Measurement, 8: 21–26, 2010
Copyright © Taylor & Francis Group, LLC
ISSN: 1536-6367 print / 1536-6359 online
DOI: 10.1080/15366361003684711

Psychology Press
Taylor & Francis Group

On Bias in Linear Observed-Score Equating

Wim J. van der Linden
CTB/McGraw-Hill

The traditional way of equating the scores on a new test form X to those on an old form Y is equipercentile equating for a population of examinees. Because the population is likely to change between the two administrations, a popular approach is to equate for a *synthetic population*. The authors of the articles in this issue of the journal try to avoid the arbitrariness in the definition of a synthetic population by equating X to Y for the population G1 that takes the new form. Let $F_X(x)$ be the distribution function of the scores on X for this population of examinees and $F_Y(y)$, the corresponding function of the scores on Y. Ignoring possible irregularities in the distributions, the equipercentile transformation can be written as

$$I_Y(x) = F_Y^{-1}(F_X(x)). \tag{1}$$

Linear equating is an approximation to this transformation based on the first two moments of the two observed-score distributions. One reason for its popularity is that the restriction to these moments reduces the need for large samples and, thus, avoids the typical issues of instability and choice of smoothing method involved in the use of (1). Continuing the authors' notation, let $\mu(X)$ and $\mu(Y)$ be the population means for the two forms and $\sigma(X)$ and $\sigma(Y)$ the two standard deviations. The version of (1) that follows from the restriction is

$$I_Y(x) = \frac{\sigma_2(Y)}{\sigma_1(X)}[x - \mu_1(X)] + \mu_2(Y). \tag{2}$$

I have been happy to notice the authors' attention to the topic of bias in linear equating. The equating literature has been dominated by an interest in the standard error of equating, but bias is the primary criterion for evaluating the success of an equating. After all, equating is an attempt to remove the bias in the score on the new test form as an estimate of the score on the old form due to scale differences between them. A focus only on the standard error of equating prevents one from noticing any remaining bias in the equated scores, or even possible new bias added to them in the equating process. The purpose of this commentary is to discuss a little further the issue of bias in linear equating.

Two other equating issues for which the authors justly ask attention in the discussion sections of their articles are the necessity to specify a synthetic population and the negligence of the

Correspondence should be addressed to Wim J. van der Linden, CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940. E-mail: wim_vanderlinden@ctb.com

impact of measurement error on the traditional regression assumptions in linear equating. The discussion of these issues is postponed until the end of this commentary, but I already notice that both are directly related to the presence of bias in observed-score equating.

THE AUTHORS' DEFINITION OF EQUATING BIAS

The authors' definition of equating bias is motivated by the database used in their empirical study. Forms X and Y as well as the examinees in the study were sampled from an initial pool of test items and examinees from the Multistate Bar Examination. In the reference condition, the two groups of examinees for the two forms were randomly sampled for equivalence. As the actually observed scores of the two groups on both test forms were available, the means and standard deviations in the linear transformation in (2) for this condition were known. The authors used this transformation as their baseline in evaluating the equating results for all the NEAT designs in their study.

More specifically, let x_i be the score on form X and $f\,(x_\mathrm{i})$ its observed frequency in the equating study for the group that took X (the authors' choice of synthetic population). In addition, let $\hat{I}_Y(x)$ be the transformation for an equating with a NEAT design that is to be evaluated. Bias was defined as the average difference between $\hat{I}_Y(x)$ and the criterion $I_Y(x)$. In formula

$$\text{Bias} = \frac{\sum_i f(x_i)\left[\hat{I}_Y(x_i) - I_Y(x_i)\right]}{\sum_i f(x_i)} \tag{3}$$

(Suh, Mroch, Kane, & Ripkey, 2009, p. 155, Eq. 2)

A FEW TECHNICAL COMMENTS

Generally, bias in equating should be conceived of as a difference between two functions: the function of x that is actually used to perform the equating and the one that should have been used. So, bias itself is also a function of x. The averaging over the score distribution on X in (3) may obscure the nature of this bias function in two different ways: First, the weighing confounds the size of the bias with the numbers of students incurring it. Second, because bias can be positive (equated scores too high) or negative (equated scores too low), the authors' definition of bias allows the two types to compensate. In principle, the expression in (3) can be close to zero, whereas the underlying bias function may display bias at nearly every score along the range of scores on X.

Secondly, as already indicated, linear equating derives its popularity from the statistical stability of the equating that is obtained by its restriction to the first two moments of the distributions on X and Y. However, the ignoring of the higher-order moments of the distributions comes generally at a price in the form of bias. In fact, the price is not different from that paid for the stability of an equating that results from the usual presmoothing of the two distributions or postsmoothing of the equating transformation. All are examples of the well-known bias-accuracy trade-off that can be met throughout statistics. The expression in (3) misses this source of bias.

A measure that would have captured the source is the difference between the actual equating function $\hat{I}_Y(x)$ and the equipercentile function in (1), that is,

$$I_Y(x) - F_Y^{-1}(F_X(x)). \tag{4}$$

The authors used linear correlation between the scores on X and Y to check the assumption of distributions of X and Y differing only in location and scale and found correlations between .77 and .79 across all conditions in their study (Suh, Mroch, Kane, & Ripkey, 2009, p. 155). For perfectly reliable scores, the correlation needs to be equal to 1.0 to satisfy the assumption. My guess is that after correction for attenuation their correlations would still suggest imperfect linearity. The larger the deviation from perfection, the larger the bias in the equated scores due to the assumed linear shape of the equating function.

The next comment is on the use of an equating function for a randomly equivalent group as a criterion for evaluating equating in a NEAT design. Unlike randomly equivalent–group designs, NEAT designs, not only have different items in the two test forms, but the two groups that take them also differ in their ability distributions. Hence, equating transformations for this type of design should adjust both for the differences between the items and the abilities. It is difficult to see why such transformations should be evaluated against one that adjusts only for the differences between the items. In fact, the better a NEAT transformation adjusts for the differences between the abilities, the worse its evaluation against the transformation for a randomly equivalent–group design. It does not seem right to penalize an equating transformation for a job it is supposed to do.

The question remains, what other criterion could the authors have used to evaluate bias in test-score equating.

A MORE FUNDAMENTAL COMMENT

A suggestion that immediately comes to mind is to use the actual observed score on the old test form, y_i, as a criterion, that is, $y_i - \hat{I}_Y(x_i)$ for each examinee. After all, the authors had these scores in their database.

The reason why this suggestion has not been followed is obvious. Figure 1 shows the distribution of the actual observed scores on an old test form Y given an observed score $X = 10$, 20, 30, and 40 on the new form. (The forms in this figure are not those used by the authors but are from another large-scale testing program, with the test length randomly reduced to the same length of 50 items.) Notice that although in each histogram the scores are for examinees with the same score on X, their scores on Y differ widely. This type of figure is seldom displayed in the equating literature because it reveals an embarrassing fact: The goal of observed-score equating is to make the scores of examinees on a new test form indistinguishable from the scores on an old form. But how could the goal ever be realized when examinees with the same score on the new form differ as widely in their actual scores on the old form as they do in Figure 1?

The reason for these wide ranges of actual scores is, of course, measurement error. In spite of its frequent references to test-score reliability (the authors have several sections in their articles where reliability is one of the key concepts), the linear equating literature typically ignores the simple fact that the observed scores of individual examinees have an error component. Instead,

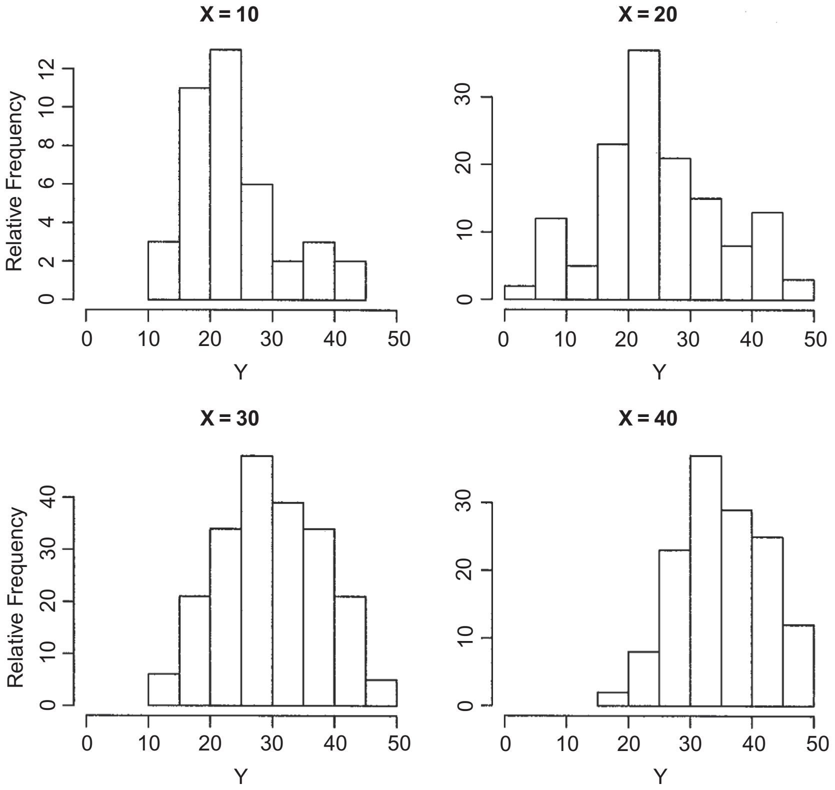

FIGURE 1 Conditional distributions of the observed scores on old form Y given the scores on new form $X = 10, 20, 30$, and 40.

these scores are pooled into a *population distribution* for each of the two groups in the equating study, whereupon the two distributions are redefined as distributions for a synthetic population, from which a single equating transformation for all examinees is derived. This pooling and synthesizing of observed scores for individual examinees, along with the derivation of a single transformation, may be the most serious source of bias in traditional score equating.

The problem has been analyzed extensively by this commentator, for instance, in van der Linden (2000, 2006). A more recent review of the problem from the perspective of the chapter on equating in Lord's (1980) well-known monograph is given in van der Linden (2010).

One way of formulating the problem is to note that traditional equating does not treat the observed scores of an individual examinee as a realization of a random variable—a basic assumption that underlies any model in test theory. Likewise, all of test theory assumes that the distribution of this variable depends on the ability of the examinee that is tested (as well as the properties of the test items). If traditional observed-score equating had been developed from these two fundamental assumptions, it would have treated score equating as the problem of finding the transformation that makes the random observed score of an examinee at a given ability level on the new test form indistinguishable from the random score on the old form.

This is precisely what Lord (1980, chap. 13) asked for when he introduced the criterion of equity of equating.

A moment's reflection shows that the goal could be realized if we knew the distributions of the observed scores on the two forms at the examinees' ability levels and applied the transformation in (1) at the level of each examinee: Let θ denote the ability of a given examinee. The unknown distributions of the observed scores on the two test forms by the examinee are the distributions of *X* and *Y* given θ. The two distributions are made indistinguishable by the transformation

$$\varphi(x;\theta) = F_{Y|\theta}^{-1}(F_{X|\theta}x)). \tag{5}$$

The fact that this is "equipercentile equating at the level of the score distributions given θ" should not come as a surprise. What is known as the equipercentile transformation in traditional equating is just an instance of the Q-Q transformation used in statistics used as a general tool to equate any two distributions (e.g., Wilk & Gnanadesikan, 1968).

The analysis reveals an inherent source of bias in any traditional form of observed-score equating: Instead of using a different appropriate transformation for each ability level, traditional equating uses a common transformation derived from some synthetic population. This common transformation necessarily compromises between the different transformations required for each ability level and, in doing so, biases each equated score. The commentator's proposal of local equating as an alternative to traditional equating (e.g., van der Linden 2010) minimizes the bias by approximating (5) as closely as possible for each individual examinee given all information available in the equating study. If the two forms fit a response model, an excellent approximation is to substitute a statistical estimate of θ into (5) for each examinee. The result is a third type of IRT observed-score equating, which is much less biased than the two earlier types suggested in Lord (1980, chap. 13). In two recent studies of local equating for the NEAT design, we explored the use of the anchor score as a proxy for the unknown ability (van der Linden & Wiberg, 2010; Wiberg & van der Linden, submitted). These studies show considerable bias for the traditional equating methods for this type of design but generally less bias for local equating, especially with increasing length of the anchor test. A more general review of the first attempts at local equating is given in van der Linden (2010).

FINAL COMMENTS

It is clear that the authors are unhappy with the need to specify weights for a synthetic population in the linear equating methods for which they reserve the collective name of parameter substitution methods (e.g., Kane, Mroch, Suh, & Ripkey, 2009, p. 129; Mroch, Suh, Kane, & Ripkey, 2009, pp. 174–178). Without much motivation, their choice is full weight for the population that takes the new form and no weight for the old population. However, the above analysis shows that the heavy reliance on population distributions in traditional equating, not only entails the need of such arbitrary choices, but is at the very heart of a serious bias problem. Because its equating transformations are derived for a population, the equated score of each individual examinee is compromised by those of all other examinees in the assumed population. Bias due to this compromise is much more embarrassing than the arbitrary choice of weights for old and new populations. But it may be reduced considerably by equating as closely as possible at the level of the individual abilities of the examinees.

The authors are entirely correct in their reference to the literature on errors in regression variables as relevant to the improvement of the current methods of observed-score equating. Basically, this literature shows that ignoring measurement error in the variables on which we regress leads to bias in the estimated regression function.

The authors use the literature to motivate their choice between equating methods based on the regression of X on the anchor score V and the other way around (Mroch, Suh, Kane, & Ripkey, 2009, Appendix A). Their goal seems to get a regression equation as close as possible to the true-score relationship between these two scores. However, the goal of observed-score equating is not to equate true scores. The attempt should be, therefore, not to get rid of measurement error, but, just as in this error-in-variable literature, to acknowledge its existence and account for it when equating observed scores.

REFERENCES

Kane, M. T., Mroch, A. A., Suh, Y., & Ripkey, D. R. (2009). Linear equating for the NAET design: Parameter substitution models and chained linear relationship models. *Measurement: Interdisciplinary Research and Perspectives*, *7,* 125–146.

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

Mroch, A. A., Suh, Y., Kane, M. T., & Ripkey, D. R. (2009). An evaluation of five linear equating methods for the NAET design. *Measurement: Interdisciplinary Research and perspectives*, *7*, 174–193.

Suh, Y., Mroch, A. A., Kane, M. T., & Ripkey, D. R. (2009). An empirical comparison of five linear equating methods for the NAET design. *Measurement: Interdisciplinary Research and Perspectives*, *7*, 147–173.

van der Linden, W. J. (2000). A test-theoretic approach to observed-scored equating. *Psychometrika*, *65*, 437–456.

van der Linden, W. J. (2006). Equating error observed-score equating. *Applied Psychological Measurement*, *30*, 355–378.

van der Linden, W. J. (2010). Local observed-score equating. (in press). In A. A. von Davier (Ed.), *Statistical models for equating, scaling, and linking*. New York: Springer.

van der Linden, W. J., & Wiberg, M. (2010). Local observed-score equating with anchor-test designs. *Applied Psychological Measurement*, *34*.

Wiberg, M., & van der Linden, W. J. *Local linear observed-score equating*. Manascript submitted for publication.

Wilk, M. B., & Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika, 55*, 1–17.