Introduction to Problems, Perspectives, and Practical Issues in Equating

Robert L. Brennan
American College Testing Program

Until a few years ago the subject of equating was largely ignored by most people in the measurement community except for those psychometricians who had specific responsibilities for equating, especially in large testing companies. Also, until a few years ago there was very little published literature available on equating—the principal exception being Angoff's (1971) treatment of the subject in his chapter on "Scales, Norms, and Equivalent Scores" in the second edition of *Educational measurement*. Indeed, for the decade of the 1970s, Angoff's chapter was the only extensive treatment of equating that was available to the measurement community at large, and it is still a major reference on the subject. In the early 1980s, however, the published literature on equating began to grow at a noticeable rate, and the importance of equating began to be recognized by a broader spectrum of people associated with testing. This is illustrated, for example, by the fact that the 1985 *Standards for educational and psychological testing* (American Psychological Association, 1985) devotes a substantial part of a chapter to equating, whereas the previous edition (American Psychological Association, 1974) does not even list "equating" in the index.

In the present author's opinion, this increased recognition of the importance of equating is attributable, at least in part, to two developments during the past decade. First, there has been an increase in the number and variety of testing programs that employ multiple forms of a test. The persons responsible for these programs have recognized that scores on such forms need to be equated so that, in some sense, examinees who take a more difficult form are not given an unwarranted advantage over examinees who take an easier form, and vice versa. Second, in order to address a number of the issues raised by testing critics in the last decade, developers and publishers of tests have sometimes found it necessary to reference, if not thoroughly discuss, the role of equating in arriving at reported scores. Such references inevitably gave equating an increased visibility among not only measurement specialists but also users of tests.

However, much of the newly generated literature on equating is very theoretical, and many of the public statements about equating are so vague that they are difficult to evaluate. Consequently, equating is too often viewed as a statistical "black box" which, without human judgment, is assumed to generate equated scores with ideal characteristics. Such is not the case. Indeed, from at least one perspective (Lord, 1980, pp. 196–198), "ideal" equating is impossible—except when it is unnecessary. The perspective on equating taken by most practitioners is not so absolute, however. Rather, most practitioners view equating as a process punctuated by three principal choices:

1. Choice of a data collection design (e.g., ran-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 11, No. 3, September 1987, pp. 221–224
© Copyright 1987 Applied Psychological Measurement Inc.
0146-6216/87/030221-04$1.45

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

dom groups design or the use of common items with nonequivalent populations);
2. Choice of an operational definition for the equating transformation (e.g., linear or equipercentile); and
3. Choice of a statistical procedure for estimating the equating transformation (e.g., Tucker or Levine equally reliable procedures for linear equating).

Of course, each of these choices or steps in equating involves certain theoretical considerations. With equal force, however, each of these steps typically involves a number of subjective judgments, and these judgments are usually best made based on experience with equating in general and the testing program in particular. Consequently, there are a number of problems and practical issues that must be addressed in the actual process of equating, and addressing these issues often requires considering different perspectives on them. Because these problems and practical issues have often been ignored in the published literature on equating, they are the principal focus of the series of papers in this issue. Of necessity, addressing these topics requires certain theoretical considerations, but theory does not dominate this series. It is hoped that the treatment of the topics raised in these papers will have two beneficial effects: (1) sensitizing researchers and practitioners to issues that should be considered in the performance of equating and, especially, the interpretation of equating results; and (2) motivating researchers to pursue solutions to the problems raised in these papers.

Overview of the Papers

In the first paper, Cook and Petersen provide a review of the literature concerning how various equating methods are affected by sampling error, sample characteristics, and characteristics of anchor test items. Their treatment of these issues involves both linear and curvilinear operational definitions of equating, as well as conventional and item response theory (IRT) methodologies.

In their review of the literature on sampling error, Cook and Petersen concentrate on studies that examine the effect on the accuracy of equipercentile equating of analytic techniques for smoothing marginal and bivariate frequency distributions. (One such study is extensively discussed in the second paper in this series.)

Characteristics of equating samples are discussed by Cook and Petersen in terms of a review of studies that examine whether an equating transformation remains the same no matter what group is used to define it. Their review suggests that group invariance of an equating transformation is affected by both the nature of the test forms (homogeneous or heterogeneous) and the nature of the examinee groups tested. For example, self-selected groups of examinees (e.g., those who elect to be tested on different dates) may vary in systematic ways that adversely affect equating results.

Cook and Petersen also provide a review of studies that examine the consequences on equating results of various characteristics of anchor test items. Anchor test items, or common items, are those that are used to link two forms of a test in an equating design. Their review suggests that the effectiveness of an anchor test depends on how closely it mirrors the full-length forms to be equated. Finally, Cook and Petersen identify gaps in the research they have reviewed and offer some suggestions to practitioners.

The second paper provides a discussion of a very extensive study by Fairbank of analytical smoothing methods for reducing sampling errors in equipercentile equating. Fairbank examines seven presmoothers, which are applied to the empirical (and usually bumpy) distributions of discrete scores before equating, and seven postsmoothers, which are applied to the resulting equipercentile points. The presmoothing methods examined include, among others, three- and five-point moving medians, three- and five-point moving weighted averages, and fitting the negative hypergeometric distribution. The postsmoothing methods include, among others, cubic splines, fitting a logistic ogive, and various regressions. These smoothing methods were studied using both simulated data and data from an operational administration of the Armed Services Vocational Aptitude Battery (ASVAB).

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Fairbank uses several statistical criteria to assess the effectiveness of the various smoothers in reducing random sampling error and bias. Based on these criteria, he concludes that fitting the negative hypergeometric distribution is the most effective presmoothing method, and orthogonal regression and cubic splines are the most effective postsmoothing methods.

In the third paper, Kolen and Brennan provide a reformulation of the Tucker and Levine equally reliable methods of linear equating with the common-item (or anchor test) nonequivalent-populations design. In this design, two groups of examinees from different populations take different test forms that have a subset of items in common. This design is frequently encountered in situations such as licensure and certification testing.

The formulation of the Tucker and Levine methods that is offered by Kolen and Brennan explicitly considers certain population notions. Particular consideration is given to the role of population weights in forming a synthetic population, which is the population for which the Tucker or Levine equating applies most directly. Also, the Kolen-Brennan formulation facilitates a consideration of the practical consequences of using these two equating methods. For example, using this formulation it is easy to show that the Levine method weights group differences more heavily than the Tucker method.

Finally, Kolen and Brennan provide formulas for decomposing observed differences in means and variances into parts associated with (1) form differences for the populations and (2) population differences on the old and new scales. These decompositions highlight one effect of different choices of synthetic population weights.

In the fourth paper, Brennan and Kolen consider a number of practical issues that arise in using or interpreting equating results. They give principal emphasis to certain issues relevant to the identification, quantification, and elimination or reduction of various sources of equating error. Among the sources of error they consider are random error attributable to using only a sample of examinees, systematic error (or bias) that results if the equating model is misspecified, error that accumulates over the multiple equatings in an equating linkage plan, and error attributable to unwarranted rounding of scores.

Also, Brennan and Kolen consider several issues involving procedures for examining the adequacy of equating. For example, they provide a discussion of the so-called ''circular equating paradigm'' for equating a test to itself. This paradigm is sometimes used to examine equating stability. The discussion presented by Brennan and Kolen suggests that results obtained using this paradigm are often ambiguous and potentially misleading.

Brennan and Kolen end their paper with brief discussions of four other topics that have practical consequences for equating. First, they discuss some of the issues that arise in equating when content specifications for test forms change over time. Second, they consider equating in contexts where cutting scores are involved. Third, they consider some of the equating issues that arise when a test form that has already been administered and equated is found to contain one or more flawed items. Finally, they consider some of the effects on equating of a security breach that gives some examinees an unwarranted advantage over other examinees.

In his discussion of these papers, Angoff treats both technical and practical issues. With respect to the Cook-Petersen and Fairbank papers, Angoff gives special and generally positive consideration to these authors' treatments of anchor (common) item characteristics, group invariance of an equating relationship, and the effect of smoothing methods on ''total'' error. Concerning the Kolen-Brennan and Brennan-Kolen papers, Angoff agrees with most of their comments and conclusions, but he is rather critical of certain statements they make about synthetic population weights and the circular equating paradigm. The last paper in this series provides a response by Brennan and Kolen to Angoff's concerns about these two issues.

References

American Psychological Association. (1974). *Standards for educational & psychological tests*. Washington DC: Author.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

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.)

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

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.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/