# 9 Tests in Transition: Discussion and Synthesis

Robert L. Brennan[1]

University of Iowa

In educational settings that focus primarily on student achievement, testing programs are almost always in a state of transition, or they should be! Over time, changes occur in curricula and student populations. It follows that if a testing program is to reflect what is happening in particular educational settings, it must evolve to align itself with those settings. Other evolutions occur when the conditions of test administration are modified. Movement to computer-based testing is an obvious example.

One of the ironies of educational measurement is that such changes in a testing program—even when they are widely viewed as improvements—might jeopardize score comparability to some extent, which is usually viewed as anything but an improvement! One route around this problem is to adopt a new scale, but for numerous reasons rescaling is often viewed as an unacceptable alternative.[2] So, frequently, it is decided to make certain adjustments to the testing program and/or psychometric "fixes" with the goal of keeping the score scale as unaltered as possible. Then the overarching question becomes, "Has the score scale been maintained adequately enough?" Psychometric evidence to address this question is primarily the focus of the chapters by Liu and Walker (Chapter 7) and by Eignor (Chapter 8).

Such psychometric evidence is generally viewed in terms of criteria for linking, for which there are many lists in the literature. For example, the list given by Liu and Walker (Chapter 7, Section 7.2.2) is as follows:

[1] Robert L. Brennan is E. F. Lindquist Chair in Measurement and Testing, and Director, Center for Advanced Studies in Measurement and Assessment, University of Iowa. The opinions expressed in this chapter are those of the author and not necessarily of the University of Iowa.

[2] Rescaling is considered in more detail in the last section of this discussion.

1. Same construct
2. Equity
3. Symmetry
4. Subpopulation invariance
5. Equal reliability
6. Same inferences
7. Same target population
8. Same measurement characteristics/conditions

The extent to which equity and subpopulation invariance are satisfied is largely a consequence of test-developer decisions that relate to the other six criteria. Both the Liu and Walker chapter and the Eignor chapter in this volume consider aspects of these other six criteria, but interestingly, Liu and Walker focus primarily on subpopulation invariance without much direct consideration of equity, whereas Eignor focuses more on equity issues without much consideration of subpopulation invariance.

In the next two sections, I provide a summary of these two chapters that is interspersed with my own comments. The final section provides a brief consideration of the need for an integration of equity and subpopulation invariance, followed by a consideration of linking versus rescaling.

## 9.1. The Liu and Walker Chapter on Test Content Changes

Liu and Walker discussed score linking issues related to test content changes, using the new SAT® to illustrate their points. Actually, in many respects, the new SAT plays such a central role in their chapter that the chapter itself might be viewed largely as a review of rationale, studies, and methodology used to support various decisions made about the new SAT.

Liu and Walker provide the following insightful focus for their chapter on score linking issues:

> At some point early in the redesign process, before we begin to investigate issues of score comparability, the testing organization must make a conscious decision about what is most important in the test revision. … The determination of this most important factor will have strong implications for the rest of the redesign process. … We need to ask ourselves: What do we want to achieve with the new test? What are the constraints? All the revisions and data collections should be guided by this redesign principle.

As Liu and Walker noted, in the context of the new SAT, the College Board stated a priori that they wanted the new critical reading (CR) test and old verbal (OV) test to be "equatable," as well as the new math (NM) test and the old math (OM) test. This a priori constraint influenced many aspects of the work done by the Educational Testing Service (ETS). Note that the new SAT also consists of a new writing (NW) test, for which a score scale had to be established, but that is not the focus of the Liu and Walker chapter.

To examine "equatability" Liu and Walker considered the following:

- Test specifications
- Item characteristics
- Empirical relationships between old and new tests
- Reliability for old and new tests
- Conditional standard errors of measurement (CSEMs) for old and new tests, and
- Subpopulation invariance for males and females.

To provide data for various empirical analyses, ETS conducted an extensive, well-designed, and well-executed field trial. The basic structure was as follows:

- Design 1: *Equivalent groups*. Each student took either a complete old SAT (OV + OM) or a complete new SAT (CR + NM + NW).
- Design 2: *Counterbalanced single group*. Each student took an old and a new component (OV and CR, or OM and NM).

The field trial, however, had one important limitation: Sample sizes were not sufficient for separate linkings for subgroups other than males and females.

### 9.1.1. Content Specifications and Item Characteristics

Liu and Walker provided a concise and excellent summary of content differences between the old and new SATs. Among the differences they cite between CR and OV are the following:

- Analogy items in OV were replaced by short reading passages in CR.
- There is a larger number of reading comprehension items in CR than in OV.
- Test length was reduced from 78 items in OV to 67 items in CR.

Among the differences that Liu and Walker cited between NM and OM are the following:

- There are no quantitative comparison items in NM.
- The content in NM was expanded to cover third-year college-preparatory math.
- Test length was reduced from 60 items in OM to 54 items in NM.

Additional differences between the old and new SATs incude the following:

- The introduction of NW that consists of both multiple-choice questions and a single essay prompt
- Section timing changes
- An increase in total testing time from three hours to 3 hours and 45 minutes

Liu and Walker concluded that "the content specifications on the new test do not suggest dramatic changes from the old test." With respect to item characteristics (deltas and biserials), OV and CR are very similar, as are OM and NM.

On balance, it appears that item statistics are more similar than are content specifications for the old and new SATs. This is not too surprising given the "redesign" context mentioned previously. Basically, most of the content changes were determined (tentatively) before the new SAT items were selected for the field trial; thus, it was possible to some extent to pick new SAT items that would likely perform similarly as a set to items in the old SAT.

### 9.1.2. Empirical Relationships

Liu and Walker used Pearson product-moment correlations $\left(r\right)$ and reductions in uncertainty $\left(\mathrm{RiU} = 1 - \sqrt{1 - r^2}\right)$ to quantify certain empirical relations between the old and new SAT. Dorans (2000, 2004d) argued that it is reasonable to require at least 50% reduction in uncertainty for test score linkage in high-stakes settings. This criterion requires that $r \geq .866$. Liu and Walker report that

$$r(\mathrm{CR,OV}) = .912 \to \mathrm{RiU} = .59 \text{ (i.e., 59\%)}$$

and

$$r(\mathrm{NM,OM}) = .922 \to \mathrm{RiU} = .61 \text{ (i.e., 61\%)}.$$

Clearly, the two $\mathrm{RiU}$ values exceed the 50% threshold, although this threshold is somewhat arbitrary. Another benchmark that can be considered is the old and new cross-test correlations

$$r(\mathrm{OV,OM}) = r(\mathrm{CR,NM}) = .79,$$

which are notably lower than $r(\text{CR,OV}) = .912$ and $r(\text{NM,OM}) = .922$, as one would hope and suspect.

Observed-score correlations can be informative for judging the adequacy of linking, but true-score correlations ($\rho$) that approach unity are essential for an argument that a linking deserves to be characterized as equating. True-score correlations depend, of course, on reliabilities.

### 9.1.3. Reliability and CSEMs

With respect to reliability ($\text{Rel}$), Liu and Walker stated that "high reliability on both tests is needed to ensure that the equated scores are informative enough to be accepted by test users (Dorans, 2004d)." They go on to report that

$$\text{Rel(OV)} \doteq \text{Rel(CR)} = (.91 - .93)$$

and

$$\text{Rel(OM)} \doteq \text{Rel(NM)} = (.91 - .93)$$

These results are encouraging in two respects. First, letting $\rho$ designate true-score correlation, these results mean that

$$\rho(\text{CR,OV}) \doteq \rho(\text{NM,OM}) \doteq 1,$$

suggesting that the old and new tests are measuring similar constructs in an overall sense. Second, because the reliabilities are approximately equal, as are the standard deviations, the $\text{CSEMs}$ are also about equal (in the low 30s on the SAT scale.) These are important results in supporting the view that the score scale is maintained reasonably well, although these results do not guarantee that scores for all examinees are interchangeable.

### 9.1.4. Subpopulation Invariance for Males and Females

Liu and Walker pointed out that "when population invariance does not hold, it tells us that the differential difficulty of the two tests to be equated is not consistent across different subgroups." Methodology for examining subpopulation invariance is evolving at a rapid rate. Perhaps the most salient initial discussion was by Dorans and Holland (2000); additional perspectives are provided by Kolen and Brennan (2004), among others.

For the new SAT, sample sizes from the field trial were adequate for examining subpopulation invariance for males and females, only. Liu and Walker provide results for OV and CR in great detail; they state that stronger results (i.e., less subpopulation sensitivity) hold for OM and NM. Two types of statistics are reported by Liu and Walker:

1. The Dorans and Holland (2000) root mean square difference (RMSD) and root expected mean square difference (REMSD) statistics, which Liu and Walker usually evaluated relative to a "difference that matters" DTM of 5 .
2. Percentage indexes:
   - Percent of formula scores for which the absolute value of the total and subgroup conversions differ by more than 5 points, which will be abbreviated PS (i.e., percent of scores), and
   - Percent of examinees for whom the absolute value of the total and subgroup conversions differ by more than 5 points, which will be abbreviated PE (i.e., percent of examinees).

An excellent feature of the Liu and Walker discussion of subpopulation invariance is that they first provide results for two parallel OV forms; these results serve as an informative baseline for subsequent results based on CR and OV. Stated briefly, the subpopulation invariance study of the two OV forms resulted in $\text{RMSD} < 5$ at all scale score levels, and for both males and females $\text{PS} = 0$ and $\text{PE} = 0$ . These results strongly suggest that the linking of two OV forms deserves to be called an equating. Ideally, it would be desirable to have similar analyses for two CR forms, but two such forms were not available for the field trial.

The linking of OV and CR for males and females resulted in $\text{RMSD} < 5$ for all but very low scale scores: for males, $\text{PS} = 3.5$ and $\text{PE} = 0.7$ , and for females, $\text{PS} = 1.2$ and $\text{PE} = 0.4$ . These results suggest minor evidence of subpopulation sensitivity with respect to gender. Liu and Walker summarize these results in the following terms: "… based on the equatability analyses discussed above, we think that the term *equating* might be defended for the linkage from new critical reading to the old verbal, and for the linkage from new math to the old math."

There is a somewhat different perspective on these analyses, however, that might lead to a slightly more tentative conclusion. The RMSD and REMSD statistics compare the male (M) and female (F) linkings to the total-group (T) linking; these statistics do not compare the male and female linkings directly. When there are more than two subgroups, comparing each of them to the total group using RMSD and/or REMSD is convenient because it gives a single result regardless of how many subgroups are involved. When there are only two subgroups, however, a direct comparison of the two linkings seems to me to be an obvious comparison to consider. (Kolen & Brennan, 2004, provide statistics for pairwise linkings.)

Figure 7.2 in Liu and Walker plots scaled-score differences for M-T and F-T when two OV forms are linked. The difference between these two plots gives the M-F scaled-score differences. It appears from Figure 7.2 that even when two OV forms are linked, the M-F differences suggest a hint of subpopulation sensitivity around scale scores of 500 and near 800, using $\mathrm{DTM}=5$ as a benchmark. Using the same benchmark, when OV and CR are linked and the M-F differences are examined, Figure 7.3 suggests that there is some evidence of subpopulation sensitivity throughout much of the scale score range.

I would argue that when we consider subpopulation sensitivity there are two questions that are typically of interest. First, how large are the differences between the linkings for the various subpopulations? In the context of the Liu and Walker chapter, this question is answered by examining directly the M-F scaled-score differences. Second, when a decision is made to use the total-group linking operationally, by how much are examinees in the various subpopulations advantaged/disadvantaged? In the context of the Liu and Walker chapter, this question is answered by examining the M-T and F-T differences. In most cases, both questions are relevant, but the answers will not be the same. There is no unqualified "correct" perspective; these are simply two different perspectives that answer different questions.

### 9.1.5. Other Comments

The Liu and Walker chapter provided an excellent discussion of numerous issues that relate to linking, and an excellent review of the linking conducted for the new SAT. For this linking, the field-test design and data collection were superb, but it is important to keep in mind the practical constraints involved in the field test. One such constraint was that the data were not collected in an operational setting. For this reason and others, conclusions about subpopulation invariance for the new SAT are necessarily somewhat tentative. Firmer conclusions will be possible when a substantial body of operational data for the new SAT is available.

In their discussion of empirical relationships, reliability, and subpopulation invariance, Liu and Walker employed numerous statistics and often drew conclusions based in part on the magnitude of such statistics compared to some benchmark. Two obvious examples are the $50\%$ RiU criterion, which requires that $r \geq .866$ for test score linkage in high-stakes settings, and $\mathrm{DTM}=5$ for the SAT. Although I believe that a $\mathrm{DTM}$ standard provides a useful benchmark, I do not think that conclusions about subpopulation invariance should be based exclusively on a $\mathrm{DTM}$ standard (see Brennan, 2006). Population sensitivity, like most other psychometric issues, is a matter of degree. Exclusive use of any single benchmark can obscure this basic fact and lead to unwarranted or too firm conclusions. I am not quarrelling with the Liu and Walker discussion of these matters, but a word of caution seems in order.

## 9.2. Eignor Chapter on Mode of Administration

Eignor discussed "linking scores derived under different modes of test administration," with almost exclusive attention given to paper-and-pencil (P&P) testing and two varieties of computerized testing: computer-adaptive testing (CAT) and other nonadaptive forms of computer-based testing (CBT). Eignor discussed these different modes of administration in the context of three types of linking (equating, calibration, and concordance) and three designs (random groups, single group counterbalanced design, and nonequivalent groups anchor-test design). See Holland (Chapter 2) and Kolen (Chapter 3) for detailed treatments of types of linking and data collection designs, respectively.

### 9.2.1. Types of Linking

In the terminology used by Eignor:

- Equating requires that the two tests (or forms) measure the same construct at approximately the same level of difficulty and with the same reliability. Eignor noted that equity is satisfied for equated scores, and it is a matter of indifference to any examinee as to which form she or he takes. In this sense, scores that deserve to be called "equated" are "truly interchangeable," to quote Eignor. As an example, Eignor cited linking a linear CBT version of an extant P&P test built to the same specifications.
- Calibration also requires that the two tests measure the same construct at approximately the same level of difficulty, but reliabilities could differ. As an example, Eignor cited linking a CAT version of an extant P&P test. Eignor argued persuasively that in this case second-order equity will not be satisfied because conditional standard errors of measurement will differ for the CAT and P&P tests.
- Concordance requires that the two tests measure similar constructs, with somewhat similar levels of difficulty and reliability. Eignor argued that "scores that have been concorded cannot be treated as being interchangeable." As an example, Eignor cited a CBT test and a P&P

test constructed to somewhat different specifications (e.g., the use of innovative item types and/or updated test content for the CBT).

It seems that the above taxonomic terms and the examples might be misaligned sometimes. For example, it is not clear that linking scores for a P&P test and a linear CBT version of it will always result in "equated" scores in the sense used by Holland (Chapter 2, Section 2.4.1) and most recent treatments of equating and linking (e.g., Kolen & Brennan, 2004). CBT constrains certain types of behavior in ways that some examinees might consider frustrating or confusing, with a potential negative impact on at least some scores. Furthermore, some examinees' scores might be influenced by differences in clarity between the presentation of items (particularly figures) in the two administrative modes. Kolen (Chapter 3, Section 3.2) explicitly included the conditions of administration as a formal component in his treatment of linking relationships. As a consequence, his treatment has direct relevance for mode of administration studies.

When Eignor argued that "scores that have been concorded cannot be treated as being interchangeable," he could mean two things. First, such scores are *not* interchangeable; second, such scores should not be used interchangeably. The first statement is unarguable in the sense that such scores are not "equated." The second statement, however, focuses on "use" of scores, which immediately engages a number of practical issues. For example, the quintessential example of concordance is the linking of ACT® and SAT scores, which traditionally results in a single table of "equivalent" scores that *are* indeed used as if they were interchangeable. In my experience, arguing against such use is a lost cause, but cautioning users about potential errors in such use is both necessary and possible.

In my opinion, Eignor's discussion of equating, calibration, and concordance is primarily in the context of equity issues (what might be called the "matter of indifference" criterion), but his chapter does not get into technical details about equity. It is difficult to treat equity in a manner that is both practically useful and technically defensible. Although much work in this area remains to be done, a particularly useful article is provided by Hanson, Harris, Pommerich, Sconing, and Yi (2001). They introduced the terms "closely equatable scores" (equating), "weakly equatable scores" (calibration), and "nonequatable" scores (concordance). They focused on construct dis/similarity, first-order equity, and second-order equity, and they considered linkage at the level of individual scores and at the level of score distributions.

### 9.2.2. Designs

The majority of the Eignor chapter focuses on three designs and examples of them that have been discussed in the literature on linking computerized and P&P tests. This is an excellent discussion that is noteworthy for its comprehensiveness and clarity, and I make no attempt to summarize it. Rather, I focus here primarily on a few issues that I think might be somewhat arguable or merit more consideration. My concerns are very minor, however, compared to the quality of Eignor's discussion of designs.

*Random groups design*. For establishing a linkage between a computerized and P&P test, often a random groups design is preferable to other designs provided, as Eignor noted that differential dropout is not a significant problem and sample sizes are sufficient. Relative to a single group design, sample size requirements for a random groups design are larger. However, relatively small sample sizes are adequate using linear linking with a random groups design.

*Single group counterbalanced design*. A distinct advantage of the random groups design is that each examinee takes only one test or test form, which means that administration conditions in the study mirror those that will be used operationally. By contrast, for the single group counterbalanced design, each examinee takes two tests or test forms, which raises the distinct possibility of contamination due to practice and/or fatigue effects. Eignor provided an excellent discussion of these effects in the context of the single group counterbalanced design.

*Nonequivalent groups anchor test design*.[3] A crucial aspect of the nonequivalent groups anchor-test design is that, for this design to work well, the anchor test needs to mirror the full-length test in all respects (see Kolen & Brennan, 2004), including mode of administration. Also, in considering this design, it is helpful to consider the location of the anchor (before, after, or embedded) and whether the anchor is part of the score (interval or external). It appears that Eignor's discussion usually makes an implicit assumption that the anchor is external.

Eignor correctly noted that

> Both groups would need to take exactly the same anchor test in exactly the same position and in exactly the same mode. (The) same mode for the anchor test across tests … makes it difficult to conduct linkings with this design when the test and the anchor are to be given consecutively in one testing session.

[3] This design is sometimes called the common-item nonequivalent groups design (see Kolen & Brennan, 2004). See Kolen (Chapter 3) for further discussion of these designs.

There is one additional and potentially insurmountable problem when computerized and P&P tests are linked using the nonequivalent groups anchor test design. The crux of the matter is that the items in an anchor-test administered via computer will not necessarily function the same way in the P&P mode, and there is no way to circumvent this potential problem using the nonequivalent groups anchor-test design. See Kolen (Chapter 3, Section 3.2) for more on the role of mode of administration.

### 9.2.3. Other Comments

In at least two places Eignor noted that "estimated true scores on the reference form (for a CAT) can … be linked or calibrated with observed scores on the test given in the paper-and-pencil mode." It is rather natural to do this because observed scores (rather than true scores) are usually reported for a P&P test, whereas for a CAT often item response theory (IRT) theta estimates are transformed to IRT estimated true scores. Logically, however, it seems rather inconsistent to link observed scores (on a P&P test) with true scores (on a CAT) when both tests are presumably measuring the same construct. Note that there is no reason to believe that this linkage would be the same as a true-score to true-score linking or an observed-score to observed-score linking.

Eignor noted that "samples used in the linking should be representative of the population," which is clearly desirable. However, very often, linking is conducted using data outside an operational administration, and in such cases, practical data collection issues often render the data quite unrepresentative of the population that will take the new CBT. When this occurs, results need to be interpreted with caution.[4]

## 9.3. Additional Perspectives

The two preceding chapters in this part are very well written and well reasoned. They are truly state-of-the-art considerations of linking scores for tests that are undergoing changes in content specifications and/or

[4] Perhaps the quintessential example of unrepresentativeness is data typically used to create ACT–SAT concordances. By definition, the self-selected group of examinees who choose to take both tests is not the group for which the concordance will be used. There is no practical way to avoid this problem, but it does limit the scope of legitimate inferences. These concerns were discussed directly in the chapters by Dorans and Walker (Chapter 10), Pommerich (Chapter 11), and Sawyer (Chapter 12) in the section on concordance.

administrative conditions. However, the state of the art is not as far advanced as we might like in all respects. In particular, it seems that we need more integration of equity and subpopulation invariance in both theory and practice. Also, the two chapters discussed here only hint at one important question that almost always arises when tests undergo transitions—namely should scores be linked or should a rescaling be undertaken?

### 9.3.1. Equity and Subpopulation Invariance

As noted previously, the Liu and Walker chapter gives considerable attention to subpopulation invariance, whereas the Eignor chapter focuses more on equity issues in a general sense. Eignor, however, did make the insightful statement that "it would be particularly interesting to see whether linkings between CATs and paper-and-pencil tests, which have been shown not to satisfy the equity requirement, also do not satisfy the population invariance property."

Stated more broadly, I suggest that a deep understanding of linking requires an integrated treatment of both subpopulation invariance and equity (as well as other criteria, of course). Such a treatment remains to be developed. In my opinion, subpopulation invariance is the simpler matter. We have more statistical and psychometric tools to quantify it and more consensus about how to study it. By contrast, it does not seem that the field of psychometrics has achieved any consensus about how to study equity, although I believe that Hanson et al. (2001) provides some useful perspectives, as does Kim, Brennan, and Kolen (2005).

A theoretically coherent and practically useful integration of equity and subpopulation invariance would be a tremendous contribution to the field of linking. In the meantime, I suggest that any linking of tests in transition should give at least some consideration to both subpopulation invariance and equity (as well as other criteria, of course), even if the treatment is not as integrated as we might like, given current limitations of the field.

### 9.3.2. Linking Versus Rescaling

One of the most sensitive and potentially volatile issues often encountered when tests undergo transition is whether scores should be linked or rescaled. The comments I offer here are intended as a brief, general consideration of this matter, not evaluative comments specifically directed at the chapters discussed here or any particular testing program.

Actually, it is not quite accurate to characterize the situation considered here simply as linking versus rescaling. Consider, for example, the rescaling of the "new" ACT first administered in 1989 (see Brennan, 1989). Separate studies were conducted that led to a rescaling for each of the four tests in the ACT. In addition, for two of the tests, there were old-scale to new-scale linkings in the sense of concordances that were made available to users to facilitate transition from the old score scale to the new score scale.[5] In addition, of course, new forms of the tests in the new ACT were linked in the sense of equated. For the purposes of this discussion, the important point is that the new scales were indeed a break with the past in the sense that particular scores on the old scales did not have the same meaning on the new scales. So, in that sense, scores were rescaled rather than linked.

As noted previously, in educational settings that focus primarily on student achievement, testing programs are almost always in a state of transition, or they should be. Sometimes the transitions are abrupt; sometimes they are more gradual. For example, the introduction of the "new" ACT in 1989 and the recentering of the SAT (Dorans, 2002) were rather abrupt changes that involved a rescaling of scores for these programs. For less abrupt changes, a central concern is often whether the linking can be defended as an "equating."

In the usual course of events, from one year to the next, changes in testing programs are typically not dramatic, and seldom does anyone quarrel with calling the linking of scores from year to year an "equating." I suggest, however, that this common view might merit some qualification from at least two (somewhat related) perspectives. First, for almost all testing programs, when any given form is equated, the links to past years are seldom older than 3–4 years, if that. So, there is only indirect evidence about the maintenance of the score scale for a longer period of time.[6] Second, over an extended period of time, even small year-to-year changes could add up to substantial differences between old and new forms.

The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999) addressed the matter of rescaling as follows:

> Standard 4.16: If test specifications are changed from one version of a test to a subsequent version, such changes should be identified in the test manual, and an indication should be given

[5] These concordances were used only for a limited period of time.

[6] The indirect evidence is based on transitivity assumptions. For example, if Form G administered in 2005 is equated to Form D administered in 2000, and Form D was previously equated to Form A administered in 1997, then we claim that Form G has been equated to Form A—but only if all relevant assumptions are fulfilled.

> that converted scores for the two versions may not be strictly equivalent. When substantial changes in test specifications occur, either scores should be reported on a new scale or a clear statement should be provided to alert users that the scores are not directly comparable with those on earlier versions of the test.

On the surface, this standard might seem unambiguously clear. In my opinion, however, this standard provides relatively little practical guidance for determining when a rescaling should be undertaken. For the reasons discussed next, I am not at all sure that this standard could be written in a manner that would provide practical guidance applicable to all testing programs.

Most of the problem is how to interpret the two key phrases: “substantial changes in test specifications” and “directly comparable.” A related problem involves the inferences drawn with test scores. For example, if comparisons are typically made among examinees within a 4-year window, it might not matter much if test specifications change substantially only over a 10-year window. On the other hand, even relatively small changes in test specifications might influence a 20-year trend line.

The phrase “directly comparable” is also problematic. A strict interpretation of that phrase would seem to be that, for each and every examinee, it is a matter of indifference which form she or he takes. In this sense, “directly comparable” means that scores are “strictly interchangeable” (a phrase used in the comment to Standard 4.16) or, stated differently, the criterion of score equity is achieved in its fullest sense. As Lord (1980) noted decades ago, however, under this criterion, equating is either impossible or unnecessary! No one would argue about the ideal being equated scores in the strict sense of “directly comparable,” but this unachievable goal does not provide practical guidance with respect to when a linking can be justified as an “equating” or when changes in a testing program are so substantial that a rescaling should be undertaken. It is also worth noting that most of the literature on linking (except for equating) has been generated since the 1999 Standards (AERA, APA, & NCME, 1999) was published. It is not clear, of` course, whether this new linking literature would cause the authors of the Standards to modify Standard 4.16.

Rescaling might be a psychometric issue, but decisions about whether to rescale are seldom made by psychometricians. In my career, on several occasions I have suggested that rescaling be undertaken for particular testing programs. Usually that advice has been rejected outright or postponed, sometimes indefinitely. Resistance to rescaling is often visceral. Some reasons for this resistance are quite understandable (e.g., time, cost, analysis difficulties, communication complexities); other reasons are more subtle or even misguided. For example, some view rescaling as an implied admission of mistakes in the previous scale. Others honestly believe that a test can be improved without in any way altering the meaning of the scores.

In the future, my guess is that rescaling will continue to be a relatively rare undertaking, and arguments about the merits of linking versus rescaling will continue. Whether scores are equated, linked in some weaker sense, or rescaled, however, the overarching consideration in my opinion is that users be given appropriate guidance about score interpretation and use. Part of that guidance ought to be explicit indications of the amount of error in scores and in the likely uses made of scores, as well as admonitions about likely misinterpretations of scores.

### *Acknowledgments*

The author gratefully acknowledges helpful comments by Michael J. Kolen on a previous version of the manuscript.