Some Practical Issues in Equating

Robert L. Brennan and Michael J. Kolen
American College Testing Program

The practice of equating frequently involves not only the choice of a statistical equating procedure but also consideration of practical issues that bear upon the use and/or interpretation of equating results. In this paper, major emphasis is given to issues involved in identifying, quantifying, and (to the extent possible) eliminating various sources of error in equating. Other topics considered include content specifications and equating, equating in the context of cutting scores, reequating, and the effects of a security breach on equating. To simplify discussion, some issues are treated from the linear equating perspective in Kolen and Brennan (1987).

The previous paper in this issue (Kolen & Brennan, 1987) provided a relatively detailed consideration of statistical issues associated with the Tucker and Levine linear equating methods for the common-item nonequivalent-populations design. In actual practice, equating necessitates not only knowledge of such statistical models but also consideration of many other issues. Frequently these issues have important practical consequences for the use and/or interpretation of equating results, no matter what statistical procedure is employed. Some of these practical issues are the focus of this paper.

Major emphasis is given below to considerations relevant to identifying, quantifying, and (to the extent possible) eliminating various sources of error in equating. Other topics considered include content specifications and equating, equating in the context of cutting scores, reequating, and the effects of a security breach on equating. To simplify the discussion, some issues are treated from the linear equating perspective in Kolen and Brennan (1987), but none of these issues is inherently linked to any particular statistical equating procedure or approach. Obviously, not all of these issues are necessarily involved in every equating context, but they arise often enough to merit more consideration than they have heretofore received in the literature.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 11, No. 3, September 1987, pp. 279–290

0146-6216/87/030279-12$1.85

Equating Errors

Any statistical or measurement procedure necessarily involves error in generalizing the obtained results to some idealized situation. Classical statistics usually concentrates on generalization from a sample of persons to a population of persons. By contrast, measurement usually concentrates on generalization from a particular test or measurement procedure to a potential universe of such tests or measurement procedures. In principle, although consideration of equating error should involve both types of generalization, "statistical generalization" has received much more emphasis in the literature. The following discussion is relevant to both types of generalization.

Provided below is an overview of currently available procedures for estimating standard errors

of equating—procedures that take into account only random sampling of examinees. Attention then turns to other types of errors in equating, including systematic bias, error that accumulates over multiple equatings, and errors associated with generalizing from equated scores obtained on one occasion to equated scores that might be obtained on similar occasions. Finally, consideration is given to a set of issues relevant to equating errors including, for example, methods for examining equating stability, rounding error, and conditions conducive to satisfactory equating.

Standard Errors of Equating for Random Sampling of Examinees

Using the delta method (see Kendall & Stuart, 1977, pp. 246–247), Lord (1950) presented standard errors of linear equating for a variety of data collection designs. (Most of Lord's results are also reported by Angoff, 1971.) Although Lord's results can be useful, they are restrictive in that their derivation requires certain normality assumptions that are described by Kolen (1985b).

Recently, Braun and Holland (1982, pp. 32–35) derived a standard error formula for linear equating using the delta method without making any normality assumptions for the situation in which randomly equivalent groups of examinees are administered the forms to be equated. Their resulting standard error expression suggests that standard errors of equating based on normality assumptions may produce misleading results when score distributions are skewed or are more peaked than a normal distribution.

Because skewed distributions are typical of many testing programs, especially in licensure and certification areas, and because Tucker equating is frequently employed in such areas, Kolen (1985b) used the delta method to derive standard errors for Tucker equating without any normality assumptions.

At about the same time, Jarjoura and Kolen (1985) derived standard errors for equipercentile equating with the common-item nonequivalent-populations design (the so-called frequency estimation method referred to by Angoff, 1971, and outlined by Braun & Holland, 1982, pp. 21–23). Kolen and Jarjoura (1987) described how these standard errors can be used in a cubic-spline analytic smoothing procedure for equipercentile equating under the common-item nonequivalent-populations design.

It must be emphasized that the aforementioned standard errors only index equating error that is due to *random* sampling of examinees; that is, all other things being equal, the magnitude of these standard errors will decrease as sample size increases. Practitioners often think of these standard errors as if they index equating error over multiple test forms or multiple administrations of the same test, which they do not.

Other Types of Errors in Equating

One type of equating error not taken into account by the above standard errors is bias in equating (or what might be called model misspecification error). One obvious issue in dealing with the problem of equating bias is specifying some criterion for equating in some context. Some perspectives on this issue are provided by Petersen, Cook, and Stocking (1983), Petersen, Marco, and Stewart (1982), and Kolen and Jarjoura (1987).

Consider, for example, the following situation. For a given integer score, i, on Form X, let $\hat{\ell}_i$ be some estimated linear equivalent, and let ℓ_i be its expectation over random samples of examinees. Also, let e_i be the presumably "true" equipercentile equivalent for integer score i. The quantity $E(\hat{\ell}_i - e_i)^2$ is a measure of squared error, associated with an integer score of i on Form X, in using the particular linear equating function as an estimate of the presumably true equipercentile equivalent. This squared error quantity can be decomposed as

$$E(\hat{\ell}_i - e_i)^2 = E(\hat{\ell}_i - \ell_i)^2 + (\ell_i - e_i)^2 \quad . \qquad (1)$$

Note that $E(\hat{\ell}_i - \ell_i)^2$ is the variance of the random errors (i.e., the square of the standard error) in using $\hat{\ell}_i$ as an estimate of ℓ_i, and $(\ell_i - e_i)^2$ is the square of the bias or systematic error in using the linear function rather than the presumably true equipercentile relationship.

When viewed from this perspective, it is clear that an equating procedure that has relatively little

random error in some context (as indexed by the square of the standard error of equating) might have considerable systematic error (as indexed by squared bias relative to some criterion). Consequently, it would seem sensible to select an equating procedure that minimizes squared total error. The problems in doing so, however, are sometimes formidable. First, of course, an appropriate criterion must be selected, which is not necessarily a simple task. Then, the data necessary to estimate squared bias must be obtained, which may require an experiment that is difficult and costly to conduct.

The random and systematic errors discussed above are the only types of errors that have received substantial attention in the equating literature, with greater attention given to random error. However, a broad perspective on error in contexts where equating is employed necessitates consideration of other potential sources of error. For example, sometimes researchers or practitioners are interested in the extent to which examinees' equated scores correspond to scores on the initial form in an equating linkage plan, or perhaps some other form of special interest. In particular, the form to which a new form is equated is frequently of little inherent interest; the old form employed is used because it is convenient. This is defensible because the old form itself has been equated (directly or indirectly) to an even older form that is of more fundamental interest, in some sense. With few exceptions (see Braun & Holland, 1982, p. 36), the error that accumulates as a result of multiple equatings in an equating linkage plan or chain has not been considered in the literature. At a minimum, it would seem sensible that the degree of confidence in the stability of equating should be inversely related to the number of equatings necessary to progress from the new form to the initial form or some form of special interest.

Another potential source of error, and one that is frequently overlooked, is the error introduced when equated scores are rounded to the units digit (or any digit, for that matter) for the ostensible purpose of simplifying the scores reported to examinees and decision-makers. As discussed rather extensively by Kolen (1986), the mere act of rounding an equated score can be a source of considerable error. Furthermore, if reported scores are rounded, it is important that the *un*rounded equated scores (or estimates of parameters based on unrounded equated scores, such as means and standard deviations) be used in arriving at the equating relationship. Otherwise, the error introduced by rounding will accumulate, and the equating chains will propagate the error to subsequent equatings.

Also, a broader conception of equating error should recognize that decision-makers almost always interpret equated scores as being more or less invariant over some time period during which examinee ability is likely to remain unchanged. This concern suggests a consideration of measurement error in the sense of generalization over occasions of administration. It might be argued that this type of error is not properly viewed as part of the total error in equating. However, practitioners and decision-makers do not really care about the finer points of what should or should not be incorporated in equating error per se. Usually, what they want is an answer to a question of the following type: "How likely is it that an examinee with a particular equated score would get the same score on a different form administered on a (slightly) different occasion?" Implicit in this question is a consideration of all of the types of error discussed above. However, there currently does not seem to be an answer to this question that is both practical and theoretically defensible.

Additional Issues Relevant to Equating Error

Congeneric model procedures. In Tucker equating the statistical assumptions and procedures make no reference to underlying true scores or measurement error variances. Rubin (1982), among others, expressed a clear preference for such procedures. Procedures that ignore true scores are relatively easy to understand and often simple to use. Also, even though the literature on comparing equating procedures is not very extensive, the literature that does exist tends to lend favorable support to these procedures.

However, there is something a bit disconcerting and seemingly inconsistent about using equating

procedures that neglect quantities such as true scores and error (of measurement), when these quantities are such an integral part of the other psychometric procedures routinely employed in the same testing programs. This point might be used as a basis for preferring Levine's procedures, but the literature provides, at best, inconclusive support for Levine's procedures in many equating situations (see, e.g., Kolen, 1985a; Petersen et al., 1982).

Petersen et al. (1982) reported results for congeneric-model procedures that were suggested by Lord and Tucker in personal communications. Also, Rock (1982) discussed a confirmatory factor analysis model for equating that employs congeneric tests, and Woodruff (1986) considered congeneric models for equating. Although some of the work in this area appears promising, congeneric-model approaches are not as yet widely accepted or used, even though they have considerable intuitive appeal. Why is this so? Perhaps there is some subtle inconsistency between such models and what occurs in practice, or perhaps one of the available models is much better than current evidence or practice suggests. In any case, congeneric models for equating are worthy of further pursuit, because they offer the possibility of systematically incorporating consideration of measurement error variance in equating contexts.

Inconsistencies between idealized equating designs and equating practice. In the common-item nonequivalent-populations design a new Form X, which is administered to a sample from Population 1, is equated to an old Form Y, which was administered to a sample from Population 2. At least that is the way the design is usually described. Actually, however, nothing in the assumptions or derivations of the Tucker or Levine equating procedures takes into account the time-dependent statement about old and new forms. Consequently, the mathematics in Kolen and Brennan (1987) would proceed unchanged without specifying which form is the new one and which is the old one. This is part of what it means to say that an equating function is invertible or symmetric, as opposed to the situation in regression analysis where the regression of Y on X is not the same as the regression of X on Y.

Furthermore, in terms of assumptions and derivations, the equating procedures discussed in the literature do not directly take into account the form(s) to which the so-called old form was itself equated. In other words, these procedures do not directly take into account the set of linkages in a full equating plan. Rather, these linkages are taken into account through the so-called transitive property of equating—namely, if X is equated to Y, and Y is equated to Z, then X is equated to Z.

With regard to these symmetric and transitive properties of equating, it is worth noting that in many equating contexts (especially in licensure and certification areas) these properties are untestable, at least in any manner that approaches optimality. The principal problems with testing these assumptions are that only one form is administered on any given test date, and it is virtually impossible to readminister an old form. These constraints are largely the result of security concerns that have been exacerbated by current movements toward (almost) full test disclosure. In short, it is often unknown whether these properties hold in practice.

Also, because the actual process of equating so often involves a clearly time-dependent plan in which old forms are indeed *old*, it may be reasonable to ask an heretical question about whether the symmetric and transitive properties of equating should be sacrosanct in all cases. For example, practitioners sometimes seem more interested in predicted true scores, in some sense, than scores that would have been obtained on some other form of a test. However, traditional predictors of true scores (e.g., regressed score estimators) are seldom sensible in equating contexts, because they assume parallel test forms, and equating is unnecessary in such circumstances. One alternative would be to use the best linear predictors of composite universe scores discussed by Jarjoura (1983), but this is a difficult method and it has not yet been employed.

Examining equating adequacy. The new *Standards for educational and psychological testing* (American Psychological Association, 1985) encourage the practice of obtaining periodic checks on the adequacy (they actually use the word "stability") of equating. It would be foolish to argue against such a well-intentioned goal, but it would

be equally foolish to claim that optimum checks on equating are achievable, given the practical constraints that typify most testing programs. Perhaps the ideal check would involve periodically readministering and reequating previously equated forms, and examining the extent to which the equating functions are invariant. However, it is often impossible, or at best impractical, to implement such a procedure.

What can be done is usually much less ambitious, and the results are likely to be much less compelling. For example, a new form can be linked (whenever possible) to two old forms. If the two links give similar results, then there is evidence of equating consistency for the particular form being equated. Also, most persons who have performed equating for a particular program over at least several years have learned to expect certain kinds of similarities in certain statistics involved in equating. If these similarities are not evident, then the statistics and their context are usually reconsidered, with the expectation of finding some concrete explanation for the inconsistency. This admittedly highly subjective check on stability often turns out to be much more valuable and informative than might be expected.

Another approach to examining equating stability—an approach that has become somewhat popular—is to use existing data and equate a form to itself through two or more forms. This is sometimes called equating in a circle, and is illustrated in Figure 1 where X is equated to Y, Y to Z, and Z to X. Under this paradigm, assuming that the process begins with raw scores on X, the third equating gives equated raw scores for X; these will equal the actual raw scores for X if equating is perfect.

Specifically, suppose the three equatings are linear and are represented by

$$y(x) = a_x(x) + b_x \quad , \tag{2}$$

$$z(y) = a_y(y) + b_y \quad , \tag{3}$$

and

$$x(z) = a_z(z) + b_z \quad . \tag{4}$$

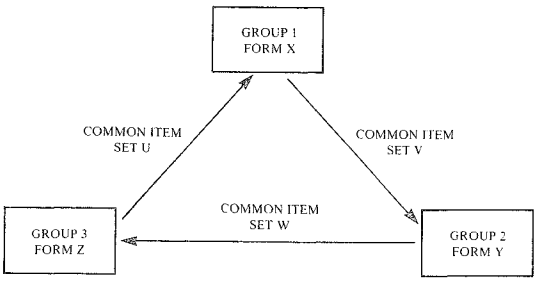
Then, it is easy to show that the equated raw scores for x are given by

$$x' = a_x a_y a_z(x) + (a_z a_y b_x + a_z b_y + b_z) \quad . \tag{5}$$

Note that $x' = x$ can be guaranteed by setting all the slopes to 1 and all the intercepts to 0. In other words, if the results of equating a test to itself through the circular equating paradigm are viewed simplistically as the criterion for good equating, then "perfect" equating can be achieved through no equating. This result is strictly an artifact of the paradigm, not a statement about the relative merits of equating versus no equating.

Sometimes a particular circular equating paradigm is used to examine equating adequacy for a variety of potential equating procedures. In this case, an equating procedure involving the estimation of only one or two moments often tends to appear better than one involving the estimation of many moments. For example, using this paradigm with real data, mean equating often appears to be better than linear equating, which in turn often appears to be better than equipercentile equating. This is largely an artifact of the paradigm, not a statement about the relative merits of mean, linear,

Figure 1
Circular Equating Paradigm



and equipercentile equating. Therefore, it is potentially quite misleading to evaluate the relative merits of a linear equating procedure versus an equipercentile procedure using the circular equating paradigm. It is more sensible to use this paradigm only for comparisons of equating procedures involving the same number of moments (e.g., comparing two different linear equating procedures), but even then, results should be interpreted cautiously. One reason for such caution is discussed below.

Another potential source of ambiguity with the criterion of equating a test to itself is that different results can be obtained depending upon *which* test is equated to itself. For example, given the situation described above, equating Form Y to itself gives

$$y' = a_x a_y a_z(y) + (a_x a_z b_y + a_x b_z + b_x) \quad , \qquad (6)$$

and equating Form Z to itself gives

$$z' = a_x a_y a_z(z) + (a_x a_y b_z + a_y b_x + b_y) \quad . \qquad (7)$$

Equations 5, 6, and 7 have the same slope but different intercepts. However, in the present authors' experience with real data, the discrepancies in intercepts are not likely to be too serious.

Conditions conducive to good equating. In actual equating practice, the task of selecting an equating procedure generally involves some mixture of knowledge of psychometrics and equating procedures, combined with some well-informed experience with equating in general and the program under consideration in particular. Indeed, in the present authors' opinion, the state of the art is not yet far enough advanced to justifiably permit statistical rules or tests to be the sole determinant in selecting an equating procedure. One reason for the need for reliance on subjective judgment is that, in practice, it is often impossible to check empirically the assumptions involved in an equating procedure. (See Braun & Holland, 1982, pp. 27–31 for methods to check some of the assumptions for the nonequivalent-populations design.)

Even so, reasonable guidelines can be identified for facilitating good equating. Table 1 lists some conditions that have been found to be conducive to satisfactory equating in many circumstances. Often, as more of these conditions are met, it becomes easier (i.e., less ambiguous) to select and defend an appropriate equating procedure. However, it should be noted that adherence to the conditions in Table 1 places a considerable burden on test development personnel.

Good equating usually requires a collaborative effort between test development personnel and psychometricians. In particular, test development personnel are seriously misinformed if they believe psychometricians can employ equating procedures to eradicate test development problems. Equally, psychometricians run a real risk of coming up with nonsensical equating results if they are ignorant of the testing program or the realistic issues that confront test development personnel. In equating there is little room for letting territorial imperatives override collaborative efforts among persons whose primary responsibilities may differ.

Some Other Issues

Content Specifications and Equating

In horizontal equating, the forms to be equated should measure the same abilities. For tests organized according to a table of specifications, this requirement means, among other things, that the same table of specifications should apply to all forms. Also, there is evidence to suggest that equating is improved when the common items linking these forms constitute a miniature version of the full-length test, especially in the sense of having the same proportional representation of items from content categories in the table of specifications (see Klein & Jarjoura, 1985). In many testing programs content specifications change over time, and logic suggests that any substantial content changes should trigger a rescaling and renorming of the test with a new "original" form to which subsequent forms are equated. However, it is sometimes not clear what should be judged to constitute a "substantial" content change, and usually it is content experts, not psychometricians, who possess the qualifications to make this judgment. Also, from a practical point of view, it is sometimes politically unpalatable to undertake a rescaling and renorming, because users often tend to resist the reeducation required to interpret and use new scores.

Table 1
Some Conditions Conducive to a Satisfactory Equating

1. **Test Structure**
   - A. Test content and statistical specifications are well-defined.
   - B. Test content and statistical specifications are stable over time.
   - C. When the test form is constructed, statistics on all or most of the items are available from pretesting or previous use.
   - D. The test is reasonably long (at least in the 35-40 item range, and preferably longer).
   - E. The scoring keys are stable.
   - F. There are two sets of common items (double-link). One link form is no more than one year in the past. Also, one of the link forms was administered in the same month as the form to be equated.
   - G. Each set of common items is at least 1/5 of the total test (at least 30 items for long tests).
   - H. Each common item set is representative of the total test in terms of content specifications and statistical characteristics.
   - I. Each common item is in approximately the same position in the old and new forms.
   - J. The common item stems, alternatives, and stimulus materials (if applicable) are identical in the old and new forms.
2. **Examinee Groups**
   - A. The examinee groups are stable over time.
   - B. The examinee groups are relatively large. Below 400 can be especially problematic.
3. **Administration Conditions**
   - A. The test and test items are secure.
   - B. The test is administered under carefully controlled (standardized) conditions.
4. **Field of Study/Training**
   - A. The curriculum, training materials, and/or field of study are stable.

Even so, changes in test specifications are perhaps more prevalent than is generally realized. Consider, for example, a longstanding program testing knowledge of law. In such a program, the content matter may change over time as older laws or interpretations of such laws are superseded by newer ones. It may even be the case that the correct answers to certain items change over time with changes in the law. The same type of situation frequently occurs in medical testing programs where new research suggests a change in the preferred treatment for certain illnesses. Furthermore, in admission testing programs like the ACT Assessment and the SAT, there have probably been enough changes in content over the years to cast at least some doubt on the meaningfulness of interpreting a current score on a recent form in terms of how an examinee would have performed on the original form.

Given the almost inevitable shifts in test content over time for most longstanding testing programs, it may be reasonable to consider developing equating procedures, or adjustments to equating results, that take into account gradual changes in content specifications without necessitating a complete break with the original score scale. It is not clear how best to do this, but it does seem that score interpretability would be enhanced if such procedures could be developed.

Equating in the Context of Cutting Scores

In many testing programs, especially licensure and certification programs, test forms are equated using a currently available procedure, but principal interest focuses on a particular score or range of scores that are the primary basis for decision making. For example, suppose that a score of 70 on the original form was defined as the passing score, and the principal purpose of conducting equating is to assure that an equated score of 70 on subsequent forms has the same meaning as the score of 70 on the original form. Currently available equat-

ing procedures can be employed for this purpose, but they do not take into account the fact that maintaining the meaning of a scaled score of 70 is likely to be much more important than maintaining the meaning of other scaled scores. Furthermore, in many licensure and certification testing programs, cutting scores are often at a point considerably below the mean, where the linear equating procedures previously discussed are likely to provide less stable results than at points closer to the mean.

In such situations it would seem desirable to have procedures that pay particular attention to equating in the region of a cutting score, even at the potential expense of poorer equating at other scores. (Indeed, in some cases, there is no need for equating at other scores.) One approach to achieving this goal might be to use the frequency estimation method for nonequivalent populations that was mentioned in the previous discussion of standard errors of equating. For example, a conceptually simple approach would be to find the new-form equipercentile equivalent of the passing score that was established for the original form, and then employ some modification of one of the linear equating procedures such that a straight line passes through this point. In effect, in such an approach, the equipercentile equivalent of the passing score would play the role of a "pivot" point for the linear relationship, just as a mean score plays such a role in the currently available linear equating procedures.

Reequating

Anyone who has been involved in equating for a substantial period of time has almost certainly encountered the following type of situation. A form of a test has been administered and equated, and subsequently it is discovered (usually through an irregularity report generated by a question from a particularly insightful examinee) that an item possesses some type of ambiguity that makes the keyed alternative technically incorrect, or that the keyed alternative is only one of two or more technically correct answers. Suppose for the sake of specificity that after reconsidering such an item, content matter specialists decide that the originally keyed alternative (say, *a*) is indeed correct, but the other alternatives (say, *b*, *c*, and *d*) can also be defended as correct, based on an obscure fact or facts. Clearly, decisions must be made about whether to give all examinees credit for the item and whether to reequate the form with that item scored correct for all examinees. (For the sake of the present discussion, assume that even examinees who omitted the item would be given credit for it.)

Suppose a firm decision on these matters is postponed until the form is reequated under the assumption of giving all examinees credit for the item. At this point, there are four conceivable ways to arrive at examinee "equated" scores:

1. Original key applied with original equating relationship;
2. Original key applied with revised equating relationship;
3. Revised key applied with original equating relationship; and
4. Revised key applied with revised equating relationship.

The first option essentially means acting as if the item is *not* flawed. The examinee who discovered the flaw may well consider this option to be unfair and, in all likelihood, the public will share the examinee's concern. However, often an examinee who is insightful enough to recognize such a flaw is also insightful enough to choose the alternative that was intended as the correct answer, rather than one of the other alternatives that are correct based on an obscure fact. If so, the first option does not really treat that particular examinee unfairly, although it would be unfair for some other unidentified examinee who chose one of the other alternatives *for a correct reason*.

The second option, using the original key with the revised equating relationship, is difficult to defend under any reasonable scenario.

The third option, using the revised key with the original equating relationship, may appear to be an option that is generous to examinees. In effect, all examinees who selected alternatives *b*, *c*, or *d* (or omitted the item) will receive a higher "equated" score, whatever the reason for selecting that alternative. However, those examinees who are given credit unjustifiably (e.g., those who had misinformation or no information about the item) will fare

better than their equally able counterparts, especially in a quota-based decision process. Thus, while this option is generous for some examinees, that very generosity may create a potential disservice to other examinees. The point here is that in evaluating the fairness or reasonableness of any of these options, it is necessary to be mindful of the consequences for not only examinees who are directly affected, but also examinees who are indirectly affected by the decision.

The fourth option, using the revised key with the revised equating relationship, essentially avoids the problems mentioned above with the third option, and the fourth option has a fair amount of face validity. Indeed, most people believe that the fourth option is obviously the best one to employ, and more often than not this appearance of face validity is judged to be an overwhelming argument in favor of the fourth option.

However, under some circumstances it can be argued that the first option may well be preferable *psychometrically* to the fourth if the goal is to be as fair as possible to *all* examinees, not just those who voice a legitimate complaint. For example, when everybody is given credit for an item, the effective test length is reduced by one item, which, on average, benefits lower-ability examinees and works to the disadvantage of higher-ability examinees. To put it another way, when all alternatives are keyed correct because an item possesses an obscure ambiguity, it is likely that many examinees will be given credit for the item who would not otherwise have answered the item correctly. This fact will cause these examinees to appear more able than they actually are, and other examinees will appear less able by comparison. Indeed, examinees who selected alternative *a* (the response originally keyed as correct) will receive a *lower* equated score under the fourth option than under the first option. Reequating cannot really eradicate these problems. Indeed, reequating can never completely remove a test development flaw; the best it can do is mitigate the impact of such a flaw.

The above points are not intended to be interpreted as arguments in favor of never rescoring or reequating when a flawed item is discovered. Even if the psychometric arguments were rather compelling, arguments from other perspectives could be even more compelling. Nor are these points to be interpreted as arguments relevant to the differential utility of benefiting lower-ability examinees versus disadvantaging higher-ability examinees. When such judgments need to be made, they should be based on a much broader set of considerations than merely psychometrics. The point here is that the issues involved in rescoring and reequating are quite complex, and certain unintended negative consequences can easily be overlooked. (These problems become even more complex when the flawed item is in a common-item equating section.)

Effects of a Security Breach on Equating

It is unfortunately true that security breaches occasionally occur in the administration of a test. If a security breach occurs in specifically identifiable test centers, then examinees in such centers should be excluded from the sample used for obtaining an equating relationship. However, on occasion a security breach may be more pervasive, or it may not be discovered until after equated scores are reported. Whatever the circumstances surrounding a security breach, it is useful to have some perspective on the *potential* effect of such an abnormality on equated scores if no corrective action were taken. In no sense should the following discussion of this issue be interpreted as support for a ''do-nothing'' position. The intent here is simply to examine some of the consequences on equated scores of a security breach that went undetected or uncorrected. In order to keep the problem manageable without losing too much generality, a simple case is considered below.

Suppose that, in the absence of a security breach,

$$\mu_s(Y) = \mu_s(X), \quad \sigma_s(Y) = \sigma_s(X) \tag{8}$$

and

$$\mu_1(V) = \mu_2(V), \quad \sigma_1(V) = \sigma_2(V) \tag{9}$$

where the notational conventions are those in Kolen and Brennan (1987). These assumptions imply that the two groups are indistinguishable and the two forms are indistinguishable in the linear equating context considered here. Also, to simplify results,

assume that, for the synthetic population, $w_1 = 1$ and $w_2 = 0$, and V is an *internal* set of common items. Under these circumstances, it is easily shown that

$$\ell(x) = x \quad , \tag{10}$$

which will be called the "true" equating relationship. Considered below are two simple cases illustrating what happens when a security breach affects only Form X.

Case 1. Suppose the security breach has a constant effect on all examinees in the equating sample for the new Form X, but the security breach affects only the common item section of Form X. Even more specifically, suppose that as a result of the security breach a *very* difficult common item becomes *very* easy for examinees who took Form X. Under these circumstances, the mean of the common items in Form X will be almost one point higher than it would have been without the security breach, but the standard deviation of the common item scores, $\sigma_1(V)$, will remain essentially unchanged. These assumptions, applied to the first five equations in Table 3 in Kolen and Brennan (1987), yield

$$\begin{aligned}\hat{\ell}(x) &\doteq x + \mu_s(Y) - \mu_s(X) \\ &\doteq x + [\mu_2(Y) + \gamma_2] - \mu_1(X) \\ &= x + [\mu_2(Y) - \mu_1(X)] + \gamma_2 \quad .\end{aligned} \tag{11}$$

Note that for the "true" equating relationship in Equation 10, the assumptions $\mu_s(Y) = \mu_s(X)$ and $\mu_1(V) = \mu_2(V)$ imply that $\mu_2(Y) - \mu_1(X) = 0$ for the "true" equating. In terms of an estimate, however, $\mu_2(Y) - \mu_1(X) \doteq -1$ because examinees in the equating sample who took Form X answered approximately one more common item correctly. Hence,

$$\hat{\ell}(x) \doteq x + (\gamma_2 - 1) \quad . \tag{12}$$

For the linear equating methods considered here, $\gamma_2 - 1$ (with an internal set of common items) is necessarily positive, and therefore, $\hat{\ell}(x)$ in Equation 12 provides an upwardly biased estimate of $\ell(x)$ in Equation 10.

In evaluating the effect of the biased estimate in Equation 12, it is necessary to distinguish between examinees who benefited from the security breach and those who did not. Consider, for example, two examinees who would have answered 9 items correctly without benefiting from the security breach [i.e., $\ell(x) = 9$], and assume that the first examinee actually did benefit from the security breach but the second examinee did not. For the first examinee,

$$\hat{\ell}(x) \doteq (9+1) + (\gamma_2 - 1) = 9 + \gamma_2 > \ell(x) \quad , \tag{13}$$

and for the second examinee

$$\hat{\ell}(x) \doteq 9 + (\gamma_2 - 1) = 8 + \gamma_2 > \ell(x) \tag{14}$$

because $\gamma_2 > 1$. In other words, both examinees have upwardly biased estimates of $\ell(x)$, but the bias is 1 point greater for the examinee who benefited from the security breach.

Case 2. As in Case 1, suppose the security breach has a constant effect on all examinees in the equating sample for the new Form X. However, for Case 2 assume that the security breach affects only the *non*-common item section of Form X. Specifically, suppose that a *very* difficult non-common item becomes *very* easy for examinees who benefited from the security breach. In this case, all of the assumptions that led to the "true" equating relationship in Equation 10 still hold (to a high level of approximation) *except* that in this case $\mu_s(X) \neq \mu_s(Y)$. Specifically,

$$\begin{aligned}\hat{\ell}(x) &\doteq x + [\mu_s(Y) - \mu_s(X)] \\ &= x + [\mu_2(Y) - \mu_1(X)] \\ &\doteq x - 1 \quad ,\end{aligned} \tag{15}$$

because $\mu_1(X)$ is inflated by examinee performance on the single, compromised non-common item.

Consider, again, two examinees who would have answered 9 items correctly without benefiting from the security breach [i.e., $\ell(x) = 9$], and assume that the first examinee actually did benefit from the security breach but the second examinee did not. For the first examinee,

$$\hat{\ell}(x) \doteq (9+1) - 1 = 9 = \ell(x) \quad , \tag{16}$$

but for the second examinee

$$\hat{\ell}(x) \doteq 9 - 1 = 8 < \ell(x) \quad . \tag{17}$$

In other words, when a security breach occurs in the non-common item section of a new form, ex-

aminees who benefit from the security breach receive their "correct" equated scores, but examinees who do not benefit from the security breach receive equated scores that are too *low*.

Even though these two cases are oversimplified, the above results illustrate some of the pitfalls of assuming that a security breach will always have a similar effect on all examinees' equated scores. The situation is much more complicated than that. The direction and magnitude of the effect of a security breach on an examinee's equated score depends on which items are affected (common or non-common items) and whether or not an examinee benefited from the security breach.

Conclusions

It has been asserted that equating is impossible unless forms are perfectly parallel, but if forms are perfectly parallel, then equating is unnecessary. Such assertions are based on a stringent definition of equating that is simply unattainable in practice. From a pragmatic point of view, therefore, the (frequently unstated) goal of equating is to arrive at a conversion equation, function, or table that *approximates* an ideal equating *in some sense*. Just as there is no such thing as *the* reliability of a test, so too there is no such thing as *the* conversion equation for a test form. There are many possible conversion equations, depending on any number of assumptions—many of which are often untestable in practice. Therefore, it behooves those who perform equating and employ the results of equating to be knowledgeable about and sensitive to the types of issues raised in this paper.

Merely raising these issues may well cause practitioners, examinees, and decision-makers understandable uneasiness. One very practical additional issue, therefore, is that psychometricians accurately communicate the limitations, restrictions, and errors in equated scores without leaving the impression that equating is a needless, useless, or hopelessly confusing process. This type of communication may be difficult, but its absence can easily lead to inappropriate, exaggerated, or misleading interpretations of scores.

References

American Psychological Association. (1985). *Standards for educational and psychological testing*. Washington DC: Author.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 508–600). Washington DC: American Council on Education. (Reprinted by Educational Testing Service, Princeton NJ, 1984).

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9–49). New York: Academic Press.

Jarjoura, D. (1983). Best linear prediction of composite universe scores. *Psychometrika, 48,* 525–539.

Jarjoura, D., & Kolen, M. J. (1985). Standard errors of equipercentile equating for the common item nonequivalent populations design. *Journal of Educational Statistics, 10,* 143–160.

Kendall, M., & Stuart, A. (1977). *The advanced theory of statistics* (Vol. 1, 4th ed.). New York: Macmillan.

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. *Journal of Educational Measurement, 22,* 197–206.

Kolen, M. J. (1985a, April). *Comparison of methods for linear equating under the common item nonequivalent populations design*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Kolen, M. J. (1985b). Standard errors of Tucker equating. *Applied Psychological Measurement, 9,* 209–223.

Kolen, M. J. (1986). *Defining score scales in relation to measurement error* (ACT Technical Bulletin No. 50). Iowa City IA: American College Testing Program.

Kolen, M. J., & Brennan, R. L. (1987). Linear equating models for the common-item nonequivalent-populations design. *Applied Psychological Measurement, 11,* 263–277.

Kolen, M. J., & Jarjoura, D. (1987). Analytic smoothing for equipercentile equating under the common item nonequivalent populations design. *Psychometrika, 52,* 43–59.

Lord, F. M. (1950). *Notes on comparable scales for test scores* (RB-50-48). Princeton NJ: Educational Testing Service.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 2,* 137–156.

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating models. In P. W. Holland & D. B. Rubin (Eds.), *Test equating*

(pp. 71–135). New York: Academic Press.

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

Rubin, D. B. (1982). Discussion of "Observed-score test equating: A mathematical analysis of some ETS equating procedures." In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 51–54). New York: Academic Press.

Woodruff, D. (1986). Derivations of observed score linear equating methods based on test score models for the common item nonequivalent populations design. *Journal of Educational Statistics, 11,* 245–257.

Author's Address

Send requests for reprints or further information to Robert L. Brennan, ACT, P.O. Box 168, Iowa City IA 52243, U.S.A.