Technical and Practical Issues in Equating: A Discussion of Four Papers

William H. Angoff
Educational Testing Service

Many of the articles on equating that have appeared during the last 35 years have been concerned with the development and exposition of alternative models of equating, their error functions, and their robustness in the face of violations of the assumptions basic to their development. The four papers presented here are somewhat different. Their purpose, generally, is to go beyond theory, to examine the implications of special problems observed in the application of equating methodology, to search for clarifications and improvements in technique, and to investigate ways in which equating methods may be applied to practical testing problems. Each paper addresses a different set of problems; the present discussion will not attempt to find common issues among them, but will consider each separately in serial order.

Cook and Petersen

The Cook-Petersen (1987) paper deals with issues related to equipercentile equating and smoothing, the invariance of equating in the face of population differences, and the properties of anchor (common) items used in equating. As the authors point out, the equipercentile method is largely empirical and data-dependent, and is driven very little by theoretical models. As a consequence, it is highly vulnerable to inadequacies resulting from overall sample size. Even when samples are large, irregularities in distributions (especially near the extremes where the frequencies are low and often zero) will cause irregularities in the observed conversion function. These irregularities are regarded as having no lawful significance; thus, in an attempt to approximate the ''true'' or ''population'' conversion, some technique is sought to smooth them out. This is done either by presmoothing methods applied to the distributions themselves prior to equating, by postsmoothing methods applied to the conversion function itself, or both. In the course of their discussion the authors offer a brief review of some of the smoothing methods that have received attention in the last 35 years.

The data-dependent nature of equipercentile equating means that its outcome is simply the result of the application of the method with little or no dependence on theory, except for the assumptions that (1) the two groups involved in the operation are drawn at random from the same population, and (2) the instruments being equated are parallel forms of the same test (the latter assumption is, of course, basic to all equating). In contrast, the linear model for equating not only makes the foregoing assumptions, but also assumes in advance of the equating operation that the conversion function is in fact linear. To the extent that the linear and equipercentile outcomes are systematically differ-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 11, No. 3, September 1987, pp. 291–300
© Copyright 1987 Applied Psychological Measurement Inc.
0146-6216/87/030291-10$1.75

ent, it is probably justifiable to say that the assumption of linearity is unsupported.

Neither method of equating is entirely satisfactory. The linear method presupposes a model that may well be incorrect; it is often used as a convenient approximation to a more nearly accurate curvilinear relationship. And while the equipercentile method makes no presupposition as to the form of the conversion (its outcome is dictated entirely by the data), it is degraded by the natural irregularities in the data themselves, and causes the investigator to ask what the "true" direction of the conversion might be. Although the linear method is both perfectly smooth and perfectly verifiable, allowing no room for subjective decision, it must be remembered that the initial assumption of linearity is itself a subjective choice.

Ultimately, the subjective factor in equating will never vanish; it can only be minimized. There has been a great deal of research in equating methodology over the last 15 to 20 years, but none in which analytical methods involving moments above the second have been investigated. It might be useful to conduct an equipercentile equating on a sample of the available cases in order to form some hypotheses regarding the nature of the function that would best fit the observed conversion. Equating models chosen to fit moments above the mean and standard deviation might then provide the desired curvilinear functions while also yielding verifiable results.

In an ideal data collection design, two test forms, X and Y, are administered to all the individuals in a group, but in such a way that their responses to the second form are unaffected by their experience of taking the first form. Inasmuch as this condition cannot be achieved in practice, researchers must resort to a design in which the entire group is separated into two random, and by definition equivalent, halves, each half taking one of the two forms (X or Y) to be equated.

There are, however, some improvements that can be introduced in this procedure. Even groups that are defined by random selection may differ to some degree; thus it is useful to administer to both groups a common (anchor) test (V) and to use the scores on this test to adjust, to the extent that its correlations with X and Y permit, for any differences remaining between them.

When the two groups are not formed by random selection methods, it is clear that no equating is possible *unless* there is an anchor test available to correct for differences between the groups. But some constraints need to be imposed on the anchor test and some principles need to be stated:

1. When the two groups (Group 1 taking Form X and Group 2 taking Form Y) are truly randomly equivalent, then any anchor test is appropriate. The more highly the anchor test correlates with X and Y, however, the more effective it will be in correcting for any differences between the two groups. Ideally, the anchor test V should be parallel to both X and Y, in which case it will correlate with them to the extent permitted by its reliability, and will therefore be maximally useful. But parallelism between V and X (and between V and Y) is not fundamentally necessary; any correlated V will help.
2. When the two groups are not randomly equivalent, then only an anchor test that *is* parallel to X and Y can give trustworthy results. The intent of this second principle is to satisfy the assumption, basic to Tucker (see Angoff, 1971) and Levine (1955) equating, that the regression of X on V in all its particulars—intercepts, slopes, and errors of estimate—is the same for Groups 1 and 2. (A similar assumption is made for the regression of Y on V.) In practice, the violation of this fundamental assumption most often occurs at the item level, especially in the equating of achievement tests in subject matter areas in which learning and forgetting are relatively rapid. If it were possible to plot the item-test regressions, item by item, for the two nonrandom groups, the effect of this violation would be clearly visible. More than likely, this effect is responsible for the differences in the pairs of item plots in Figures 1 through 4 in the Cook-Petersen article. Cook and Petersen found in their study that when the anchor items appearing in the two forms were administered to two groups of students tested at the same point in their coursework

and at the same time in relation to their degree of sophistication in the subject, the plots of deltas fell closely on the line drawn through the points, showing correlations very nearly unity (.98 to .99). However, when the groups were chosen at different times during the school year, the points were much more scattered, showing delta-plot correlations ranging from .74 to .92. These results can be construed as evidence that an interaction has taken place between performance on the items and the time of testing relative to instruction, in turn suggesting a pattern of differential learning and forgetting on these items.

With the Cook-Petersen data in hand as an illustration of what can go wrong in the use of common items, this writer would urge researchers who carry out such equating operations to conduct a differential item difficulty study for each equating, either by means of delta-plot procedures as shown by Cook and Petersen or by any other appropriate "item bias" procedure, in an effort to reveal inadequacies in the anchor test items.

What was revealed by the Cook-Petersen study of item difficulties is not, as they suggest, that the groups were at different *ability levels*, but that the groups showed different *patterns of performance* on the anchor items. Even had the groups been selected differently—for example, one from a high-scoring Fall population and the other from a low-scoring Spring population—in such a way as to guarantee nearly equal overall means and standard deviations in the two groups, there is a good likelihood that a significant interaction between item performance and group would still have been found, indicating that the items were in fact not "common items" at all in the psychological and educational sense, but had somehow interacted with the time of testing and were therefore useless for equating.

In general, it is clear that considerable care must be exercised in the use of anchor test items. Test items administered in different settings, while superficially the same items, do not in fact always present the same psychological task to all groups. It is often the case that speededness differences, order differences, context differences, etc. will affect the apparent difficulty of the task experienced by the examinee. True, when the anchor items are presented in precisely the same order as a separately-timed test to the different groups, some of these problems vanish. But when the anchor items are scattered through Forms X and Y (and also represent operational scorable units in those forms), there is a potential danger that they will be differentially difficult for those groups.

With regard to the precept that equating should be invariant with respect to the populations on whom the equating is based, it would seem on the face of it that when the two tests in question are measures of the same function (as they should always be), there should be one and only one "true" or "population" conversion; variations in conversions found from different samples should be no greater than those expected from the standard error functions. (It is only when the two forms are measures of different psychological traits that the conversions may be expected to differ as a function of the population used in the equating; see Angoff, 1966.) It was to test this hypothesis—the invariance of equating results with respect to population differences—that the Angoff and Cowell (1986) study was designed, and in part to test this hypothesis that the Kingston, Leary, and Wightman (1985) study was designed. In the Angoff-Cowell study the populations of interest were intentionally chosen to maximize the differences on the test selected for analysis. Yet even in this situation, expressly designed to strain the population-independence hypothesis and test its limits, it was found that a general conversion applied quite well to variously defined subgroups of the population. In only one instance did some question remain regarding the invariance hypothesis, and that involved the equating of a heterogeneous test in a highly selected subpopulation. The results of the study confirmed this writer's view that it is quite reasonable to hypothesize a single "true" conversion relating two parallel forms, existing independent of the particular populations used to generate the conversion, and applicable to all populations.

Fairbank

Fairbank's (1987) paper describes a well-designed and well-executed study of smoothing methods applied either to the distributions of scores prior to equating, or to the conversion function itself after equating. The presmoother of choice, he finds, is the negative hypergeometric distribution, a function developed by Keats and Lord (1962) for tests scored by number correct, generated with knowledge of the number of items and the mean and variance of scores. The preferred postsmoother was found to be the cubic spline, described by Kolen (1983).

There is little to comment on in Fairbank's paper in the context of equating issues. It differs from the other three papers in that it is a report of a comparative study of several methods of treating data. Three points, however, may be worthy of further consideration. One is that the hypergeometric distribution, valuable though it is, does have the limitation of being appropriate only to data from tests scored by number correct. It would be very useful to have a procedure that can also be adapted to data scored by other procedures. A second point is that there are situations in which the negative hypergeometric fails to fit the data well, yielding consistently lower (or higher) frequencies than those observed below the mean and consistently higher (or lower) frequencies above the mean. Lord (1969), after a series of trials, found that a procedure he refers to as Method 20, which makes use of a larger number of parameters, gives substantially improved fit.

The final point concerns the effect of the equating operation on the error of measurement and the effect of the smoothing function on both the equating operation and error of measurement. Assuming that the standard error of measurement (SEM) for a test is 100 points, the standard error of equating might be in the neighborhood of 15 points at the mean, depending on the method and the number of cases used for equating. [It should be noted here that the standard error of equating for Design I linear equating (see Angoff, 1971, pp. 94–97) rises rapidly at distances removed from the mean; at two standard deviations (SDs) from the mean, for example, the standard error of equating would be closer to 26 points.] This would mean that the overall error at the mean is $(100^2 + 15^2)^{1/2} = 101.1$, not much larger than the SEM itself. If the negative hypergeometric is introduced as a smoothing function, the standard error of equating is reduced by 10%, following Fairbank's findings, to perhaps 13.5 points at the mean, resulting in an overall error of $(100^2 + 13.5^2)^{1/2} = 100.9$, again not much different from the overall error before smoothing was introduced. This might suggest that the smoothing effort was not sufficiently effective to justify its use. However, three observations are worth making here:

1. Although the error of equating apparently has little effect on the overall error at the mean, the effect at some distance from the mean is greater.
2. The errors under consideration here include the errors of measurement for an individual examinee in addition to the errors of equating. The error patterns for *groups* of individuals are much different. As the group becomes larger, the standard error of measurement becomes smaller; in the case of very large groups, it is vanishingly small, leaving the error of equating as the dominant factor. This becomes clear when it is recognized that equating error is only observable over many replications of the equating study itself; it remains entirely unaffected once it has been determined and is applied to the scores of a tested group. It is embedded permanently in the data, where its effect closely resembles a bias; it affects the statistics of large groups as greatly as individual scores. Given these considerations, the value of the smoothing effort takes on a different complexion, suggesting that it is more useful than it may otherwise appear.
3. Although a smoothing function may not be overly useful in the main body of the data, say from −2 to +2 SDs from the mean, it is exceedingly useful in that region of the distribution where the data are scant and the investigator needs the kind of objective advice that only an analytic method of smoothing can provide. Thus, although the relative improvement may appear small near the center of the distribution and in the context of individual scores, it is in fact much more prominent, both at the

extremes of the distribution and in the context of group data, than may appear at first glance.

Kolen and Brennan

The Kolen-Brennan (1987) paper provides a highly insightful and useful reformulation of the Tucker and Levine equating models, and highlights similarities and differences between them that had not been fully elucidated in previous expositions. In doing so, the paper considers a more general definition of the group in terms of which the population parameters (the mean and variance of scores) on the two forms to be equated (Forms X and Y) may be taken to form the linear equation converting scores from one scale to the other. In previous formulations of the Tucker and Levine equations (e.g., Angoff, 1971) this group is taken as the aggregate of the individual groups taking both forms, in which each component individual group is weighted equally. In such formulations, the method that yields the desired population parameters calls for making estimates symmetrically for both Form X and Form Y.

In the Kolen-Brennan formulation, the "combined" or, in their terms, "synthetic" group is formed by giving full weight to the "new" group (the group taking the new form, Form X) and zero weight to the "old" group (that taking the old form, Form Y). (In effect, then, the synthetic group is not the combined or aggregate group at all.) As a consequence, the slope and intercept parameters for the conversion equation are taken from the *observed* mean and variance for the new group on Form X, and from the *estimated* mean and variance for that group on Form Y. In point of fact, despite the zero weight given to the old group, the data for that group are not entirely ignored; some of those data—the mean and variance on the anchor test, V—are used in forming the estimates of the population parameters for Form Y.

The slope (A) and intercept (B) parameters in the conversion equation, $Y = AX + B$, may be expressed, as formulated by Kolen and Brennan, in the following terms:

$$A = \frac{\hat{\sigma}_1(y)}{\sigma_1(x)} \tag{1}$$

$$B = \hat{\mu}_1(y) - \frac{\hat{\sigma}_1(y)}{\sigma_1(x)}\,\mu_1(x) \quad , \tag{2}$$

in which the subscripts 1 and 2 refer, respectively, to the new and old groups, and a caret denotes an estimated value. The authors might just as easily have derived the equation, $Y = A'X + B'$, in which

$$A' = \frac{\sigma_2(y)}{\hat{\sigma}_2(x)} \tag{3}$$

$$B' = \mu_2(y) - \frac{\sigma_2(y)}{\hat{\sigma}_2(x)}\,\hat{\mu}_2(x) \quad , \tag{4}$$

resulting from giving full weight to the old group (2) and zero weight to the new group (1), and yielding perhaps similar, but not identical, values for the parameters. As may be seen above, the formulation proposed by the authors (Equations 1 and 2) comes from the acceptance of the Form X mean and standard deviation as directly observed in Group 1, but the Form Y mean and standard deviation as *estimated* in Group 1. The alternative formulation (Equations 3 and 4) entails the acceptance of the Form Y mean and standard deviation as observed in Group 2, but the mean and standard deviation on Form X as estimated in Group 2.

Kolen and Brennan explain their choice of the first of the two formulations by saying that in an ongoing testing program, in which new forms are normally administered in every major test administration, the group of interest is the current group taking the new form, and that interpretations of the conversion relationship between the two forms are best made in terms of that current group. Although this is an understandable choice, given their implicit philosophy of equating and given their assumptions, the present writer, who subscribes to a somewhat different philosophy, differs with their view that such a choice is needed.

As was suggested earlier in this paper, the ultimate aim of equating is to provide an equation that describes the nature of the conversion of units from one instrument to another, without regard to the nature of the particular groups to whom these instruments were administered in the equating study. In this sense, the determination of the score conversion is analogous to the process of converting physical units from one scale to another. Such conversion equations are independent of the medium—and, ideally, the method—chosen as the experimental means for determining the conversion. That is, it should not matter whether the experiment designed for developing a Celsius/Fahrenheit temperature conversion equation was water, oil, ice, molten steel, or any other substance, for that matter; the conversion remains $F = 1.8C + 32$.

That this principle also applies to tests may be seen in the results of the study of population independency with respect to equating (Angoff & Cowell, 1986). In that study, subgroups that differed widely in type (ethnic background, sex, field of study) and in level of performance were chosen to form the databases for equating parallel forms. In spite of these differences, however, the conversions yielded by samples drawn from these various subgroups were found to differ nonsignificantly, for the most part, from an overall "population" conversion. In general, the conclusion drawn here—the hypothesis that there is one "true" conversion relating two parallel forms—is quite reasonable. Variations and discrepancies from that true conversion are random variations, not systematic, because with parallel forms and random groups (or approximations to random groups), there can be no interaction of test forms and groups.

This being the case, it appears that in selecting one of two possible ways to form their conversion, Kolen and Brennan propose an approach to equating based on philosophical grounds that are not fully supported by data such as those just described. Except in the rarest of circumstances, the two possible choices that they consider will necessarily yield different results. It would seem to this writer that the method of choice should have been one that provides a unique solution, one that weights the two sets of data equally. In such a procedure (which Kolen and Brennan decided not to use), the "synthetic" group would in fact be the combined group, some of whom would be taking Form X, others taking Form Y, and all taking Form V, the anchor test. The two groups entering into the combination are given equal weight (as mentioned above), and estimates of mean and variance on the two forms (X and Y) are made symmetrically, and in identical fashion, using all the data available.

Recall that in the Kolen-Brennan procedure, estimates are made for one form, but observed data are accepted for the other form. Not only is it true that this process does not use all of the available data (in itself a cause for some concern), it is also true that this asymmetry in treating the data probably yields population parameters that are biased with respect to one another, and thus results in a biased conversion equation. This is because the estimates of the mean and variance for Group 1 on Form Y are regressed estimates, limited by the correlation between Forms Y and V. In contrast, the use of the mean and variance for Group 1 on Form X is *not* a regressed estimate; it is an observed value.

It is true that in the data provided in Kolen and Brennan's illustration, the effect of this bias was very small. But in their illustration the two groups scored so nearly equally on Form V that the adjustments made in the estimates were necessarily very small. In other instances, however, where Groups 1 and 2 differ substantially and the correlations, r_{xv} and r_{yv}, are lower than they are in these data, this bias is likely to be of far greater consequence. Moreover, the estimates in the Kolen-Brennan formulation must deal with the *entire* difference between the two groups as observed for Form V; if the correlation, r_{yv}, is relatively low, the estimate is a poor one. In the typical Tucker equating, in which estimates are made symmetrically on Form X and Form Y for a combined group, each estimate must deal with only half of the difference between the groups. Even in this case, because the estimates are made symmetrically there is little consequent risk of bias.

Admittedly, the foregoing are merely this writer's opinions and speculations. What would probably contribute most to this discussion is the outcome of an empirical study in which both procedures would be examined and compared for random error, and particularly for bias.

The remainder of the Kolen-Brennan paper provides an interesting development in an effort to decompose the difference between $\mu_1(x)$ and $\mu_2(y)$ and between $\sigma_1^2(x)$ and $\sigma_2^2(y)$ into (1) differences associated with test forms, and (2) differences associated with groups. The purpose of these procedures is to learn how the two forms differ as to level and range of difficulty, and how the two groups differ as to level and dispersion of ability.

With regard to the first of these questions, the answer in the past has always been sought in the conversion function itself, which does provide some clues about the relative difficulties of the two forms, not only for the particular groups engaged in the equating data, but for any and all groups. Suppose, for example, the conversion equation is found to be $Y = 1.2X - 3.8$. This equation says that Form Y is more difficult than Form X (yields lower raw score values) below an X score of 19, but is easier than Form X (yields higher raw score values) above the X score of 19; that is, Form Y is more difficult for lower-scoring groups and easier for higher-scoring groups than Form X. (At score 19 the two forms are equally difficult.) What their relative difficulties are for any group in question depends on the level of performance of that group.

The slope parameter of the equation, $Y = 1.2X - 3.8$, indicates that any given group may be expected to have a standard deviation 20% larger on Form Y than on Form X; this will likely be true for any and all regions of the scale. However, if the conversion function were curvilinear, the relative sizes of standard deviations on the two forms would depend on the level of performance of the group in question.

If the two groups have taken different forms, the statistics earned on Form X for, say, Group 1 can be converted to the scale of Form Y, allowing direct comparison. If Group 1 earns a mean of 25 on Form X and Group 2 earns a mean of 27 on Form Y, this mean score of 25 is converted to its equivalent on the scale of Y, and the mean converted score for Group 1 is found to be 26.2, or .80 points lower on the Form Y scale than the observed mean of 27 for Group 2. The difference between mean scores on the scale of X is .67, also in favor of Group 2. And it should come as no surprise that the difference of .80 on the Y scale is 1.2 times the difference of .67 on the X scale, a piece of information already conveyed in the slope of the conversion equation.

A similar approach is applied to the standard deviations. If, for example, the SD for Group 1 is 5.5 on X and the SD for Group 2 is 6.1 on Y, and the SD of 5.5 is converted to its equivalent on the scale of Y, the converted SD for Group 1 is found to be 6.6, about 8% larger than the SD for Group 2. Similarly, the SD for Group 2 on the scale of X is about 5.1, with the unsurprising result that the SD for Group 1 is still about 8% larger than the SD for Group 2.

Thus, although this writer has no quarrel regarding the logic of the analyses given in this section of the Kolen-Brennan paper on decomposition (except for the problems they themselves recognize in decomposing the variances), some doubt remains as to the usefulness of their development when the answers to their questions regarding the comparisons between forms and between groups can be provided much more directly and easily from the conversion function itself.

Brennan and Kolen

The last paper in this group, by Brennan and Kolen (1987), offers a number of perspectives on the practicalities of equating, and makes clear that the literature on equating provides little or no guidance in the solution of several equating problems that often emerge in the conduct of an operational testing program. One problem is that the use of an inappropriate model for equating (e.g., a linear model, when an equipercentile model is more appropriate) may cause a bias in the equating results, often far more serious than the random errors of equipercentile equating itself. A second problem is that equating theory has so far provided little help in dealing with the matter of successive equatings—although there is some discussion of this problem in the literature, and some suggested solutions (see Angoff, 1971; Wilks, 1961) apparently are used by the authors themselves.

Brennan and Kolen are quite properly mindful of the accumulation of errors that can, and does, develop in the course of the administration of a testing program, and they seek ways of minimizing these errors. It should again be noted that unlike individual errors of measurement that cancel out and tend to vanish in the aggregate when data for large groups are formed, errors of equating do not vanish; they appear in means and other group statistics as prominently as they do in an individual score (in some senses, more prominently), and remain in the data where they have the effect of a bias. Therefore, in major assessment programs, where the groups are large and group differences are small, it is easily possible for means of groups, for example, to reverse themselves solely as a result of equating error. This is quite possibly the reason for some of the strange results of city testing programs sometimes reported in the news media.

Brennan and Kolen worry about the practicalities of checking on the ''stability'' of an equating program, in which scores on Form A and scores on Form H, for example, are considered equivalent, not because they have been equated directly but because they have been equated, albeit indirectly, over the course of time, through seven intermediate links and six intermediate forms: A to B to C to D to E to F to G to H. In this sense, ''stability'' means to them, as it does to this writer, simply that there has been no drift in the process. The obvious way to provide a check here is to conduct the additional equating of A to H directly and compare this result with the result of the operational equating over the successive links. This can be done either in a special administration or in an operational equating, in which an equating test, originally administered with the old Form A, is now administered with the new Form H in a design satisfying a Tucker or Levine equating model.

The process of equating ''in a circle'' as a check on stability, or, more generally, as a check on bias, is a process that causes Brennan and Kolen some doubts. The fact is, however, that the doubts they express seem not to be fully justified. If the values of the slope and intercept parameters in their Equations 5, 6, and 7 were specified in detail, the slopes in all three equations would equal unity, and (contrary to their first sentence following Equation 5) the intercepts in all three equations would be zero. This is precisely what would be expected when a test is equated to itself: $x' = ax + b$, where $a = 1$ and $b = 0$. What makes this technique of equating in a circle so useful is that, unlike the case in most situations, it provides advance knowledge of what the errorless result should be, and this knowledge can be used in evaluating the error in an actual equating chain. Indeed, the technique of equating a test to itself could be used quite profitably in making comparisons of the two definitions of the ''synthetic'' population, described in the Kolen-Brennan paper.

Brennan and Kolen's discussion of practical conditions conducive to good equating is excellent and strikes a responsive chord with all investigators who conduct continuing equating programs. The tensions they describe in their efforts to produce meaningful equating results are quite recognizable: (1) the understandable interest expressed, in the normal administration of a continuing testing program, in introducing new concepts and changing statistical specifications in the tests over the course of time; (2) the occasionally conflicting needs of the test development and statistical staffs in the pursuit of their respective responsibilities; and (3) the occasionally inevitable changes in test content resulting from changes in the world around us.

In view of these pressures for changes in the tests, a realistic view of equating is one that would parallel in some respects the so-characterized constant and continuing meaning of the cost-of-living index, which, while usefully equated over the short term, cannot possibly have meaning as a continuing scale over the course of decades. Clearly, the national market basket experiences continual changes, with new items added (e.g., television sets and food processors) and old items subtracted (e.g., iceboxes and corrugated washboards). The same is true for tests, especially achievement tests, that perforce must change with changes in the curriculum. Therefore, depending on the rapidity of change in the tests over the course of time, scores on the scale may be considered comparable over a limited time span, without necessarily believing that they have the same meaning over an indefinitely long interval.

Brennan and Kolen suggest quite properly, for tests employing cutting scores, that the equating be designed in such a way as to concentrate the group on which the equating is based in the region of the cutting score. Such a procedure would maximize the reliability of equating where it is needed most, at the expense—no real expense, in fact—

of scores in the region in which no decisions are made.

Brennan and Kolen consider other problems that develop in the course of an operational program, problems that call for the reequating of a test form. Sometimes, they point out, it is discovered after a test has been administered that there is a flaw in one or more items, rendering the previously agreed-upon key(s) impermissible. Sometimes it is discovered that an item (or more than one item) previously thought to be secure is in fact insecure, and has become known to a significant subset of the examinees. In these situations a number of considerations, sometimes conflicting with one another, must be balanced against one another in reaching a solution short of an entire readministration of the test. These considerations include many issues: issues of equity, of the public's *perceptions* of equity (with different publics and different individuals quite possibly perceiving the matter of equity differently), of the real and perceived integrity of the testing program and of tests generally, and other issues such as cost, schedule, available solutions, number of items involved, size of the error and size of the correction, or number of people affected.

There is no one solution, obviously, that will be satisfactory in all circumstances, and no one solution that will satisfy all the demands implicit in these issues. But in general, issues of equity must, of course, come first. When only one item is affected, the easiest and most effective solution may be to rescore all papers with the flawed item omitted, and then reequate the revised form. This reequating may be carried out very simply by calculating the proportion correct of the flawed item as originally scored and ''equating'' the corrected form (without the item) to the original form (containing the item), and to ignore for the sake of expediency the trivial effect on the standard deviation. In essence, what is sought here is a conversion of the corrected form to the reporting scale as though the corrected form had been administered in the first place. A more ambitious, but not significantly different, solution would be to rescore as before and to reequate the corrected new form to its predecessor form, Form Y, directly—as though the original new form had never existed. This is the solution of choice when it is found that several items are flawed, but not, generally, when only one item is involved.

The Brennan-Kolen paper might have gone on to consider other errors in the testing process including, for example, a mistiming error on the part of the supervisor. Consider a situation in which too little time has been given for the test. Should this error call for a reequating? How can this be carried out if the affected group is a small one, say consisting of only 11 examinees? How serious must the mistiming be before taking this kind of action, or, indeed, any kind of action? Should the decision to take action depend on the particular speededness characteristics of the test? Should everyone tested be given a bonus in points because of the undertiming? If so, how much of a bonus? Suppose the solution were sought in a reequating, and a literal reading of the reequating results calls for *subtracting* points for some examinees? How can this sort of action be justified? On the other hand, suppose the error was an overtiming. Are points to be subtracted from the students' scores? How would the examinees affected react to this? And again, how many points should be subtracted?

Still more situations can be considered. What is the correct approach if there is an unexpected power failure in the testing center, and the lights go out for a period of time; or if the school band strikes up a march under the windows during the testing period; or if one of the examinees experiences sudden illness during the testing, distracting the other students for a period of 10 minutes? Occasionally such problems can be solved satisfactorily through statistical means, but when the problems are such that the statisticians themselves lose confidence in the power of their methods to provide a solution, test administrators will regretfully (or perhaps with some relief!) seek administrative remedy, such as a rescheduling of a new administration, a solution to be considered in any case.

The administration of an operational testing program brings with it many problems (call them ''challenges''), and it is interesting, even gratifying, to those of us who are engaged in this work that useful solutions to some of these problems may be found in the technology of equating. The contributions that these four papers have made to the technology are excellent. Further contributions at their high level of quality would be welcome.

References

Angoff, W. H. (1966). Can useful general-purpose equivalency tables be prepared for different college admissions tests? In A. Anastasi (Ed.), *Testing problems in perspective* (pp. 251–264). Washington DC: American Council on Education.

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.)

Angoff, W. H., & Cowell, W. R. (1986). An examination of the assumption that the equating of parallel forms is population-independent. *Journal of Educational Measurement, 23,* 327–345.

Brennan, R. L., & Kolen, M. J. (1987). Some practical issues in equating. *Applied Psychological Measurement, 11,* 279–290.

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11,* 225–244.

Fairbank, B. (1987). The use of presmoothing and postsmoothing to increase the precision of equipercentile equating. *Applied Psychological Measurement, 11,* 245–262.

Keats, J. A., & Lord, F. M. (1962). A theoretical distribution for mental test scores. *Psychometrika, 27,* 59–72.

Kingston, N., Leary, L., & Wightman, L. (1985). *An exploratory study of the applicability of item response theory methods to the Graduate Management Admissions Test* (RR-85-34). Princeton NJ: Educational Testing Service.

Kolen, M. J. (1983). *Effectiveness of analytic smoothing in equipercentile equating* (ACT Technical Bulletin No. 41). 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.

Levine, R. S. (1955). *Equating the score scales of alternate forms administered to samples of different ability* (ETS Research Bulletin No. 23). Princeton NJ: Educational Testing Service.

Lord, F. M. (1969). Estimating true-score distributions in psychological testing (An empirical Bayes estimation problem). *Psychometrika, 34,* 259–299.

Wilks, S. S. (1961). *Scaling and equating College Board Tests*. Princeton NJ: Educational Testing Service.

Author's Address

Send requests for reprints or further information to William H. Angoff, Mail Stop 03T, Educational Testing Service, Princeton NJ 08541, U.S.A.