or is unnecessary. Lord then formulated two alternative methods of equating known as item response theory (IRT) observed-score equating and true-score equating. The former deals only indirectly with measurement error by using a parametric estimate of the observed-score distributions for the two tests rather than sample distributions. The latter ignores measurement error altogether. Interestingly, Lord

W.J. van der Linden
CTB/McGraw-Hill, 20 Ryan Ranch Rd., Monterey, CA 93940, USA
e-mail: wim_vanderlinden@ctb.com

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_13, © Springer Science+Business Media, LLC 2011

appeared unable to express a preference for either of these approximate methods, and the chapter ended entirely open, with an intriguing question that I will discuss below.

It is clear that Lord (1980) was aware of the need of observed-score equating as well as the popularity of the methods practiced in his days. On the other hand, although the presentation of the two approximate methods indicates that he was willing to strike a balance between practice and what psychometric theory allows us to do, the open end of the chapter suggests that he was unable to do so.

Lord's (1980) attitude toward observed-score equating reminds me of a cartoon I once saw, in which one scientist said, "Look at the nice application I have!" and the other responded, "Yes, but does it work in theory?" In a field such as test theory, where numbers do not mean anything unless they can be proven to behave according to a model for their formal properties, our affinity should definitely go to the second scientist.

## 13.2 Lord's Analysis of Equating

Lord's (1980) treatment of equating is based on the conceptualization of measurement that underlies IRT—the main topic of his monograph. Key in the conceptualization is the observation that responses to test items reflect not only the ability the test measures but also the properties of the items. Equating is an attempt to disentangle these abilities and item properties at the level of the observed scores on different test forms.

I will follow Lord's (1980) notation and use $\theta$ to denote the ability parameter. In addition, $X$ and $Y$ denote the number-correct scores on two different tests that measure the same $\theta$, and X and Y denote the tests themselves. For convenience, throughout this chapter, tests X and Y are assumed to have equal length. Because $X$ and $Y$ are dependent both on the abilities of the test takers and the properties of the items, an equating problem exists. Suppose that test Y is the newer form and $Y$ has to be equated back to $X$. The goal is to find the transformation $x = \varphi(y)$ from $Y$ to the scale of $X$ that guarantees that the transformed scores on test Y are indistinguishable from the scores on test X.

This conceptualization does not restrict the generality of our analysis in any way; it would do so only if a specific response model were assumed and the results depended on the properties of this model. As each of the mainstream response models used in the testing industry involves a different parameterization of the items, in order to maintain generality, we therefore deliberately avoid specifying any item parameters.

Each equating study involves the choice of a sampling design, but the current conceptualization is also neutral with respect to this choice. For any response model with adequate person and item parameters, we can estimate the parameters in the presence of structurally missing responses. Except for a mild requirement of "connectedness" (van der Linden, 2010), equating based on such models, therefore, does not require a specific equating design.

### *13.2.1 Equating Without Measurement Error*

Lord (1980) introduced the equating problem by considering the case of two perfectly reliable scores $X$ and $Y$ (“case of infallible measures”), a condition under which observed scores are fixed quantities and the distinction between observed and true scores disappears. If $X$ and $Y$ are perfectly reliable scores for tests measuring the same $\theta$, each of these three quantities orders any given population of test takers identically. Consequently, the scores on tests X and Y for any test taker always have the same rank in their distributions for the population of choice. This equivalence of rank establishes an immediate equating relation—if an examinee takes one of the forms, we know that he or she always would obtain the score on the other form associated with the same rank in the population.

In more statistical terms, let $F(x)$ be the (cumulative) distribution function of the scores on test X and $G(y)$ the distribution function of the scores on test Y for an arbitrary population of test takers. Both functions are assumed to be monotonic. For convenience, we also will ignore problems due to the discreteness of number-correct scores throughout this chapter. Let $y$ be the quantile in the distribution on test Y for an arbitrary cumulative proportion $p$ of the population; that is,

$$G(y) = p. \tag{13.1}$$

The equivalent score $\varphi(y)$ on test X follows then from

$$F(\varphi(y)) = p. \tag{13.2}$$

.

Or, making $\varphi(y)$ explicit,

$$\varphi(y) = F^{-1}(G(y)). \tag{13.3}$$

This transformation is the well-known equipercentile transformation in the equating literature. It is typically estimated by sampling the same population twice, administering test forms X and Y to the two samples, estimating the distributions functions of $X$ and $Y$ from the samples, and establishing the relationship by varying $p$ in Equations 13.1 and 13.2 systematically. As the focus of this chapter is not on sampling issues, I do not discuss these issues further.

For perfectly reliable scores, the same transformation from $Y$ to $X$ in Equation 13.3 is obtained for different populations of test takers; that is, use of the equipercentile transformation guarantees population invariance. This invariance is a practical feature, in that it does not seriously restrict equating studies in the choice of their subjects. Also, the choice of population cannot bias the equating in any way: No matter the selection of test takers, the equating errors

$$e_1(x) \equiv \varphi(y) - x \tag{13.4}$$

are always equal to zero for each individual test taker. These two features are documented in the following theorem:

**Theorem 1.** For perfectly reliable test scores $X$ and $Y$, the equipercentile transformation $\varphi(y)$ in Equation 6.3 is (a) unbiased and (b) invariant across populations of test takers with distributions that have the full range of $X$ and $Y$ as support.

These attractive properties of population invariance and error-free equating are immediately lost when we move from the ideal world of infallible measurements to the real world of test scores with errors.

### *13.2.2 Equating With Measurement Error*

In the case of fallible measures, test takers no longer have fixed observed scores on test forms X and Y, but their scores vary across replicated administrations of these tests. Statistically, we therefore should view the observed scores $x$ and $y$ for a test taker as realizations of random variables $X$ and $Y$.

Several things change when scores with measurement error have to be equated. First, it no longer holds that the actually observed scores $X = x$ and $Y = y$ on an administration of the two tests order a given population of test takers identically. Measurement errors distort the ranks of the test takers in the distributions of $X$ and $Y$ for any population; that is, test takers are likely to have a higher rank in one observed-score distribution than dictated by their $\theta$S but a lower rank in another. The principle of equivalence of rank of the scores on test forms X and Y, on which the equipercentile transformation in Equation 13.3 was based, is thus violated and the transformation is no longer valid.

Second, the goal of equating is to find the transformation $\varphi(y)$ from $Y$ to the scale of $X$ that guarantees identical scores. But the criterion can never be met for the case of random errors because these errors introduce nonzero components in the definition of equating error $e_1(y)$ in Equation 13.4. In fact, the problem is even more fundamental in that the definition in Equation 13.4 itself is no longer sufficient: Test scores are now to be viewed as random variables, and it is not enough to just evaluate a single realization of them when the interest should be in their full distribution. Lord (1980) was aware of this problem and replaced the criterion in Equation 13.4 for the case of equating with measurement error by the more general criterion of equity, which he defined intuitively as follows: "If an equating of tests X and Y is to be equitable to each applicant, it must be a matter of indifference to applicants at every given ability level $\theta$ whether they are to take test X or Y" (p. 195).

Lord's formal definition of equity generalizes Equation 13.4 to the requirement for the *full distributions* of the scores on X and Y given $\theta$ and stipulates that

$$f_{\varphi(Y)|\theta} = f_{X|\theta}, \text{ for all } \theta, \tag{13.5}$$

where $f_{\varphi(Y)|\theta}$ and $f_{X|\theta}$ are the probability functions of the transformed scores on test Y and the scores on test X for the chosen population (Lord, 1980, Equation 13.3). This definition of equity is based on a clear concern about fairness of equating: If the two distributions would differ, a test taker might be disadvantaged by taking one test rather than the other. For instance, a high-ability test taker with a larger variance for his or her observed score on test Y than on test X runs a larger risk of not passing a certain cutoff score on the former than the latter.

Thirdly, and lastly, the feature of population invariance of the equipercentile transformation is immediately lost when $X$ and $Y$ have measurement error. This can be shown by deriving their distributions for an arbitrary population with ability distribution $f(\theta)$ as

$$f_X(x) = \int f_{X|\theta}(x)f(\theta)d\theta; \tag{13.6}$$

$$f_Y(y) = \int f_{Y|\theta}(y)f(\theta)d\theta. \tag{13.7}$$

The equipercentile transformation is applied to the marginal distributions $f_X(x)$ and $f_Y(y)$. As test forms X and Y have different items, $f_{X|\theta}(x)$ and $f_{Y|\theta}(y)$ are different. Any change of $f(\theta)$, therefore, has a differential effect on $f_X(x)$ and $f_Y(y)$, and produces a different equating transformation. This is hard to accept for individual test takers who expect their test scores to be adjusted for the differences between the *items* in tests X and Y but actually get a score that depends on the abilities of the *other test takers* who happen to be in the chosen population.

### *13.2.3 Lord's Theorem*

We are now able to discuss Lord's theorem:

**Theorem 2.** Under realistic conditions, scores $X$ and $Y$ on two tests cannot be equated unless either (i) both scores are perfectly reliable or (ii) the two tests are strictly parallel [in which case $\varphi(y) = y$].

As the equipercentile transformation in Equation 13.3 was derived for the case of perfectly reliable scores, the sufficiency of this condition for equipercentile observed equating is obvious. To prove the sufficiency of the second condition (strictly parallel tests), Lord (1980) used the criterion of equity in Equation 13.5 and showed that the criterion only holds for monotonic transformations $x = \varphi(y)$ when the two tests are item-by-item parallel, in which case $\varphi(y) = y$. I will skip the formal proof and refer interested readers to Lord (1980, Section 13.3).

It is important to observe that Lord's proof shows that the only *monotonic* transformation from $Y$ to $X$ for which equity is possible is the identity transformation when the two tests are strictly parallel. It thus makes no sense to look for any other monotonic transformation than the equipercentile transformation that might result in equitable equating. In fact, the following example makes us even wonder if *any* transformation could ever produce an equitable equating for all test takers: Suppose the test scores that need to be equated are for tests with Guttman items at two different locations $\theta_1 < \theta_2$. All $n$ items in test X are located at $\theta_1$, all $n$ items in test Y at $\theta_2$ For test takers with $\theta < \theta_1$, the distributions of $X|\theta$ and $Y|\theta$ are degenerate distributions at $x = 0$ and $y = 0$, respectively; for test takers with $\theta > \theta_2$, they are degenerate distributions at $x = n$ and $y = n$. Hence, for these two groups of test takers, the two tests automatically produce identically distributed scores. However, for $\theta_1 < \theta < \theta_2$, the distributions of $Y|\theta$ remain at $y = 0$ but those of $X|\theta$ are now at $x = n$. For these test takers, the number-correct scores have to be mapped from 0 on test Y to $n$ on test X. Thus, in order to produce an equitable equating, we have to choose between this extreme transformation (and forget about the test takers below $\theta_1$ and above $\theta_2$) and the identity transformation (and forget about those between $\theta_1$ and $\theta_2$).

### *13.2.4 Two Approximate Methods*

Lord (1980) then offered two approximate methods of equating. One method is IRT true-score equating. Let $i$=1, ... , $n$ denote the items in form X and $j$=1,...,$n$ those in form Y. Each of the mainstream response models for dichotomously scored items specifies a probability for the correct response as a function of $\theta$. We use $P_i(\theta)$ and $P_j(\theta)$ for the response probabilities on the items in form X and form Y, respectively. The (number-correct) true scores on forms X and Y are given by

$$\xi = \sum_{i=1}^{n} P_i(\theta), \tag{13.8}$$

$$\eta = \sum_{j=1}^{n} P_j(\theta). \tag{13.9}$$

If the item parameters have been estimated from response data with enough precision, the only unknown quantity in Equations 13.8 and 13.9 is $\theta$. (Because the response model is usually not identified, for the item parameters to be on the same scale they have to be estimated simultaneously from response data for an appropriate sampling design.) Variation of the unknown $\theta$ creates a relation between $\xi$ and $\eta$ that represents $\xi$ as a (monotonic) function of $\eta$. Ignoring the differences between observed scores $X$ and $Y$ and their true scores $\xi$ and $\eta$, IRT true-score equating uses this function to equate $Y$ to $X$.

The other method is IRT observed-score equating. The method is based on an approximation of Equations 13.6 and 13.7 by

$$\hat{f}_X(x) = N^{-1}\sum_{a=1}^{N} f(x\,|\,\widehat{\theta}_a), \tag{13.10}$$

$$\hat{f}_Y(y) = N^{-1}\sum_{a=1}^{N} f(y\,|\,\widehat{\theta}_a), \tag{13.11}$$

where $\hat{\theta}_a$ are the ability estimates for a sample of test takers $a = 1, \ldots, N$. The two estimated marginal distributions of forms X and Y are then used to derive the equipercentile transformation.

### 13.2.5 *An Intriguing Question*

Lord (1980) was doubtful about the use of the method of true-score equating: “We do not know an examinee’s true score. We can estimate his true score….However, an estimated true score does not have the properties of true scores; an estimated true score, after all, is just another kind of fallible observed score” (Lord, 1980, p. 203). But he also had his doubts about the method of IRT observed-score equating: “Is this better than applying…true-score equating…to observed scores $x$ and $y$?”.

Lord (1980) then explained the reason for his inability to choose between the two approximate methods: “At present, we have no criterion for evaluating the degree of inadequacy of an imperfect equating. Without such a criterion, the question cannot be answered” (p. 203). The same uncertainty is echoed in the final section of the chapter, which Lord (1980) began by admitting that practical pressures often require that tests be equated at least approximately. He then summarized as follows: “What is really needed is a criterion for evaluating approximate procedures, so as to be able to choose from among them. *If you can’t be fair (provide equity) to everyone, what is the best next thing?*” (p. 207).

This final question is intriguing. At the time, Lord already must have worked on his asymptotic standard error of equipercentile equating, which was published 2 years later (Lord, 1982b), so he clearly did not refer to this development. Rather than something that only evaluates the effect of sample size (as a standard error does) but leaves the equating method itself untouched, he wanted a yardstick that would allow him to make a more fundamental comparison between alternative equating methods and to assess which would be closest to equity (provide “*the next best thing*”).

## 13.3 Local Equating

Local equating is an attempt to answer Lord’s question. Its basic result is a theorem that identifies an equating that would provide full equity and immediately suggests how to evaluate any actual equating method against this ideal. Also, the theorem involves a twist that forces us to rethink much of our current theory and practice of equating—a process that has led me both to better understanding of the fundamental nature of the observed-score equating problem and a more intuitive appreciation of the idea of local equating. It also suggests new equating methods that better approximate the equity criterion than equipercentile equating. In this section, I review the theorem and provide alternative motivations of local equating. A few new equating methods based on the idea of local equating are discussed in the next section

### *13.3.1 Main Theorem*

The theorem follows directly from the equity criterion in Equation 13.5. Lord (1980) expressed the criterion as an equality of conditional probability functions. Equivalently, it could be expressed as an equality of the conditional distribution functions $F_{\varphi(Y)|\theta}$ for the equated scores on test Y and $F_{X|\theta}$ for the observed score on test X. However, rather than as an equality, we express the criterion as a definition of equating error,

$$e_2(x;\theta) \equiv F_{\varphi(Y)|\theta} - F_{X|\theta}, \tag{13.12}$$

and require all error to be equal to zero for all $\theta$. The transformations $x = \varphi^*(y)$that solve this set of equations are the error-free or true equating transformations.

Thus, it should hold that

$$F_{X|\theta}(x) = F_{\varphi(Y)|\theta}(\varphi(y)), \quad \theta \in R. \tag{13.13}$$

Solving for $x$ by taking the inverse of $F_{X|\theta}$,

$$x = \varphi^*(y;\theta) = F^{-1}_{X|\theta} F_{\varphi(Y)|\theta}(\varphi(y)), \quad \theta \in R. \tag{13.14}$$

However, because $\varphi(\cdot)$ is monotone, $F_{\varphi(Y)|\theta}(\phi(y)) = F_{Y|\theta}(y)$. Substitution results in

$$\varphi^*(y;\theta) = F^{-1}_{X|\theta}(F_{Y|\theta}(y)), \quad \theta \in R, \tag{13.15}$$

as the family of true equating transformations.

Surprisingly, Equation 13.15 involves the same type of transformation as for the equipercentile equating in Equation 13.3, but it is now applied to each of the conditional distributions of $X|\theta$ and $Y|\theta$ instead of only once to the marginal distributions of $X$ and $Y$ for a population of test takers. The fact that the derivation leads to an entire family of transformations reveals a rather restrictive implicit assumption in Lord's theorem, as well as all of our traditional thinking about equating: namely, that the equating should be based on a single transformation for the entire population of choice. Relaxing the assumption to different transformations for different ability levels opens up a whole new level of possibilities for observed-score equating that is waiting to be explored. The following theorem is offered as an alternative to Lord's (for an extended version, see van der Linden, 2000):

**Theorem 3.** For the population of test takers $P$ for which test scores $X$ and $Y$ measure the same ability $\theta$, equating with the family of transformations $\varphi^*(y;\theta)$ in Equation 13.15 has the following properties: (i) equity for each $p \in P$; (ii) symmetry in $X$ and $Y$ for each $p \in P$; and (iii) invariance within $P$.

**Proof.** (i) For each $p \in P$ there is a corresponding value of $\theta$, and for each $\theta$ the transformation in Equation 13.15 matches the conditional distributions of $\varphi$*(Y) and $X$ given $\theta$. (ii) The inverse of $F^{-1}_{X|\theta}F_{Y|\theta}(y)$ is $F^{-1}_{Y|\theta}F_{X|\theta}(x)a$, which is Equation 13.15 for the equating from $X$ to $Y$. (iii) The conditional formulation of Equation 13.15 implies independence from the distribution of $\theta$ over $P$. As a consequence, the family holds for any subpopulation of $P$.

In addition to equity, the family of transformations thus has the properties of symmetry and population invariance—other criteria identified by Lord (1980, Section 13.5) as essential to equating. The criterion of symmetry is usually motivated by observing that it would be hard to understand why a reversal of the roles of $X$ and $Y$ should lead to a different type of equating. It should—and does—hold for the definition of the true equating transformations in Equation 13.15. When selecting an actual method in an equating study, we sometimes are faced with trade-offs between the three criteria, and it then makes sense to sacrifice some symmetry to get closer to the more desirable property of equity. As we shall see later, the same choice is made for some of the traditional methods of equating.

As for the issue of population invariance, the criterion of equity in Equation 13.5 is defined conditional on $\theta$. Hence, if the criterion holds, it automatically holds for any subpopulation of $P$ as well. But the criterion also implies the definition of the family of transformations in Equation 13.15. It follows that *equity is a sufficient condition for population invariance* within $P$. This conclusion implies that an effective attempt to get closer to population invariance is approximating equity.

Also, note that the theorem defines the ultimate population $P$ for which the invariance holds as the population of persons for which tests X and Y measure the same $\theta$. We have a clear empirical criterion to evaluate membership of $P$: the joint fit of the response model in the testing program for the two tests. Besides, although the definition of $P$ excludes arbitrariness, it is nevertheless open in that it not only includes all past or current test takers whose response behavior fit the model but encompasses future test takers for which this can be shown to hold as well. Finally, unlike traditional observed-score equating, the definition of $P$ does not entail any necessity of random sampling of test takers.

The error definition in Equation 13.12 implies the ideal or true equating that provides equity but also offers the "criterion for evaluating approximate procedures" that Lord (1980) wanted so badly: For any arbitrary transformation $\varphi(y)$, the criterion is just the difference between the conditional distribution functions for the equated scores $\varphi(Y)|\theta$ and the scores $X|\theta$ in Equation 13.12. Observe that the difference is a function of $x$ and that we have a different function for each $\theta \in R$. Also, because of its conditioning on $\theta$, the evaluation is population invariant within $P$—an evaluation of the equated scores $\varphi(Y)$ for any subpopulation of $P$ automatically holds for any other subpopulation.

Alternatively, we can compare any given transformation $\varphi(y)$ directly with the family of true transformations $\varphi^*(y;\theta)$ in Equation 13.15. This comparison leads to the alternative family of error functions:

$$\begin{aligned} e_3(y;\theta) &= \varphi(y) - \varphi^*(y;\theta) \\ &= \varphi(y) - F^{-1}_{X|\theta}(F_{Y|\theta}(y)), \quad \theta \in R. \end{aligned} \tag{13.16}$$

Of course, the results from both evaluations are equivalent: An equating transformation is error free if and only if its equated scores are. A critical difference between Equations 13.12 and 13.16, however, exists with respect to the scale on which they are defined: The error functions in Equation 13.12 are functions of $x$ but those in Equation 13.16 are functions of $y$. The former are convenient when we have to evaluate an equating from a test Y with a variable composition to a fixed form X, for instance, from an adaptive to a linear test. For a more extensive discussion of these two alternative families of error functions, see van der Linden (2006a, b).

The definition of equating error is only the first step toward a standard statistical evaluation of observed-score equating. For the implementations of local equating discussed later in this chapter, the error functions above will be used to define the *bias* and *mean-square error* functions of an equating, that is, the expectations of the error and squared error over essential random elements in the implementation. These additional steps take the evaluation of equating to the same level as, for instance, the standard evaluation of an estimator of an unknown parameter or a decision rule in statistics.

In principle, we are now ready to look for equating methods that approximate the family in Equation 13.15 as closely as possible and evaluate these methods using these statistical criteria. The challenge, of course, is to find a proxy of the unknown $\theta$ that takes us as closely as possible to the true member in the family for each test taker. Before exploring the possibilities, I motivate the idea of local equating from a few alternative points of view.

### 13.3.2 Alternative Motivations of Local Equating

All of current observed-score equating is based on the use of a single transformation. However, the example at the end of the discussion of Lord's theorem above already hinted at the fact that no transformation whatsoever could ever establish an equitable equating at each ability level for a population of test takers. The following thought experiment illustrates this point again (van der Linden & Wiberg, in press). Suppose a person $p$ with ability level $\theta_p$ takes test form Y, and a test specialist is asked to equate his or her observed score $y_p$ to a score on test form X. For the sake of argument, suppose the specialist is given the full observed-score distributions for $\theta_p$ on both tests, that is, $F_X(x|\theta_p)$ and $F_Y(y|\theta_p)$. For this single-person population, an obvious choice from a traditional point of view is to use the equipercentile transformation $x = \varphi_p(y) = F^{-1}_{X|\theta_p}(F_{Y|\theta_p}(y))$ to equate the observed score $y_p$ to a score on form X. Now suppose a second person $q$ with another ability level takes the same test, and the same specialist is asked to equate this person's observed score $y_q$. The specialist, who is also given the distribution functions for $q$, is then faced with the choice between (a) using a separate equipercentile transformation for $q$ or (b) treating the two test takers as a new population and using the equipercentile transformation for the marginal distributions of it. The first option would only involve establishing another individual transformation $\varphi_q(y)$, analogous to $\varphi_p(y)$. The result would be an equitable, symmetric, and population-invariant equating for both test takers. The second option would require the marginal distribution functions for the population, which is the average of the separate functions for $p$ and $q$. Letting $F'_X(x)$ and $F'_Y(y)$ denote the two averages, the alternative equating transformation would be $x = \varphi'(y) = {F'_X}^{-1}(F'_Y(y))$. This second option would miss all three features. In fact, its problem would become even more acute if we kept adding test takers to the population: For each new test taker, the equating transformation would change. Even more embarrassing, the same would happen to the equated scores of all *earlier* test takers.

Clearly, traditional equipercentile equating involves a compromise between the different transformations required for the ability levels of each of the test takers in an assumed population. In doing so, it makes systematic errors for each of them. In more statistical terms, we can conclude that the use of a single equating transformation for different ability levels involves equatings that are structurally biased for each of them. The error function in Equation 13.12 reflects the size of the bias for each individual test taker.

The history of test theory shows an earlier occasion where a similar choice had to be made between a one-size-fits-all approach and one based on individual ability levels—the choice of the standard error of measurement for a test. The classical standard error was a single number for an entire population of test takers derived from the reliability of the test. It was quickly recognized that this error was a compromise between the actual errors at each ability level and was thus always biased. For example, a test that matches an individual test taker's ability level is known to be more informative than one that is much too difficult or too easy—a fact that should be reflected in the standard errors for the individual test takers. The classical standard error is now widely replaced by the conditional standard deviation of the observed score given ability, that is,

$$[Var(X \mid \theta)]^{1/2}. \tag{13.17}$$

The family of true equating transformations $\varphi^{*}(y;\theta)$ in Equation 13.15 is based on the full conditional distributions of the observed test scores, of which this conditional standard error represents the dispersion.

Interestingly, the family $\varphi^{*}(y;\theta)$ also can be shown to generalize Lord's first approximate equating method—the true-score equating in Equations 13.8–13.9. The equating following from this set of two equations is usually presented as a table with selected pairs of values of $\eta$ and $\xi$ used to equate the observed scores on form Y to $X$. It is tempting to think of this format as the representation of a single equating transformation. However, this conclusion would overlook that Equations 13.8–13.9 actually are a system of *parametric* equations, that is, a family of mappings with $\theta$ as index. When applied to equate $Y$ to $X$, it becomes the family of true equating transformations in Equation 13.15 with its distributions degenerated to their expected values $E(X \mid \theta)$ and $E(Y \mid \theta)$:

$$E(Y|\theta) \to E(X|\theta), \quad \theta \in R. \tag{13.18}$$

Obviously, much is to be gained when we avoid this degeneration and turn to an equating based on the full conditional distributions of $X$ and $Y$.

It is also instructive to view Lord's (1980) second approximate method in Equations 13.10–13.11 from the perspective of the family of true equatings in Equation 13.15. This method substitutes ability estimates $\widehat{\theta}_a$ for the test takers in a sample $a = 1, \ldots, N$ into the set of equations for the marginal distributions of $X$ and $Y$ in Equations 13.6–13.7. However, as already indicated, the factors $f_{X|\theta}(x)$ and $f_{Y|\theta}(y)$ in these equations are for different items, and any change of population $h(\theta)$ (or sample of test takers in this approximate method) has a differential effect on $f_X(x)$ and $f_Y(y)$ and therefore produces a different equipercentile transformation. An effective solution to this problem of population dependency is to just ignore the common second factor $h(\theta)$ in the integrands in Equations 13.6–13.7 and base IRT observed-score equating only on their first factors $f_{X|\theta}(x)$ and $f_{Y|\theta}(y)$ for the estimates $\widehat{\theta}_a$, precisely the choice made in the first local equating method discussed later in this chapter.

On the other hand, the traditional approach to the problem of population dependency has been to identify some special population $h(\theta)$ and use this as a standard for the equating. Two versions of the approach exist. One is based on the idea of a synthetic population to be derived from the two actual populations that take tests X and Y. Braun and Holland (1982, Section 3.3.2), who introduced the notion, defined it as any population with a distribution function equal to a linear combination of the functions for the two separate populations. More formally, if $F_X(\theta)$ and $\mathrm{F}_Y(\theta)$ are the distribution functions for the populations who take tests X and Y, the synthetic population has distribution function

$$wF_X(\theta) + (1-w)F_Y(\theta), \tag{13.19}$$

with $0 \leq w \leq 1$ a weight to be specified by the testing program. The definition could be justified by two-stage sampling of the test takers in the equating study from the separate populations for tests X and Y with weights *w* and 1-*w*. However, this type of weighted sampling is rarely used in this context. More importantly, equatings are always required only for the scores of the population that takes the new test form, Y, and any nonzero weight *w* would detract from this goal (van der Linden & Wiberg, in press).

The other approach recognizes this fact and uses the population for Y as the standard. It does so by identifying the critical variables on which the populations for X and Y differ and using them to resample the population for X to match the population for Y. The two matched distributions are then used in the actual equating. For evaluations of this approach with matched samples, see, for instance, Dorans (1990); Dorans, Liu, and Hammond (2008); Liou, Cheng, and Li (2001); and Wright and Dorans (1993).

The use of synthetic or matched populations does not take population dependency away. For each test taker, it still holds that the equated score depends on the abilities of the other test takers in these synthetic or matched populations. Rather than equating $X$ and $Y$ for populations with identical ability distributions, as these two approaches attempt, we should equate them for identical abilities (i.e., condition on ability). Finally, notice that the use of a matched population also implies loss of symmetry of the equating. Except for the case of weight $w = .5$, the same holds for the use of a synthetic population.

At first sight, local equating may seem liable to two different objections, one involving an issue of fairness and the other being more philosophical. The former has to do with the fact that local equating implies different equated scores for the same score $Y = y$ by test takers with different abilities. This different treatment of equal observed scores seems unacceptable. However, the following example shows that actually the opposite holds: Consider the case of two test takers *p* and *q* who both have a score of 23 items correct on a 30-item test Y. Traditional equipercentile equating routinely would give both test takers the same equated score, a higher score than 23 if test Y appears to be more difficult than test X and a lower score if it appears to be easier. Now, suppose we are told that *p* and *q* have the observed-score distributions in Figure 13.1. As the figure reveals, the score observed for *q* was in the lower tail of *q*'s distribution. However, *p* had better luck; *p*'s score was in the upper tail of the distribution. Would it be fair to give the two individuals the same equated score on test X? Or should we adjust their equated scores for measurement error? After all, we do live in a world of fallible measures.

The critical question, of course, is where our knowledge of the abilities and the observed-score distributions of the test takers could come from. The true challenge to local equating lies in the answer to this question, not in any of the conceptual or more formal issues we have dealt with so far. But in fact we often know more than we realize. For instance, in observed-score equating, we generally ignore the information in the response patterns that leads to the observed scores. The example in Figure 13.1 typically arises when two test takers have equal number-correct scores but one fails on some of the easier items and the other on some of the more difficult ones. We immediately return to this important question in the next section.

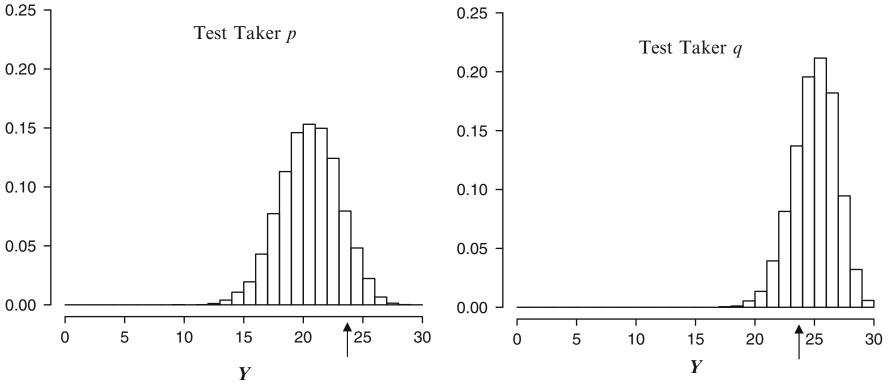


**Fig. 13.1** Example of two test takers $p$ and $q$ with different abilities but the same realized observed score $Y = 23$

The more philosophical issue regards the question of how we seriously could propose using different equating transformations for a single measurement instrument. No one would ever consider doing this, for instance, when a tape measure appears to be locally stretched and its (monotonically) distorted measurements need to be equated back to those by a flawless measure. The idea of using different transformations to equate identical measurements on the distorted scale back to the standard scale would seem silly. Why, then, propose this for number-correct score equating in testing? This question is problematic because of its implicit claim of the number-correct score as a measure with the same status as length measured by a tape measure. Number-correct scores are entirely different quantities, though. Unlike length measures, they are not fundamental measures, which always can be reduced to a comparison between the object that is measured and a concatenation of standard objects (e.g., an object on one scale and a set of standard weights on the other). They are also not derived from such measures. (For a classic treatment of fundamental and derived measurement, see Campbell, 1928). More surprisingly, perhaps, although defined as counts of correct responses, number-correct scores are not counting measures, either. They would only be counting measures if all responses were equivalent. But they are not—each of them always is the result of an interaction between a different combination of ability level and item properties.

This last fundamental fact was already noted when I introduced Lord's (1980) notation for observed-score equating in the beginning of this chapter and stated, "Because $X$ and $Y$ are dependent both on the abilities of the test takers and the properties of the items, an equating problem exists." An effective way of disentangling ability and item effects on test scores is to model them at the level of the item-person combinations with separate item and person parameters, as IRT does. Observed-score equating is an attempt to deal with the same problem at the level of test scores in the form of a score transformation. But before applying any transformation to adjust for the differences between the items in different tests, we have to condition on the abilities to get rid of their effects. Monotonic transformations

$x = \varphi(y)$ that adjust simultaneously for item and ability effects on observed test scores on tests X and Y do not exist.

## 13.4 A Few Local Equating Methods

According to Lord's theorem, observed-score equating is possible only if the scores on forms X and Y are perfectly reliable or strictly parallel. On the other hand, Theorem 3 in this chapter shows that equating under regular conditions is still possible, provided we drop the restriction of a single transformation for all ability levels.

It may seem as if Theorem 3 only replaces one kind of impossible condition (perfect reliability or strictly parallelness) by another (known ability). However, an important difference exists between them. *Post hoc* changes of the reliability and the degree of parallelness of test forms are impossible; when equating the scores on a test form, we cannot go back and make them more reliable or parallel. As a result, Lord's theorem leaves us paralyzed; it offers no hint whatsoever as to what to do when real-world tests have less than perfect reliability or are not parallel. On the other hand, we can always try to approximate the family of true equating transformations in Equation 13.15 using whatever information is available in the test administration or equating study. Clearly, the closer the approximation, the better the equating. In fact, even a rough estimate or a simple classification of the abilities may be better than combining them into an assumed population before conducting the equating.

The name *local equating* is derived from the attempt to get as close as possible to the true equating transformations in Equation 13.15 to perform the equating. The error definitions in Equation 13.12 or Equation 13.16 can be used to evaluate methods based on such attempts in terms of their bias and mean standard error using a computer simulation with response data generated for known abilities under a plausible model.

Now that we know the road to equitable, population-independent equating, and have the tools to evaluate progress along it, we are ready to begin a search for Lord's "next best thing." The local equating methods below are first steps along this road. I only review their basic ideas and show an occasional result from an evaluation. More complete treatments and discussions of available results are found in the references.

### *13.4.1 Estimating Ability*

The first method is a local alternative to the IRT observed-score equating method in Equations 13.10–13.11. It follows the earlier suggestion to obtain population-independent equating by ignoring the common second factor $h(\theta)$ in Equations 13.6–13.7 and basing the equating entirely on their first factors, $f_{X|\theta}(x)$ and $f_{Y|\theta}(y)$.

The main feature of this method is estimation of $\theta$ under a response model that fits the testing program, substituting the estimate in the true equating in Equation 13.15. In fact, the procedure is entirely analogous to the use of the conditional standard error of measurement in Equation 13.17, which also involves substitution of a $\theta$ estimate when used in operational testing.

For dichotomously scored test items, the conditional distributions of $Y$ and $X$ given $\theta$ belong to the generalized binomial family (e.g., Lord, 1980, Section 4.1). Unlike the regular binomial family, its members do not have distribution functions in closed form but are given by the generating function

$$\prod_{i=1}^{n} [Q_i(\theta) + tP_i(\theta)], \tag{13.20}$$

where $P_i(\theta)$ is the success probability on item $i$ for the response model in the testing program and $Q_i(\theta){=}1{-}P_i(\theta)$. Upon multiplication, the coefficients of the factors $t^1$, $t^2$,... in the expression are the probabilities of $X = 1, 2, \ldots$. The probabilities are easily calculated for forms X and Y using the well-known recursive procedure in Lord and Wingersky (1984). From these probabilities, we can calculate the family of true equating transformations in Equation 13.15. Thus, the family can be easily calculated for any selection of $\theta$s as soon as the items in forms X and Y have been calibrated for the testing program.

The estimates of $\theta$ can be point estimates, such as maximum-likelihood estimates assuming known item parameters or Bayesian expected a posterior estimates. But we could also use the full posterior distribution of $\theta$ for the test taker's response vector on form X to calculate his or her posterior expectation of the true family in Equation 13.15. However, this alternative is more difficult to calculate and has not shown to lead to any significant improvement over the simple procedures with a point estimate of $\theta$ plugged directly into Equation 13.15. More details on this local method are given in van der Linden (2000, 2006a).

The local method of IRT observed-score equating lends itself nicely to observed-score equating problems for test programs based on a response model. Another natural application is the equating of an adaptive test to a reference test released to its test takers for score-reporting purposes. In adaptive testing, $\theta$ estimates are immediately available. Surprisingly, this proposed equating of the number-correct scores on an adaptive test is entirely insensitive to the fact that different test takers get different selections of items; the use of the true equating transformations for the test takers' item selections at their $\theta$ estimates automatically adjusts both for their ability differences and the selection of the items (van der Linden, 2006b).

Observe that two different summaries of the information in the response patterns on test Y are used: number-correct scores and $\theta$ estimates. The latter picks up the information ignored by the former. The earlier example for the two test takers with the same number of items correct on Y in Figure 13.1, used to illustrate that they nevertheless deserved different equated scores, was based precisely on this alternative type of IRT observed-score equating.

For later comparison, it is also interesting to note the different use of the conditional distribution functions in traditional and local IRT observed-score equating. In both versions, $\theta$ estimates of the test takers and the conditional distributions of $X$ and $Y$ given these estimates are calculated from Equation 13.20. In the traditional version, the conditional distributions are then averaged over the sample of test takers to get an estimate of the marginal distributions for assumed populations on $X$ and $Y$, and the equipercentile transformation is calculated for these marginal distributions (e.g., Zeng & Kolen, 1995). In the local version of the method, no averaging takes place, but different equipercentile transformations are calculated directly for the different conditional distributions of $X$ and $Y$ given $\theta$.

Figure 13.2 shows a typical result from a more extensive evaluation of the method of local IRT observed-score equating against traditional equipercentile equating in van der Linden (2006a). The curves in the two plots show the bias functions based on the responses on two 40-item tests X and Y simulated under the three-parameter logistic response model for the simulated values $\theta = -2.0, -1.5, \ldots, 2.0$ (curves more to the left are for lower θ values). The bias functions were the expectations of the error functions in Equation 13.16 across the simulated observed-score distributions on test Y given $\theta$. (The mean standard error functions in this study, which were the expectations of the squares of the same errors, are omitted here because they showed identical patterns of differences.) For the local method, the bias was ignorable. But the bias for the traditional method went up to 4 score points (i.e., 10% of the score range) for some combinations of θs and observed scores. For an increase in test

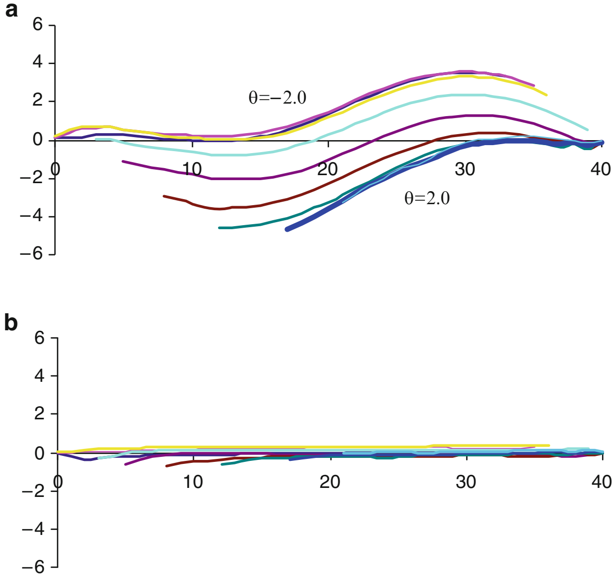


**Fig. 13.2** Bias functions for (**a**) traditional equipercentile and (**b**) local item response theory (IRT) observed-score equating for $\theta = -2.0(.5)20$

length, the bias for the traditional method became even worse, but for the local method it vanished because of better estimation of $\theta$. For the same reason, the bias decreased with the discrimination parameters of the items in test Y. Likewise, the local method appeared to be insensitive to differences in item difficulty between tests in X and Y because $\theta$ estimates have this property.

### 13.4.2 Anchor Score as a Proxy of Ability

The traditional methods for observed-score equating for a nonequivalent-groups-and-anchor test (NEAT) design are chain equating and equating with poststratification. The former consist of equipercentile equating from $Y$ to the observed score $A$ on an anchor test A for the population that takes form Y with subsequent equating from $A$ to $X$ for the population that takes form X. The equating transformation from $Y$ to $X$ is the composition of the separate transformations for these two steps. In equating with poststratification, the conditional distributions of $X$ and $Y$ given $A = a$ are used to derive the distributions on forms X and Y for a target population, usually a population that is a synthesis of those that took the two forms as in Equation 13.19, and the actual equating is equipercentile equating of the distributions for this target population (von Davier, Holland, & Thayer, 2004b, Section 2.4.2).

As a local alternative to these traditional methods, it seems natural to use the extra information provided by the anchor test to approximate the true family of equating transformations in Equation 13.15. For simplicity, we assume an anchor test A with score $A$ that is not part of $Y$ (“external anchor”). For an equating with an internal anchor, we just have to add the score on this internal anchor to the equated score derived in this section.

For an anchor test to be usable, $A$ has to be a measure of the same $\theta$ as $X$ and $Y$. Formally, this means a classical true score $\tau_A \equiv E(A)$ that is a monotonic increasing function of the same ability $\theta$ as the true scores for $X$ and $Y$. The exact shape of the function, which in IRT is known as the *test characteristic function*, depends on the items in A as well as the scale chosen for $\theta$. It should thus hold that $\tau_A = g(\theta)$ where $g$ is an (unknown) monotonically increasing function and $\theta$ is the same ability as for $X$ and $Y$.

An important equality follows for the conditional observed-score distributions in the true equating transformations in Equation 13.15. For instance, for the distribution of $X$ given $\theta$ it holds that

$$f(x \mid \theta) = f(x \mid g^{-1}(\tau_A)) = f(x \mid \tau_A). \tag{13.21}$$

Similarly, $f(y|\theta) = f(y|\tau_A)$. Thus, whereas $\theta$ and the true score on the anchor test are on entirely different scales, the observed-score distributions given these two quantities are always identical.

This fact immediately suggests an alternative to the local method in the preceding section. Instead of using an estimate of $\theta$ for each test taker, we could use an estimate of $\tau_A$ and, except for estimation error, get the same equating. An obvious estimate of $\tau_A$ is the observed score $A$. The result is a simple approximation of the family of true transformations in Equation 13.15 by

$$\varphi(y;a) = F^{-1}_{X|a}(F_{Y|a}(y)), \quad a = 0, ..., m, \tag{13.22}$$

where $m$ is the length of the anchor test and $F_{X|a}(x)$ and $F_{Y|a}(y)$ are the distribution functions of $X$ and $Y$ given $A = a$. Local equating based on this method is easy to implement; it is just equipercentile equating directly from the conditional distributions of $Y$ to those of $X$ given $A = a$.

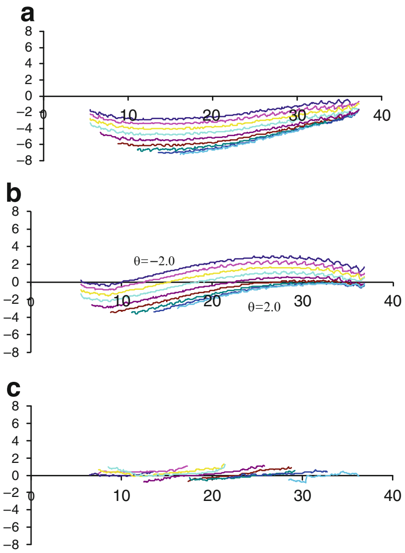
It is interesting to compare the use of the different observed-score distributions available in the NEAT design between the two traditional methods and this local method:

1. In chain equating, the equipercentile transformation is derived from four different population distributions, namely, the distributions of $X$ and $Y$ for the populations that take tests X and Y and the distributions of $A$ for the same two populations.
2. In equating with poststratification, the conditional distributions of $X$ and $Y$ given $A = a$ are used to derive the marginal distributions of $X$ and $Y$ for a target population. The equating transformation is applied to these two distributions.
3. The current method of local equating directly uses the conditional distributions of $X$ and $Y$ given $A = a$ to derive the family of equating transformations in Equation 13.22.

The only difference between the previous method of local equating based on maximum likelihood or Bayesian estimation of $\theta$ and the use of the anchor test scores $A = a$ as a proxy of $\theta$ resides in the estimation or measurement error involved. (I use the term *proxy* instead of *estimate* because, due to scale differences, $A$ is not a good estimate of $\theta$.) These errors have two different consequences. First, for both equatings they lead to a mixing of the conditional distributions in Equation 13.15 that actually should be used. For direct estimation of $\theta$, the mixing is over the distribution of $\theta$ given the estimate, $\theta|\widehat{\theta}$. But for Equation 13.22, it is over $\theta \mid A = a$ where $\theta = g^{-1}(\tau_a)$. The former can be expected to be narrower than the latter, which is based on less accurate number-correct scoring. The impact of these mixing distributions, which generally depend on the lengths of X and A as well as the quality of their items, requires further study. But it is undoubtedly less serious than the impact of mixing the conditional distributions on forms X and Y over the entire marginal population distribution $f(\theta)$ in Equations 13.6–13.7, on which the traditional methods are based. Second, in the current local method, the conditional distributions of $X$ and $Y$ given $A = a$ are estimated directly from the sample, whereas in the preceding method they are estimated as the generalized binomial distributions in Equation 13.20. For smaller sample sizes, the former will be less accurate.

Figure 13.3 shows results from the evaluation of the chain-equating, poststratification and local method for a NEAT design in van der Linden and Wiberg

**Fig. 13.3** Bias functions for (**a**) chain equating, (**b**) poststratification equating, and (**c**) local equating for the nonequivalent-groups-and-anchor test (NEAT) design for $\theta = -2.0(.5)20$



(in press). The results are for a study with the same setup as for Figure 13.2 but with a 40-item anchor test added to the design. Again, the local method outperformed the two traditional methods. But it had a slightly larger bias than the local method in the preceding section, because of the less favorable mixing of the conditional distributions of $X$ and $Y$ given $\theta$ when $A$ is used as a proxy of $\theta$. However, the more accurate the proxy, the narrower the mixing distributions. Hence, as also demonstrated in this study, the bias in the equated scores vanishes with the two main determinants of the reliability of $A$—the length of the anchor test and the discriminating power of its items. In this respect, the role of the anchor test in the current method is entirely comparable to that of test form Y from which $\theta$ is estimated in the preceding method.

For testing programs that are response-model based, Janssen, Magis, San Martin, and Del Pino (2009) presented a version of local equating for the NEAT design with maximum-likelihood estimation of $\theta$ from the anchor test instead of the use of $A$ as a proxy for it. The empirical results presented by these authors showed bias functions for this alternative method that are essentially identical to those in Figure 13.2 and better than those in Figure 13.3. Janssen et al. also explained this difference in performance by the fact that maximum-likelihood estimation of $\theta$ from A did a better job of approaching the intended conditional distributions of $X$ and $Y$ given $\theta$ than the use of number-correct anchor scores.

The study that produced the results in Figure 13.3 did not address the role of sampling error in the estimation of the conditional distributions of $X$ and $Y$ given $A = a$. For small samples, the error will be substantial. A standard approach to small-sample equating for NEAT designs, especially if the main differences between the distributions of the observed scores on forms X and Y are in their first and second moments, is linear equating in the form of Tucker, Levine, or linear chain equating (Kolen & Brennan, 2004, Ch. 4). The use of local methods for linear equating is explored in Wiberg and van der Linden (2009). One of their methods uses the conditional means, $\mu_{X|a}$ and $\mu_{Y|a}$, and standard deviations, $\sigma_{X|a}$ and $\sigma_{Y|a}$, of $X$ and $Y$ given $A = a$ to conduct the equating. The result is the family of transformations

$$x = \varphi(y; a) = \mu_{X|a} + \frac{\sigma_{X|a}}{\sigma_{Y|a}}(y - \mu_{Y|a}), \quad a = 0, ..., m. \tag{13.23}$$

In an empirical evaluation, the method yielded better results than the traditional Tucker, Levine, and linear chain equating methods but also improved on Equation 13.22 because of its reliance only on estimates of the first two moments instead of the full conditional distributions of $X$ and $Y$ given $A = a$.

So far, no explicit smoothing has been applied to any local equating method. The application of smoothing techniques should reduce the impact of sampling error in the estimation of the conditional distributions of $X$ and $Y$ given $A = a$ for the NEAT design to be considerable, especially for the techniques of presmoothing of observed-score distributions proposed in von Davier et al. (2004b, Chapter 3).

### 13.4.3 $Y = y$ as a Proxy of Ability

The argument for the use of anchor score $A$ as a proxy for $\theta$ in the previous section holds equally well for the realized observed score $Y = y$. The score can be assumed to have a true score $\eta$ that is a function of the same ability $\theta$ as the true score on form X; see Equation 13.8–13.9. Again, scale differences between conditioning variables do not matter, and we can just focus on the distributions of $X$ and $Y$ given $\eta$ instead of $\theta$. As $Y = y$ is an obvious estimate of $\eta$, it seems worthwhile exploring the possibilities of local equating based on the conditional distributions of $X$ and $Y$ given $Y = y$ that is, use

$$\varphi(y) = F^{-1}_{X|y}(F_{Y|y}(y)), \quad y = 0, ..., n. \tag{13.24}$$

In an equating study with a single-group design, the distributions of $X$ given $Y = y$ can be estimated directly from the bivariate distribution of $X$ and $Y$ produced by the study. The distributions of $Y$ given $y$ are more difficult to access. In fact, they are only observable for replicated administrations of form Y to the same test takers. However, Wiberg and van der Linden (2009) identified one case for which replications are unnecessary—linear equating conditional on $Y = y$. For this case, the general form of the linear transformation for observed-score equating specifies to

$$x = \varphi(y) = \mu_{X|y} + \frac{\sigma_{X|y}}{\sigma_{Y|y}}(y - \mu_{Y|y}), \quad y = 0, ..., n. \tag{13.25}$$

As classical test theory shows, $\mu_{Y|y} = y$. Hence, the family simplifies to

$$x = \varphi(y) = \mu_{X|y}, \quad y = 0, ..., n. \tag{13.26}$$

For all test takers with score $Y = y$, this local method thus equates the observed scores on Y to their conditional means on X.

In spite of the standard warning against the confusion of equating with regression in the equating literature (e.g., Kolen & Brennan, 2004, Section 2.3), the local linear equating in Equation 13.26 has the same formal structure as the (nonlinear) regression function of $Y$ on $X$. Actually, however, Equation 13.26 is a family of degenerate mappings with index $y$, just like the family for IRT true-score equating in Equation 13.18. (In fact, the equating in Equation 13.26 follows directly from Equation 13.18 if we substitute $y$ as proxy for $\theta$.) Although it is thus incorrect to view Equation 13.26 as a direct postulate of the use of the regression function of $X$ on $Y$ for observed-score equating, the formal equivalence between the two is intriguing. Apparently, the fact that we allow for measurement error in $X$ and $Y$ when equating does force us to rethink the relation between equating and regression.

An evaluation of Equation 13.26 showed a favorable bias only for the higher values of $y$ (Wiberg & van der Linden, 2009). Because the responses were simulated under the three-parameter logistic model, the larger bias at the lower values of $\theta$ should be interpreted as the effect of guessing for low-ability test takers—a phenomenon known to trouble traditional equipercentile equating as well. This bias problem has to be fixed before practical use of the local method in this section can be recommended.

### *13.4.4 Proxies Based on Collateral Information*

In principle, every variable for which the expected or true scores for the test takers are increasing functions of the $\theta$ measured by forms X and Y could be used as a proxy to produce an equating. The best option seems collateral information directly related to the performances by the test takers on X or Y, such as the response times on the items in Y or scores on a earlier related test. However, the use of more general background variables, such as earlier schooling or socioeconomic factors, should be avoided because of the immediate danger of social bias.

Empirical studies with these types of collateral information on $\theta$ have not yet been conducted. Of course, different sources of collateral information will yield equatings with different statistical qualities. But the only thing that counts is rigorous evaluation of each of these qualities based on the definitions of equating error in Equations 13.12 and 13.16. These evaluations should help us to identify the best feasible method for an equating problem.

## 13.5 Concluding Remarks

The role of measurement error has been largely ignored in the equating literature. When I had the opportunity to review two new texts on observed equating that now have become standard references for every specialist and student in this area, I was impressed by their comprehensiveness and technical quality but missed the necessary attention to measurement error. Both reviews ended with the same conclusion: “It is time for test equating to get a firm psychometric footing” (van der Linden, 1997, 2006c).

It is tempting to think of measurement error as “small epsilons to be added to test scores” and to believe that for well-designed tests the only loss involved in ignoring their existence are somewhat less precise equated scores. This chapter shows that this view is incorrect. Equating problems without measurement error are structurally different from problems with error; the score distributions for the former imply single-level modeling; those for the latter hierarchical modeling. Lord’s (1980) discussion of observed-score equating for the cases of infallible and fallible measures already revealed some of the differences: Without measurement error equating is automatically equitable and population independent, but with error these features are immediately gone. This chapter has added another difference: Without measurement error the same transformation suffices for any population of test takers, but with error the transformations become ability dependent and we need to look for different transformations for different ability levels.

The statistical consequence of ignoring such structural differences is not “somewhat less precise equated scores” but bias that, under realistic conditions, can become large. This consequence is not unique to equating; it has been well researched and documented in other areas, a prime example being regression with errors in the predictors, which have a long history of study as “errors-in-variables” problems in econometrics.

As the review of the local equating methods above suggests, the main change for equating to allow for measurement error is a shift from equating based on marginal distributions for an assumed population to the conditional distributions given a statistical estimate or a proxy for the ability measured by the tests. In principle, the formal techniques required for distribution estimation, smoothing, and the actual equating, as well as the possible designs for equating studies, remain the same. Thus, in principle, in order to deal with measurement error we do not have to reject a whole history of prolific equating research, only to redirect its application.