6 Potential Solutions to Practical Equating Issues

Alina A. von Davier[1]

Educational Testing Service

I work on problems in statistics that I can solve.
—Rupert Miller (Stanford, Department of Statistics) to Paul Holland, circa 1964

6.1. Introduction

Test equating methods are used to produce scores that are interchangeable across different test forms that are built to the same specifications (Holland, Chapter 2; Holland & Dorans, 2006; Kolen, Chapter 3). It is the most stringent form of score linking because it claims score *interchangeability*, not merely comparability, as do concordances and predictions (see Holland & Dorans and Holland, Chapter 2, for more details and definitions of types of score linking). Other types of score linking might use the same computations as test equating but do not result in scores that are interchangeable. A linking typically does not qualify as an equating when the test forms are not constructed to the same specifications or when the test forms measure different constructs. Test equating places several stringent requirements on the content and statistical properties of the test forms and on the samples of test-takers involved and is vulnerable to deviations from these requirements. These deviations might result in scores that are not interchangeable. In these circumstances, the intended test equating becomes a weaker form of test linking and the lack of interchangeability of scores can lead to unintended unfairness to some test-takers.

A good equating is like good cooking: It starts with good ingredients, the right tools, sound knowledge, and a bit of talent. Some "stumbling blocks to equating" (Cook, Chapter 5) appear when the assumptions

[1]The opinions expressed in this chapter are those of the author and not necessarily of Educational Testing Service.

required by an equating method are not fulfilled—for example, when the population invariance assumption fails (Dorans & Holland, 2000). Other stumbling blocks arise when the samples available for equating are too small and when large differences exist in the abilities of the groups that take the two test forms to be equated. In these situations, the equating issues are further exacerbated by poor test or anchor-test construction. In an attempt to address these stumbling blocks, researchers have measured the impact on equating of failures of assumptions (population invariance studies, studies on the quality of the anchor) and have developed new strategies to cope with design and data difficulties (equating with small samples, new approaches to anchor-test construction, and new equating models).

This chapter outlines some of this new research and discusses how it can improve test equating practice.

Before embarking on this investigation of the usefulness of new methodologies, we need to remember that, so far, no systematic theory of test equating has been outlined. Over the years, methods have been developed in response to the need to create comparable test scores in practical circumstances. In order to evaluate these methods, Dorans and Holland (2000), Holland and Dorans (2006), Kolen and Brennan (2004), and Lord (1980) have laid out a framework that defines a good equating procedure. This framework is based on the following five *requirements* on the test forms and on the equating functions: the same construct, equal reliability, symmetry, equity, and population invariance requirements. "This is not much of a theoretical underpinning for test equating," said Dorans and Holland (2000, p. 283). Moreover, many of these requirements are vague or arguable. In addition, in most situations, a failure of any of these requirements is hard to detect using the available data. The combination of the lack of a theory and difficulties in detecting bad equating results in practical settings create a challenging situation for a practitioner.

The research overviewed in this chapter is mostly focused on observed-score equating methods and investigates the following equating issues:

1. The population sensitivity of equating functions
2. Small samples equating
3. Addressing the differences in ability of the two groups of test-takers by matching on an anchor test and by constructing the anchor test in nontraditional ways
4. Addressing the stability of the equating results by implementing new equating models such as kernel equating (KE) and by applying the KE framework

The rest of the chapter is structured in six sections, with the first introducing the notation, the next four addressing the above-described issues, and, finally, providing the conclusions and discussion.

This overview of problems and solutions in equating does not directly address the conflicts that might arise between the demands of the testing industry and strong statistical and psychometric practice. To paraphrase the motto of the chapter: "I work on statistical problems that I can solve."

6.2. Observed-Score Equating Methods

In this section I introduce notation and lay out a framework for the discussion of equating. See also Kolen (Chapter 3) for a related discussion.

There are two test forms to be equated, X and Y, and a target population, T, on which this is to be done. The data are collected in such a way that the differences in the difficulty of the test forms and the differences in the ability of the test-takers that take the two forms are not confounded. There are two classes of data collection designs for equating: (a) designs that allow for common people (equivalent groups, single group, and counterbalanced designs) from a single target population of examinees T (see Livingston, 2004, for a slightly different view and definition of a target population) and (b) designs that allow for common items (the nonequivalent groups with an anchor-test design or NEAT design, also referred to as the common-item or anchor-test design) where the tests, X and Y, are given to two samples from two test administrations (populations), P and Q, respectively, and a set of common items, the "anchor test," is given to samples from both these populations). See also Figures 3.5 and 3.6 in the chapter by Kolen (Chapter 3, Section 3.5). The target population, T for the NEAT design, is assumed to be a *weighted average* of P and Q. P and Q are given weights that sum to 1. This is denoted by $T = wP + (1 - w)Q$.

Many observed-score equating methods are based on the *equipercentile equating function*. It is defined on the target population, T, as

$$e_{Y;T}(x) = G_T^{-1}(F_T(x)), \tag{6.1}$$

where $F_T(x)$ and $G_T(y)$ are the cumulative distribution functions (cdfs) of X and Y, respectively, on T. In order for this definition to make sense and to ensure that the inverse equating function also exists, it is also assumed that $F_T(x)$ and $G_T(y)$ have been made strictly increasing and continuous or "continuized."

Several important observed-score equating methods might be viewed as only differing in the way the continuization is achieved. The traditional equipercentile equating method (percentile rank method) uses linear

interpolation of the discrete distribution to make it piecewise linear and, therefore, continuous. The KE method uses Gaussian kernel smoothing to approximate the discrete histogram by a continuous density function.

Equipercentile equating leads to linear equating if one assumes that $F_T(x)$ and $G_T(y)$ are continuous and have the same shape while differing in mean and variance. The linear equating function, $\text{Lin}_{Y;T}(x)$, is defined by $\text{Lin}_{Y;T}(x) = \mu_{YT} + \sigma_{YT}((x - \mu_{XT})/\sigma_{XT})$, where μ_{XT}, and μ_{YT}, and σ_{XT}, and σ_{YT}, are the means and standard deviations of X and Y on T, respectively.

In von Davier et al. (2004b), it is shown that any equipercentile equating function can be decomposed into the corresponding linear equating function and a nonlinear part.

The next four sections describe several stumbling blocks to equating, some of the research conducted to address them, how the results of these research studies might improve equating practice, and identify research that still needs to be conducted.

6.3. Addressing the Fairness Issue: Population Invariance of Equating Functions

The practical equating concern addressed in this section is the lack of fairness towards subgroups of examinees that may occur when the assumption of population invariance of an equating function does not hold across subpopulations. I discuss this topic from several perspectives.

6.3.1. Definitions and Measures of Population Differences in Equating

One of the five requirements of score equating functions mentioned earlier is that equating should be population invariant; that is, the function computed should not be sensitive to the examinees whose data are used to compute it. Because strict population invariance is often impossible to achieve, Dorans and Holland (2000) introduced a measure of the degree to which an equating function is sensitive to the population on which it is computed. The measure, the root mean square difference (RMSD), compares linking functions computed on different subpopulations with the linking function computed for the whole population. The RMSD index was initially developed for the single group and equivalent groups designs. It was extended to other equating designs and methods in von Davier et al. (2004a).

Although the concept of invariance in equating and linking can be traced back to 1950 (Kolen, 2004b), in recent years there was a significant increase in this research. Most of the studies have focused on the detection

of population differences in equating and linking (Angoff & Cowell, 1986; Dorans & Holland, 2000; Harris & Kolen, 1986; Segall, 1995; von Davier et al., 2004a) and on the development of tools for making decisions (Dorans & Feigenbaum, 1994; Holland, Liu, & Thayer, 2005; Liu & Holland, 2006; Moses, 2006; von Davier & Manalo, 2006).

Dorans (2004e) introduced *score equity assessment* (SEA) to describe studies of test fairness that include differential prediction and differential item functioning (DIF). He provided an overview of the evolution of fairness assessment and placed the study of the population sensitivity of equating functions at the core of score equity. He recommended the routine investigation of subgroup dependence of the equating functions. I also believe that measures of population sensitivity of equating results should be routinely employed in operational work (similarly to the way that DIF analyses are now routine operational procedures). This is especially important when new tests or changes to the tests are introduced. The procedure could be automated and embedded in system software and might provide a flag if the population invariance assumption is violated at particular score points. However, establishing a flag requires a criterion. In the following subsections more details on establishing criteria are presented.

How could population invariance indexes help practitioners achieve better equating results? Such indexes are a first step in the process of ensuring fair equating results. The next subsection discusses how to judge the information provided by population invariance indexes.

6.3.2. Criteria for Detecting Subpopulation Differences in Equating Functions

There are at least three different questions one might ask about a particular measure of population sensitivity: (a) Does the amount of observed population sensitivity matter? (b) Is the amount of observed population sensitivity statistically significant or is it just noise? (c) What characteristics of the data, tests, and test-takers lead to population dependence?

To address question a, we might make use of the *difference that matters* (DTM), introduced by Dorans and Feigenbaum (1994). The DTM for a testing program depends on its reporting scale. For example, if the unit of a score scale is one point, then a difference between equating functions larger than a half-point on this scale means a change in the reported score, and this fact might establish the DTM for that particular program. All differences in equating results can be compared to the DTM to judge if they matter. However, the population invariance index, RMSD, introduced

by Dorans and Holland (2006), needs to be compared to a *standardized* DTM, which is the DTM divided by the same quantity as the denominator in the RMSD. Some of the recent studies that made use of the DTM criterion for detecting population sensitivity are: Dorans, Holland, Thayer, and Tateneni (2003), Liu, Cahn, and Dorans (2006), von Davier and Wilson (2005), and Yang (2004).

The studies that address question b focus on computing the accuracy of the population invariance indexes. Moses (2006) computed the standard errors (*SE*) for the RMSD index for the KE and showed how to compute the analytical formulas for the *SE* in the KE framework, using a standard large-sample approach. Other approaches compute the empirical *SE* of the RMSD for various equating functions using jackknife techniques (von Davier & Manalo, 2006).

Some studies (Holland et al., 2005; Liu & Holland, 2004) examined how population invariance indexes vary with differences between the tests and the subpopulations of test-takers. This allows us to define "a large value" of these indexes in terms of known factors that influence these indexes (question c).

How do these different criteria help practitioners achieve better equating results? All three types mentioned are valuable and are not mutually exclusive. Each provides information that can aid important decisions for ensuring a fair assessment. For example, the difference between the DTM and the *SE* is similar to the difference between clinical significance and statistical significance as used in medicine: One can have a statistically significant population dependence that will not matter to the test-takers or might have a DTM that is not statistically significant given the data at hand. One the other hand, comparing an RMSD index value to those typically found for parallel tests of given reliability can indicate when a observed RMSD value is typical of that type of testing program.

6.3.3. Implications of Population Sensitivity of Equating Functions

What should be done when the population invariance assumption is violated? This case can easily arise with concordances (see Dorans & Holland, 2000; Holland & Dorans, 2006; Dorans & Walker, Chapter 10; Pommerich, Chapter 11; Sawyer, Chapter 12). However, suppose that it occurs in an equating situation.

The psychometrician can consider examining potential violations of the equating requirements by applying the above described criteria. There are several areas that might be investigated: (a) *Test development*. Should population dependence be expected given the manner in which the tests are constructed? Do the tests measure the same construct? Are the tests

equally reliable? (b) *The characteristics of the population dependence*. Is this the first occurrence of a subgroup dependence of the equating function in this assessment? How much does the equating function depend on the subpopulations? At which scores does this dependence occur? Does the dependence matter to the test-takers? (c) *The statistical significance*. Is the observed population dependence statistically significant? Are the subgroups large enough that the equating functions for the subgroups are reliably different?

If this is a first-time occurrence and if no explanation can be found given the testing process, the psychometrician might decide to monitor past and future forms of this particular assessment. If this population dependence recurs or if it is too serious to be ignored, then more radical solutions might be considered. Linking functions between two tests can be computed and the scores on the tests can be linked using them, even when population invariance fails to hold to a sufficient degree. In this situation, however, it is appropriate to claim less for the linking between the two tests: The link might be appropriate for the target population as a whole but inappropriate for some identifiable subgroups. In particular, in order to be fair to different groups of examinees, it might be necessary to consider using *different* links between the tests for different subpopulations of examinees.

Holland and Dorans (2006) gave the following example. Suppose that there are two subgroups of test-takers, two tests to be linked, and one subgroup of test-takers has lower scores on *X* than the other subgroup but that the *reverse* holds for the other test, *Y*. They concluded that when a reversal holds, the lower scoring group is always disadvantaged by the use of the total-group linking function. When tests that are built to the same specifications are equated, the possibility of reversals is rare. For the forming of concordances, however, reversals are more likely and should be monitored for major subgroups.

Dorans (2004e) recommended using SEA and population dependence of equating functions "to distinguish between equating and weaker forms of linking" and said:

> Some have argued in the K-12 arena that scores from different tests are simply exchangeable. Despite cogent arguments to the contrary (see Feuer, Holland, Green, Bertenthal, & Hemphill, 1999), this belief lingers. […] Does it matter to a boy or a girl […] which test or version of a test they take? If the answer is yes […], then the presumption of exchangeability is not supported by the data. Inferences that depend on this presumption may be suspect. Some weaker form of linking is more appropriate, and separate concordances for males and females are more equitable than ignoring existing linking differences. (p. 65)

However, the use of *different* links (in situations where equating is actually expected) for different subpopulations of examinees is a controversial solution (see Petersen, Chapter 4). Is it fair to have two people taking the same test, performing similarly, and receiving different scores based on the subgroups to which they belong? This concern needs to be balanced with the unfairness that reversals can create.

6.3.4. Discussion and Future Research Directions

The suggestions for the above-described strategies are not only statistical but also involve program policy. A particular program will need to weigh the consequences of any decision for the test-takers and test users. It is better to avoid such situations by careful planning of test development and equating designs that lead to fair equating results. For more details, see Dorans (2004e), Petersen (Chapter 4, Section 4.2), and Kolen (2003).

To make the study of population sensitivity more practical, I recommend continuing to search for indexes of population dependence that do not require the various subgroups equatings. When there are multiple subpopulations, examination of the subgroups equatings with the existing indexes is time- and labor-intensive. Dorans and Holland (2000) provided an example of such a simplifying method. See Holland et al. (2005) for an illustration of how this simplified method can reduce computations without losing sensitivity to population differences in equating.

6.4. Addressing the Small-Samples Issue: Synthetic Linking Functions

The equating of test scores is subject to sample characteristics. If the sample is large, the equating relationship in the sample might represent accurately the equating relationship in the population. The smaller the sample, the more likely that the equating function computed for that particular sample will differ from the equating function in the population. Both sampling error and bias can influence the quality of the equating. Hence, the impact of small sample size on equating is compounded when the samples are not representative.

The practical equating issue addressed here is what to do when the samples are small.

The research in this area has focused on three topics: the use of presmoothing of the discrete data prior to equating (Livingston, 1993; Skaggs, 2004), the use of the identity function instead of equating (Harris & Crouse, 1993; Skaggs, 2004), and the use of a weighted average of the

identity and a linear equating function without presmoothing (Kim, von Davier, & Haberman, 2006).

Livingston (1993) examined the effectiveness of log-linear presmoothing (Holland & Thayer, 1987, 2000) with small samples in an equivalent groups design with an anchor test. He found that the benefits of presmoothing were greatest when the sample was small, but that the number of moments in the observed distribution that should be preserved in the smoothed distribution might depend on the sample characteristics.

Skaggs (2004) studied equating of the passing score using samples ranging from 25 to 200 in an equivalent groups design with no anchor. He observed that the standard errors of equating became smaller as the sample size increased, but that the equating bias did not change much as a function of sample size. For samples as small as 25, no equating is likely to do less harm to examinees than some form of equating, but for samples in the 50–75 range, some form of equating was preferable to no equating. Generally, using log-linear models that fit the first two or three moments of the observed distribution produced smaller standard errors than did the unsmoothed equating, as Livingston (1993) found.

Kim et al. (2006) focused on the NEAT design, which is relatively uncommon in the literature on small-samples equating. In the NEAT design, the anchor test is supposed to adjust for the differences in ability in the two groups. However, in small samples, this adjustment might not be accurate. They introduced a compromise between the identity function (no equating) and an estimated equating function computed on the small sample. The *synthetic linking* function is defined as the weighted average of an estimated equating function and the *identity function* ($ID(x) = x$) or no-equating.

$$\text{syn}_Y(x) = we_y(x) + (1 - w)ID(x), \tag{6.2}$$

where w is a weight between 0 and 1. They showed that under an appropriate choice of the weight w, the synthetic function meets the symmetry requirement of an equating or linking function mentioned earlier.

The identity function might be a good choice when test specifications are well defined and the test forms are close to being parallel (see also Lawrence & Dorans, 1990; Skaggs, 2004), even when the equating samples are neither representative nor large enough. The mean of the equating results from the synthetic equating function is the weighted average of the mean of the identity and of the estimated equating function. This will reduce the bias in the identity equating function. At the same time, the new linking function will always contain less noise than the estimated equating function:

$$\begin{aligned}\mathrm{Var}(\mathrm{syn}_y(x)) &= w^2\mathrm{Var}(e_y(x)) + (1-w)^2\,\mathrm{Var}(ID(x)) + 2w(1-w)\mathrm{Cov}(e_y(x), ID(x)) \\ &= w^2\mathrm{Var}(e_y(x)).\end{aligned} \quad (6.3)$$

One limitation of this approach is that the two tests should have the same length for the identity function to make sense. In addition, if the test forms are not nearly parallel, the bias introduced by the identity function might be too large.

In Kim et al. (2006), two types of real test data were used that differed in the reliability of the tests and the anchor. For the estimated linking function they used chained linear equating. Chained linear equating was also used as the criterion based on about 10,000 cases. Smaller samples were randomly drawn from the two (nonequivalent) groups. When sample sizes were small (less than 25), the synthetic function did a better job than the estimated chained linear function. For samples as small as 10 or 25, the synthetic equating function was preferable to either not equating or using the chain linear method alone.

If historical data exist, *w*, in the synthetic function, can be viewed from the perspective of variance components. The weight on the identity should increase as sample variance increases and as year-to-year test variability decreases. In the absence of historical data, the weight can be a function of the difference in the abilities in the two groups, the correlation of the tests and the anchor, the reliabilities and the difference in difficulty of the two forms, and the sample size (see also Kolen & Brennan, 2004, p. 289).

Equating with small samples requires the user to depend on *assumptions* because there is less guidance from the data. The synthetic equating function illustrates how to use assumptions to achieve more stable results. When the test forms are constructed to be nearly parallel, the bias introduced by an identity equating is not expected to be large. The synthetic function allows more flexibility than simply not equating when the samples are small. In a similar way, presmoothing with log-linear models makes assumptions to compensate for the lack of data.

However, assumptions can be wrong, so it is important to know their consequences. Would using empirical data to construct a replacement for the identity function be better? Would the equating results be more stable if a log-linear model is used that fits only the mean of the sparse observed distributions? Perhaps collateral information about the test items could be used to augment the total-test scores, as Mislevy, Sheehan, and Wingersky (1993) proposed?

More research is needed before we can conclude whether the use of the synthetic function relying on the identity function makes matters better or worse. Follow-up studies of the work of Kim et al. (2006) can investigate the synthetic function under various circumstances, including those in

which the identity function might introduce a significant bias. It is natural to suggest comparing versions of the synthetic function to the use of log-linear models with few parameters in terms of bias and variability.

6.5. Addressing Differences in Ability in the Two Populations of the NEAT Design

The practical equating issue here is to equate scores for test forms that are taken by groups that exhibit *large* differences in ability (see Cook, Chapter 5). In the NEAT design, the anchor test, taken by the two groups of test-takers, is used to adjust for the differences in ability in the two groups. Previous research in this area has focused on three topics: the use of the anchor test to create similar or matched groups (Kolen, 1990; Lawrence & Dorans, 1990; Livingston, Dorans, & Wright, 1990), the use of other variables to create matched groups (Liou, Cheng, & Li, 2001; Wright & Dorans, 1993), and, recently, the creation of anchor tests that maximizes their correlation with the tests to be equated (Sinharay & Holland, 2006).

When there are only small differences between the two samples of examinees used in the NEAT design, all linear equating methods tend to give similar results, as do all nonlinear equating methods (see Kolen, 1990; von Davier, 2003; von Davier et al., 2004a). To the extent that a NEAT design is almost an equivalent groups design with anchor test, the need for the anchor test is minimized. This is the main argument behind the *matching-on-the-anchor procedure*. When matching on the anchor is carried out, the distributions of the anchor in the two matched groups will be the same (Kolen, 1990; Lawrence & Dorans, 1990; Livingston et al.). If the distributions of the anchor in the two groups are the same, all comparable (equipercentile versus linear) observed-score equating methods will give the same result (von Davier, 2003). However, Cook and Petersen (1987) and Livingston et al. (1995) noted that although all the equating functions agree, their agreement might correspond to an incorrect equating function due to bias.

In order for the matching-on-the-anchor procedure to work, the anchor has to behave in the two groups similarly to the two tests, X and Y (see also Cook, Chapter 5). Other research focused on matching groups on variables other than the anchor (Wright & Dorans, 1993). Matching both on the anchor and on other variables seems to be promising.

When the two samples are very different in performance, the use of the anchor test becomes critical; it is the *only* means of separating the *differences between the abilities* of the two groups of examinees from the *differences between the two tests* that are being equated (see Holland & Dorans, 2006). The most important properties of the anchor test are its

integrity and stability over time and its correlation with the scores on the two tests being equated (Holland & Dorans). It is important for the correlation to be as high as possible. Because of their part-whole relationship with the other tests, internal anchors have high correlations with the total tests.

Petersen et al. (1989, p. 246) and von Davier et al. (2004a, p. 33) indicated that the higher the correlation between scores of an anchor test and scores on the tests to be equated, the better the anchor test is for equating. The importance for equating of this correlation raises the question: Does the usual advice of making the anchor test a "mini-version" of *X* and *Y* actually increase this correlation? The requirement that the anchor test should be representative of the content of the total test has been shown to be an important requirement by Klein and Jarjoura (1985). If the difficulties of the items in the full tests are spread over a range of values, does that mean that the difficulties of the anchor-test items should be spread over the same range? The results reported in Sinharay and Holland (2006) suggested that this might not be true. These authors examined whether the spread of item difficulties should be the same as that of *X* and *Y*. They show that an anchor test with a narrow spread of item difficulty might perform as well (in terms of accuracy and precision in equating) as one consisting of items with a wider spread of difficulties. In a series of simulation studies, they explored the relations between scores on a total test and an external anchor test for different types of anchor test, based on generated data from one- and two-dimensional logistic item-response models. Their main finding is that an anchor test with a narrow spread of item difficulties located near the mean of the difficulties of the total tests has the highest correlation with the total tests for almost all of the situations considered.

How can this research improve test equating? When there are large differences in ability in the two populations in the NEAT design, equating can be a challenge.

Matching on the anchor and/or on other variables that correlate with the tests are procedures that require more research and the results need to be interpreted carefully. As mentioned earlier, all of the equating functions might agree to an incorrect (biased) equating function. If a demographic variable is used, then one might ask if the result is an equating and if the test scores are interchangeable. What role would the subpopulation dependence of the equating function play in choosing a person characteristic as a matching variable? More research is necessary to shed light on these issues.

One interpretation of Sinharay and Holland (2006) is that although it is important that anchor tests match the content and overall difficulty of the total tests, it is less important to match the spread of the item difficulties of

the total tests. If further research bears out their preliminary findings, then their work suggests that test developers need not attempt to make the distribution of item difficulties look like a miniversion of the distributions for the total tests and can focus on matching test content and overall difficulty.

6.6. Addressing the Stability of Equating Results: Kernel Equating and Applications

In the practice of equipercentile equating, psychometricians have typically used the *percentile rank method* (that uses linear interpolation to make the cdfs in Equation 6.1 continuous) for equating test forms with score distributions that differ in shape. One of the consequences of this method is that the linearly interpolated cdfs and the equating function have kinks; that is, the functions are not smooth (see Kolen & Brennan, 2004, Figures 2.4, 2.5, and 2.10). Moreover, if there are no examinees at a particular score, the percentile rank method is not well defined. In order to address these issues, past research focused on procedures for smoothing the data prior to equating (presmoothing), procedures for smoothing the equating function (postsmoothing), alternative procedures for continuizing the cdfs, and new equating functions.

In my opinion, the issue of stability and quality of equating results is best addressed by providing the following: (a) a coherent and formal equating process; (b) better methods of continuizing the discrete distributions, F and G, in order to be able to compute the equating function from Equation 1; (c) useful measures of statistical accuracy; and (d) equating models that are appropriate for particular test designs. In the next subsection I will briefly describe the kernel equating method and indicate how it accomplishes the four above mentioned aspects.

6.6.1. The Gaussian Kernel Method

Holland and Thayer (1989) and von Davier et al. (2004b) viewed all observed-score test equating as having five steps or parts, each of which involves distinct ideas: (1) presmoothing of the score distributions; (2) estimation of the score probabilities on the target population; (3) continuization of the presmoothed discrete score distributions; (4) computing the equating function; and (5) computing the standard error of equating and related accuracy measures. They applied this framework to describe kernel equating (KE); see von Davier et al. (2004b) for details and for a detailed description of KE.

The main advantage of the KE framework is that it brings together these steps into an organized whole rather than treating them as disparate problems. KE exploits presmoothing by fitting log-linear models to score data, and it incorporates the presmoothing into Step 5 of the framework, where KE provides new tools for comparing two or more equating functions and to rationally choose between them.

Kernel equating is an equipercentile equating procedure in which the discrete score distributions are made continuous using Gaussian kernel smoothing rather than linear interpolation. By varying the bandwidth values in Step 4, KE can approximate the traditional equipercentile and the linear equating methods. The bandwidths are positive constants that manipulate the weight placed on the Gaussian kernel and that can be chosen to achieve various purposes. When "optimal" bandwidths are chosen, KE will closely approximate the traditional equipercentile equating method. When the bandwidths are large (10 times the standard deviation of the scores or larger), the continuized distributions will be nearly Gaussian and the KE functions are effectively linear. Thus, linear equating can be regarded as special case of equipercentile equating in the KE framework.

In the KE framework, von Davier et al. (2004b) introduced the standard error of the difference (the SEED) between two equating functions. The SEED has several practical uses such as rationalizing the linear/nonlinear decision, implementing a new approach to the counterbalanced design, comparing chained and poststratification equating methods in the NEAT design, or aiding the comparison among other observed-score equating methods (von Davier & Kong, 2005). The various uses of the SEED do not require KE, but the SEED is a natural part of the KE framework and von Davier et al. (2004b) showed how to apply it for these purposes.

Several research studies (Han, Li, & Hambleton, 2005; Mao, von Davier, & Rupp, 2006; von Davier et al., 2006) focused on evaluations of KE and on comparisons of KE with other observed-score and true-score equating methods. Among other things, these studies indicate that KE can closely approximate traditional equating methods well. These studies used the newly developed KE-Software (Educational Testing Service, 2006).

6.6.2. Applications of the KE Framework

Recent studies have taken advantage of the formal and coherent formulation of the KE process and have focused on the application of KE to particular equating issues.

Moses, Yang, and Wilson (2005) explored the use of KE for integrating and extending two procedures (Hanson, 1996; Lawrence & Dorans, 1990)

proposed for assessing the statistical equivalence of two test forms in which the same items have been scrambled into different orders.

Other applications of the KE framework are Moses (2006), which computed the standard error of population invariance indexes, and Moses and Holland (in press), which extended the KE computations to situations in which the data are not presmoothed.

The KE framework is also used to c*onstruc*t hybrid equating function that combine a linear equating function from one source with an equipercentile function from another. An example is a nonlinear generalization of the Levine linear observed-score equating function. The Levine linear method does not yet have a curvilinear analogue, and there is no version of KE that approximates the Levine function. Nevertheless, the Levine linear method is often computed in practical applications for comparison purposes. Under some circumstances, it is more accurate than other linear methods (see Petersen, Marco, & Stewart, 1982).

von Davier, Fournier-Zajac, and Holland (in press) used the KE framework to construct a hybrid equating based on the Levine linear method. The new function preserves the nonlinear characteristics from the KE poststratification and the linear form from the classical Levine observed-score equating.

With the five steps of the KE framework identified, other research has focused on replacing the original proposals from von Davier et al. (2004b) with alternatives to create new equating processes. One of these proposes alternative continuization methods: Wang (2004) continuized the discrete probability distribution by using the polynomial log-linear function (from the presmoothing step), divided by the area under it, in order to ensure that it is a probability distribution function. The method is called the continuized log-linear (CLL) method. As a potential alternative to the Gaussian kernel, Holland (personal communication, July 26, 2005) discussed the possibility of using a logistic kernel. One of the advantages of the logistic kernel is that the analytical form of the derivatives required for computing the SEE and the SEED is very simple. At the same time, given the modular character of KE-Software (Educational Testing Service, 2006), it would be very easy to implement it in parallel with the Gaussian kernel.

6.6.3. Discussion and Future Research Directions

How do these new equating models address the issues of stability of the equating results? The nonlinear Levine function is a new equating method that might allow the known benefits of the Levine linear function (Petersen et al., 1982) to apply to cases where nonlinear equating is required. There

are situations in which the tests and the anchor are very carefully constructed, but the two test score distributions differ in shape (see von Davier, Holland, et al., 2006). In such a case, a nonlinear version of the Levine function is desirable.

One reason for seeking alternatives to continuization with a Gaussian kernel is that the use of the Gaussian kernel leads to lower values of the higher order cumulants of the continuous distribution than those of the original discrete distribution (Holland & Thayer, 1989; von Davier et al., 2004b). So far, this reduction in the cumulants has not been shown to have any practical implication. The Wang (2004) proposal of CLL might provide a possible alternative to kernel smoothing because it directly computes the cdfs from the fitted loglinear model.

The new accuracy method introduced in the KE framework, the SEED, has direct practical uses: It can aid the decision between linear and nonlinear equating functions, between equating functions that are based on different assumptions, such as the poststratification and chained equating (see Kolen & Brennan, 2004; von Davier et al., 2004b), or between the linear methods used in the NEAT design. The SEED is a statistical tool that has the potential of being extended to other applications—possibly as a decision aid between log-linear models.

The KE method has been around for almost 20 years, and despite the obvious theoretical and practical advantages, it is still not part of the operational practice. Many practitioners are intimidated by the theoretical description of KE. Actually, many practitioners do not explicitly use linear interpolation, but a conversion table, with averaged values between score points. The KE method, although a differentiable function that differs from the linear interpolation, agrees closely to the equipercentile function, which uses linear interpolation at almost all score points when an appropriate bandwidth is selected. This is fortunate and unfortunate at the same time. It is fortunate to have the equating functions agreeing, but it is unfortunate because it gives practitioners no reason to change. Researchers and policy makers need better arguments to convince practitioners, such as emphasizing the availability of KE accuracy and diagnostic measures, the modularity of the KE framework that translates into a modular software package, and the easy-to-use interface of KE-Software (Educational Testing Service, 2006). Moreover, the KE framework has the potential of introducing automatic procedures with incorporated automatic decision steps to reduce the routine work of the psychometricians and data analysts.

In my opinion, studies of alternative continuization methods and of hybrid functions are of a more theoretical than practical interest in the near future. From the practical point of view, I believe that research focused on decision aids and automatic equating procedures is necessary. Developing or refining indexes, such as the SEED for aiding in the process of

comparing equating function, indexes for deciding among log-linear models in the presmoothing procedures, or attempts to improve the fit of the loglinear models (and therefore to improve the stability of equating results) in regions of the score range that matter to a particular program are of importance in equating practice. In addition, we should focus on expanding the research on the KE method to scale drift and to tests with complicated distributional shapes. Additionally, researchers should focus on finding more efficient ways to teach and explain the KE method and to engage more practitioners in evaluating procedures and approaches.

6.7. Discussion

This chapter summarizes my selection of the current research directions in equating that show some potential in addressing issues encountered in the practice of equating.

For the sake of coherence, I decided to focus on observed-score equating only. Equating that uses item information and is based on item response theory models has its own challenges, including those mentioned in Cook and Petersen (1987), von Davier and Wilson (2006), Hambleton, Swaminathan, and Rogers (1991), Kolen and Brennan (2004), Lord (1980), and Petersen et al. (1983, 1989).

Here I discussed several equating topics: the population dependence of equating functions, the equating in small samples, the adjustments needed when the groups of test-takers differ in ability, and the stability of equating functions provided by the KE method.

The main point that this chapter makes is that there is a continuous effort to address scientifically the practical issues of equating and that research does not take place in an ivory tower, but is responsive and related to practical problems.

Another point made here is that the equating process always involves policy decisions in addition to the statistical ones and that the responsibility for fair assessments needs to be shared between the leaders of the program and the psychometricians who advise them.

Currently, when more and more standardized testing is used for assessing competencies in different domains nationally and internationally, we are also discovering more challenges to ensuring that the process and the results are fair and accurate. In turn, these challenges and these new social implications open the door to more research in support of fair assessments, both in improving the test construction process and in advancing the statistical methods involved.

***Acknowledgments*.** The author thanks Neil Dorans, Mary Pommerich, Paul Holland, Dan Eignor, and Tim Moses for their comments on the previous versions of the chapter and for interesting discussions on different topics addressed in this presentation.