An NCME Instructional Module on

Traditional Equating Methodology

Michael J. Kolen
The American College Testing Program

This instructional module is intended to promote a conceptual understanding of test form equating using traditional methods. The purpose of equating and the context in which equating occurs are described. The process of equating is distinguished from the related process of scaling to achieve comparability. Three equating designs are considered, and three equating methods—mean, linear, and equipercentile—are described and illustrated. Special attention is given to equating with nonequivalent groups, and to sources of equating error.

Student X takes a college admissions test for the second time and earns a higher score than on the first testing. Why? We might conclude that this higher score reflects a higher level of achievement. What if, however, Student X had been administered exactly the same test questions on the second testing as on the first testing? Then rather than indicating a higher level of achievement, Student X's score on the second testing might be inflated because X had already been exposed to the test items. Fortunately, most college admissions testing programs use a new test form (set of test questions) on each test date, so it would be unlikely for Student X to be administered the same test questions on two test dates.

The use of different test forms on different test dates suggests another potential problem, as illustrated by the following situation. Students Y and Z are applying for the same college scholarship that is based, in part, on scores on a test. Students Y and Z take the test on different test dates, and Student Y earns a higher score than Z. Is Student Y higher achieving than Z? What if Y took an easier test form than Z? If so, then Y could be unfairly advantaged relative to Z, because Y took an easier form. However, most college testing programs *equate* test forms. In equating, test scores are adjusted based on the difficulty of the form administered. If the test forms were successfully equated, then the difference observed between Y and Z could not be attributed to Y being administered the easier form.

The process of equating is used in situations where multiple forms of a test exist, and examinees taking different forms are compared to each other. Even though test developers attempt to construct test forms that are as similar as possible to one another in content and statistical specifications, the forms will still differ somewhat in difficulty. Equating is intended to adjust for difficulty differences, allowing the forms to be used interchangeably. After successful equating, examinees can be expected to earn the same score regardless of the test form administered.

There are processes similar to equating that are better referred to as *scaling to achieve comparability,* as suggested in the *Standards for Educational and Psychological Testing* (APA, 1985). One of these processes is vertical scaling (frequently referred to as vertical "equating"), which is often used with elementary achievement test batteries. In these batteries, students typically are administered test questions matched to their current educational level (e.g., grade), but scores over test questions matched to different educational levels are all reported on the same score scale (e.g., grade equivalents). This procedure allows the scores of examinees at different levels to be compared, and allows for the assessment of an individual's growth over time. Because the content of the tests administered to the students at various educational levels is different, scores on tests intended for different educational levels cannot be used interchangeably, even though they are reported on the same score scale.

Other examples of scaling include converting scores on one test to the score scale of another test, and scaling the tests within a battery so they all have the same distributional characterists. As with vertical scaling, solutions to these problems do not allow the tests to be used interchangeably because the content of the tests is different; that is, *equating adjusts for differences in difficulty—not for differences in content.*

Many of the procedures used in equating are also used in scaling to achieve comparability. In this module the focus is on equating. Angoff (1984) and Petersen, Kolen, and Hoover (in press) present more detailed discussions of equating and

Michael J. Kolen is Director, Measurement Research Department, American College Testing Program, P.O. Box 168, Iowa City, Iowa 52243. He specializes in educational measurement and statistics.

Series Information

ITEMS is a series of units designed to facilitate instruction in educational measurement. These units are published by the National Council on Measurement in Education. This module may be photocopied without permission if reproduced in its entirety and used for instructional purposes. Fred Brown, Iowa State University, has served as the editor for this module.

related issues, and Skaggs and Lissitz (1987) have reviewed the research on equating methods. Refer to these references for more in-depth treatments of the topics presented in this module.

Purpose and Context

As previously indicated, equating has the potential to improve score reporting and interpretation for testing programs that possess *both* of the following characteristcs: (a) alternate forms are administered and (b) examinees administered different forms are evaluated with reference to the same standard or norm group.

There are at least two alternatives to equating in situations where these two characteristics hold. First, raw scores can be reported, regardless of the form administered. As was the case with examinees Y and Z (considered earlier), this approach can cause problems because examinees administered an easier form will be advantaged and those administered a more difficult form will be disadvantaged. In addition, if raw scores are reported, it is difficult to disentangle test form differences from examinee group differences. As an example, suppose that the mean score on a 40-item test increased from 27 one year to 30 another year, and that different forms of the test were administered in the two years. Can we say what caused this increase? Without additional information it would be impossible to know whether this 3-point score increase was attributable to differences in the difficulty of the two forms, differences in the achievement level of the groups, or some combination of these two factors.

A second alternative to equating is to convert raw scores to other types of scores so that certain characteristics of the score distributions are the same across test dates. For example, in a testing program that tests twice a year, say in February and August, the February raw scores might be converted to scores having a mean of 50 among February examinees and the August raw scores converted to have a mean of 50 among August examinees. Suppose, in addition, that an examinee somehow knew that August examinees were higher achieving, on average, than February examinees. In which month should the examinee take the test to maximize her score? Because the August examinees are higher achieving, it would be more difficult to get a high converted score in August than in February, so it would be to the examinee's advantage to take the test in February. Therefore, under these circumstances, examinees who take the test with a lower achieving group are advantaged and examinees who take the test with a higher achieving group are disadvantaged. Furthermore, trends in average examinee performance cannot be addressed using this alternative because the average (converted) score will be the same, regardless of the achievement level of the group tested.

TABLE 1

Hypothetical Conversion Tables for Three Test Forms

Form 1		Form 2	Form 1		Form 3	Form 2	
Raw	Scaled	Raw	Raw	Scaled	Raw	Raw	Scaled
•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•
30	15	30	29	15	30	29	14
29	15	29	28	14	29	28	14
28	14	28	27	14	28	27	13
27	14	27	26	13	27	26	13
26	13	26	25	13	•	•	•
25	13	•	•	•	•	•	•
•	•	•	•	•	•	•	•

When equating is successful, equated scores are not affected by the problems that occur with these two alternatives, because equating will adjust for differences in the difficulty of test forms. Unfortunately, it is not always possible to conduct an adequate equating. If certain assumptions are not met, application of equating methods can make matters worse than using either of the alternatives just discussed.

Scaling/Equating Process

Equating can be viewed as an aspect of a more general scaling/equating process. In this process, a scale for reporting scores is established at the beginning of a testing program (or at the time that a test is revised). This score scale is chosen to enhance the interpretability of scores by incorporating useful information into the score scale so as to avoid misleading interpretations. Incorporating normative information is one way to enhance score interpretability. For example, IQ scores are scaled to have a mean of 100 at each age level for a nationally representative group of individuals, and grade equivalents are scaled to indicate the typical performance of students in a particular grade.

Score scales typically are established using a single test form. For subsequent test forms, the scale is maintained through an equating process that places scores from subsequent forms on the score scale that was established initially. In this way, a scaled score of 26 means the same thing regardless of the test form administered.

The hypothetical conversions shown in Table 1 illustrate the scaling/equating process. The first two columns show the conversion of Form 1 raw scores to scaled scores. For example, a raw score of 28 on Form 1 converts to a scaled score of 14. (At this point we need not be concerned with what particular method was used to convert raw scores to scaled scores.) Note that the first two columns do not involve equating—only scaling.

Now assume that an equating process indicates that Form 2 is uniformly one raw score point easier than Form 1. Then, for example, a raw score of 29 on Form 2 would reflect the same level of achievement as a raw score of 28 on Form 1. This conversion of Form 2 raw scores to Form 1 raw scores is shown in the second set of columns in Table 1. What scaled score corresponds to a Form 2 raw score of 29? The answer is a scaled score of 14, because a Form 2 raw score of 29 corresponds to a Form 1 raw score of 28 which, from the first pair of columns, corresponds to a scaled score of 14.

To carry the example one step further, assume that Form 3 is found to be uniformly one raw score point easier than Form 2. Then, as illustrated in Table 1, a raw score of 30 on Form 3 corresponds to a raw score of 29 on Form 2, which corresponds to a raw score of 28 on Form 1, which corresponds to a scaled score of 14. Later on, additional forms can be converted to scaled scores by a similar chaining process. (A new form also could be directly equated to the original form.) The result of a successful scaling/equating process is that reported scores (i.e., scaled scores) on all forms are on the same scale and, therefore, can be used interchangeably.

Equating Methodology

Three interrelated issues must be considered when equating tests. First, a design for collecting the data for equating is needed. A variety of designs for data collection are used, and practical concerns usually enter into the choice of the design. Second, what constitutes correspondence between scores on alternate forms needs to be defined. In traditional equating methods, score correspondence is defined by setting certain characteristics of score distributions equal for a specified group of examinees, for example, the means and standard deviations of two forms might be set equal for a particular group of examinees. (Item response theory methods use a different definition of score correspondence. These methods are not considered here, but will be considered in a forthcoming instructional module by Cook and Eignor, in press.) Third, the statistical methods used to estimate the defined score correspondence must be specified.

Equating Designs

A variety of designs are used for collecting data for equating, and the choice of a design involves considering both practical and statistical issues. Three commonly used designs are illustrated in Figure 1. Assume that a conversion for Form 1 raw scores to scaled scores has been developed, and that Form 2 is a new form to be equated to Form 1.

Single group design. In the single group design the same examinees (Group A) are administered both Form 1 and Form 2. What if Form 1 were administered first to all examinees, followed by Form 2? If fatigue were a factor in examinee performance, then Form 2 could appear to be relatively more difficult than Form 1 because examinees would be tired when administered Form 2. On the other hand, if familiarity with the test increased performance, then Form 2 would appear to be easier than Form 1. To avoid effects such as fatigue and practice, the order of administration of the two forms usually is counterbalanced. In one method for counterbalancing, one-half of the test booklets are printed with Form 1 following Form 2 and the other half are printed with Form 2 following Form 1. In packaging, booklets having Form 1 first would be alternated with the booklets having Form 2 first. When the booklets are handed out, the first student gets Form 1 first, the second student Form 2 first, the third student Form 1 first, and so on. This *spiraling* process helps to ensure that the examinee group receiving Form 1 first is comparable to the examinee group receiving Form 2 first.

Suppose that the single group design is used to equate two forms of a 100-item test, and that the mean for Form 1 is 72 and the mean for Form 2 is 77. Assume also that a large representative examinee group is used, and that counterbalancing effectively controls factors such as fatigue and practice. What can be concluded about the relative difficulty of the two forms? Because the mean for Form 2 is five points higher than the mean for Form 1 for the *same* examinees, we can conclude that Form 2 is on average five raw score points easier than Form 1.

In addition to the need to control for factors such as practice and fatigue, other practical problems can restrict the usefulness of the single groups design. Because two forms must be administered to the same students, testing time is doubled, which often is not feasible. In addition, it is often not possible to administer more than one form at a single administration.

Random groups design. The random groups design is the second design shown in Figure 1. A spiraling process typically is used to implement this design, where alternate examinees in a test center are administered Form 1 and Form 2. This spiraling process leads to comparable (randomly equivalent) groups taking Form 1 and Form 2. As with the single group design, if large representative groups of examinees are used, then the difference between means on the forms is a direct indication of the average difference in difficulty between the forms.

From a practical standpoint, the random groups design is often preferable to the single group design because each examinee takes only one form of the test, thus minimizing testing time. Like the single group design, the random groups design requires two (or more) forms to be available and administered at the same time, which may be difficult in some situations. Because different examinees take the two (or more) forms in the random groups design, larger samples are needed for the random groups design than for the single group design where the examinees serve as their own controls.

Common item nonequivalent groups design. The third design is the common item nonequivalent groups design. This design typically is used when test security or other practical concerns make it impossible to administer more than one form per test date. In this design, Form 1 and Form 2 have a set of items in common, and different groups of examinees are administered the two forms. There are two variations of this design. When the score on the set of common items contributes to the examinee's score on the test, the set of common items is referred to as *internal.* Typically, internal common items are interspersed among the other items in the test. When the score on the common items does not con-

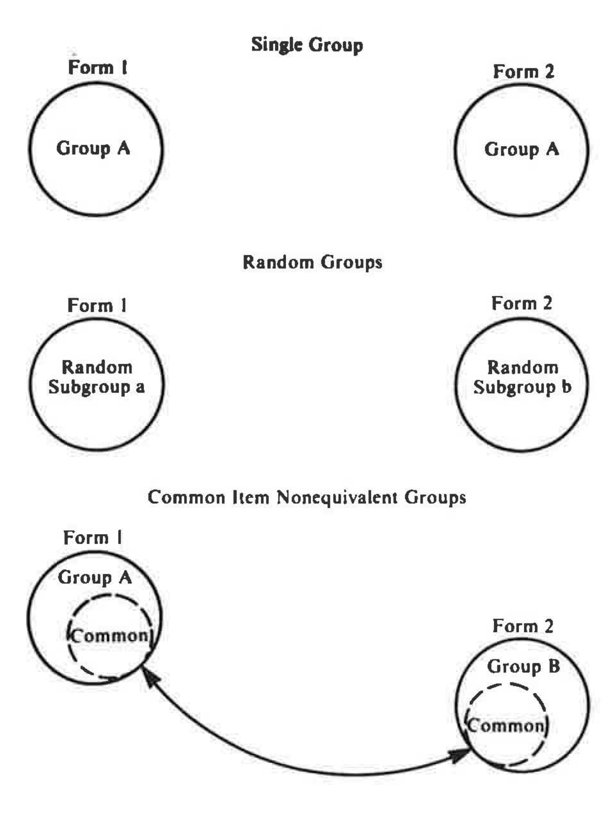


FIGURE 1. *Illustration of three data collection designs*

TABLE 2

Means for Two Forms of a Hypothetical 100-Item Test With 20 Common Items

Group	Form 1	Form 2	Common Items
A	72	—	13 (65%)
B	—	77	15 (75%)

tribute to the examinee's score on the test, the set of common items is referred to as *external.* Typically, external common items are administered as a separately timed section of the test.

Suppose that the common item nonequivalent groups design is used to equate two forms of a 100-item test, and that the raw score mean for Group A on Form 1 is 72 and the raw score mean for Group B on Form 2 is 77. From these data what can be concluded about the relative difficulty of the two forms? Although, on average, Group B scored 5 points higher on Form 2 than Group A did on Form 1, we don't know whether this difference is due to Group B being higher achieving than Group A, Form 2 being easier than Form 1, or some combination of these two factors.

To see if information on the common items will be of some help, refer to Table 2 where the means for the forms and for the common items are shown. Note that Form 1 and Form 2 each contain 100 items and there are 20 common items. The means for the common items suggest that Group B is higher achieving than Group A, because members of Group B, on average, correctly answered 75% of the common items whereas members of Group A correctly answered only 65% of the common items. That is, on average, Group B correctly answered 10% (75% minus 65%) more of the common items than did Group A.

Now reconsider the question, "Which of the two forms is easier?" To provide one possible answer to this question note that test takers correctly answered, on average, 5 (5%) more of the total 100 items (77 minus 72) on Form 2 than on Form 1. Because this difference (5%) is less than the difference for the common items (10%), we might conclude that Form 2 is *more difficult* than Form 1. By this reasoning, if the two forms were administered to the same group of examinees, Form 2 would be expected to have a mean 5 points lower (and, thus be 5 points more difficult, on average) than Form 1. This reasoning is a considerable oversimplification of how the equating actually would be accomplished; in fact, an equating method might even lead to the opposite conclusion about which form was more difficult. Still this example illustrates that the major task in conducting equating with nonequivalent groups is to disconfound group and form differences.

The common item nonequivalent groups design is widely used in practice. A major reason for its popularity is that this design requires only one test form to be administered per test date, allowing for equating to be conducted using scores from an operational administration of the test. In addition, with an external set of common items, all items that contribute to an examinee's score (the noncommon items) can be disclosed following the test date. The ability to disclose items is important for some testing programs because some states have mandated disclosure for certain tests.

The administrative flexibility offered by being able to use nonequivalent groups is gained at some cost. Strong statistical assumptions are required to disconfound group and form differences. Although a variety of solutions have been proposed, no statistical procedure can provide completely appropriate adjustments for this design when the examinee groups differ (Petersen, Kolen, and Hoover, in press).

A variety of approaches exist for dealing with the problems associated with this design. One important consideration is that the set of common items be proportionally representative of the total test forms in content and statistical characteristics. That is, the common item set should be constructed to be a "mini version" of the total test forms. Table 3 provides data on a hypothetical test as an illustration of the need for content representativeness. In this example, on average, Group A correctly answers 70% of the Type I items and 80% of the Type II items. If the total test contains half Type I items and half Type II items, then Group A will earn an average score of 75% correct (1/2 [70%] + 1/2 [80%] = 75%) on the whole test, and Group B will earn an average score of 75% correct (1/2 [80%] + 1/2 [70%] = 75%) as well. Thus, the two groups will be at the same level of achievement on the total test. Now assume that two forms of the test are to be equated. What would happen if the common item set contained three-fourths Type I items and one-fourth Type II items? In this case, on average, Group A will correctly answer 72.5% of the common items (3/4 [70%] + 1/4 [80%] = 72.5%) and Group B will correctly answer 77.5% of the common items (3/4 [80%] + 1/4 [70%] = 77.5%). Thus, on this set of common items Group B would appear to be higher achieving than Group A, even though the two groups actually were at the same level on the total test. Thus, a content representative set of common items should be used, (see Klein and Jarjoura, 1985, for an illustration of the need for content representativeness in an actual testing program).

Additional ways to improve equating with nonequivalent groups include: (a) using long sets of common items, which usually allows for better content representativeness; (b) placing common items in approximately the same position in both forms, because item position often affects item difficulty; and (c) using two common item sets that are common to two different forms—"double links"—to provide a consistency check on the equating process and to help keep test forms "on scale." These and other related issues are discussed by Brennan and Kolen (1987), and by Cook and Petersen (1987).

Types of Conversions

In traditional equating methods, score correspondence is established by setting characteristics of the score distributions equal for a specified group of examinees. In *mean equating* the means on the two forms are set equal for a particular group of examinees; that is, the Form 2 scores are

TABLE 3

Average Percent Correct on Two Item Types for Two Groups

Item Type	Group A	Group B
I	70%	80%
II	80%	70%

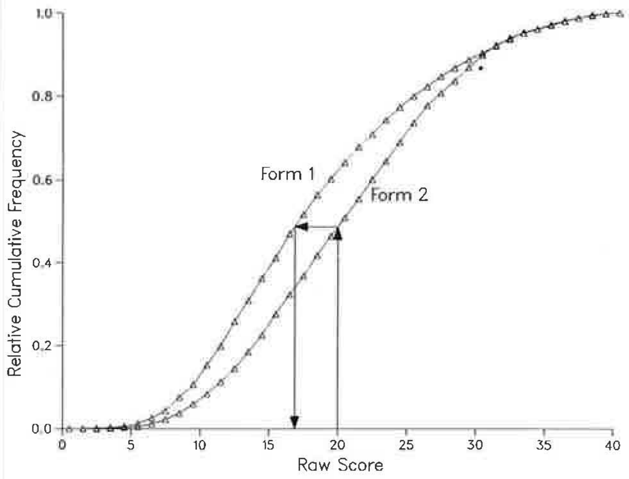


FIGURE 2. *Relative Cumulative Frequencies (Percentile Ranks/100) for Two ACT Assessment Mathematics Forms*

converted so that their mean will equal the mean of the scores on Form 1. In *linear equating* the means and standard deviations on the two forms for a particular group of examinees are set equal. In this method, Form 2 scores are converted so as to have the same mean and standard deviation as scores on Form 1. In *equipercentile equating* the Form 2 distribution is set equal to the Form 1 distribution for a particular group of examinees. Form 2 scores converted using equipercentile equating have approximately the same mean, standard deviation, and distributional shape (skewness, kurtosis, etc.) as do scores on Form 1. Any of these types of conversions can be used with any of the equating designs described previously. Although the equating relationship always is derived for a specified group of examinees, Angoff and Cowell (1986) and Harris and Kolen (1986) indicate that similar conversions can be expected for a wide range of examinee groups when equating alternate forms.

Mean equating. Consider the example described earlier for the single group design in which the mean on Form 1 was 72 and the mean on Form 2 was 77. In mean equating, a Form 1 score of 72 would be set equal to a Form 2 score of 77, so that a score of 77 on Form 2 would be judged to reflect the same level of achievement as a score of 72 on Form 1. In mean equating, the difference observed at the mean (in this example, 77 − 72 = 5 points) is defined to be constant throughout the score scale. So, for example, a Form 2 score of 70 would be considered as indicating the same level of achievement as Form 1 score of 65.

To express mean equating in the form of an equation, first set equal those scores on the two forms that are an equal distance away from their respective means:

$$X_1 - \overline{X}_1 = X_2 - \overline{X}_2,$$

where X_1 is a score on Form 1, X_2 is the corresponding score on Form 2, $\overline{X}_1$ is the mean on Form 1, and $\overline{X}_2$ is the mean on Form 2. Then solve for X_1:

$$X_1 = X_2 - \overline{X}_2 + \overline{X}_1,$$

which is the equation for finding the Form 1 score corresponding to a particular Form 2 score. For the example:

$$X_1 = X_2 - 77 + 72 = X_2 - 5.$$

What is the Form 1 equivalent of a Form 2 score of 70? It can be found by plugging 70 into the preceding equation. Thus, $X_1 = 70 - 5 = 65$, which was indicated earlier.

Linear equating. Mean equating assumes that the difference in difficulty between the forms is constant throughout the entire score range. In many cases the difference in relative difficulty between two forms is better considered to be variable along the score scale. For example, Form 1 might be relatively more difficult than Form 2 for low achieving students than for high achieving students. Linear conversions allow the relative difficulty of the forms to vary along the score scale.

The linear conversion is defined by setting standardized scores on the two forms equal, so that:

$$\frac{X_1 - \overline{X}_1}{S_1} = \frac{X_2 - \overline{X}_2}{S_2},$$

where S_1 is the standard deviation for Form 1 and S_2 is the standard deviation for Form 2. Solving for X_1:

$$X_1 = \frac{S_1}{S_2} X_2 + \left[\overline{X}_1 - \frac{S_1}{S_2} \overline{X}_2 \right] = AX_2 + B,$$

where,

$$A = \frac{S_1}{S_2} \text{ and } B = \overline{X}_1 - \frac{S_1}{S_2} \overline{X}_2.$$

The constant *A* is often referred to as the *slope* of the linear conversion and *B* as the *intercept.*

Suppose that for the previously described mean equating example the standard deviations are $S_1 = 9$ and $S_2 = 10$. The linear conversion is

$$X_1 = \frac{9}{10} X_2 + \left[72 - \frac{9}{10}(77) \right] = .9\, X_2 + 2.7\,.$$

First apply this equation at the mean X_2 value of 77. In this case $X_1 = .9(77) + 2.7 = 72$, which is the mean on Form 1. This result illustrates that linear and mean equating give the same conversion at the mean score. What if X_2 is 67? Then $X_1 = .9(67) + 2.7 = 63$. What if X_2 is 87? Then $X_1 = .9(87) + 2.7 = 81$. Thus, the difference in test form difficulty varies with the score level; in this example the difference for scores around 85 (e.g., 87 − 81 = 6) is greater than the difference for scores around 65 (e.g., 67 − 63 = 4). Recall that this difference would be constant using mean equating.

Equipercentile equating. Equipercentile equating provides for even greater similarity between distributions of equated scores than does linear equating. In equipercentile equating, scores on Form 1 and Form 2 with the same percentile rank for a particular group of examinees are considered to indicate the same level of performance.

The process of equipercentile equating is presented graphically in Figures 2 and 3. The equating shown in these figures was based on an administration of two forms of the 40-item ACT Mathematics test to over 3000 examinees using the random groups design. The first step in this graphical process is to plot the relative cumulative frequency distributions (percentile ranks/100) for each form. Form 1 and Form 2

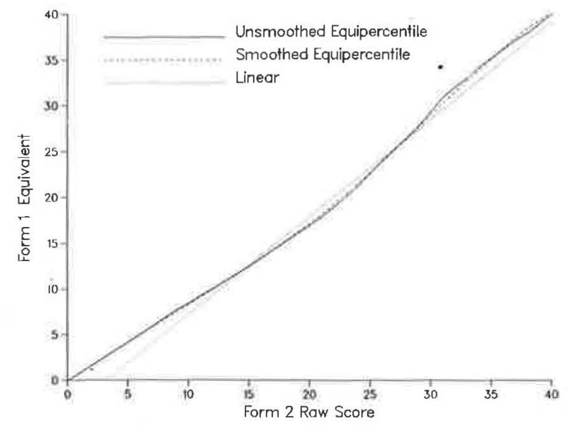


FIGURE 3. *Equating Relationships Between Two ACT Assessment Mathematics Forms*

scores with the same percentile rank are considered to be equivalent. In Figure 2, the arrows indicate that a Form 2 score of 20 has a percentile rank just below 50. The Form 1 score with this same percentile rank is approximately 17. Thus, a Form 2 score of 20 is considered to indicate the same level of achievement as a Form 1 score of 17.

The equipercentile equating score correspondences are shown as a solid line (unsmoothed equipercentile) in Figure 3. For example, Figure 3 indicates that a Form 2 score of 20 corresponds to a Form 1 score of approximately 17. Looking at either Figure 2 or Figure 3, what Form 1 raw score corresponds to a Form 2 score of 25? In Figure 2, a Form 2 score of 25 has a percentile rank of about 70 as does a Form 1 score of about 23. Thus, a Form 1 score of 23 and a Form 2 score of 25 reflect approximately the same level of achievement. This correspondence is also illustrated in Figure 3.

In addition to the unsmoothed equipercentile results, smoothed equipercentile equating results are shown in Figure 3. Smoothing, which can be accomplished by hand or by using analytic methods, is used to reduce sampling error. It is presumed that the bumpiness in the unsmoothed function, such as that which occurs near a Form 2 score of 30 in Figure 3, results from sampling error, and that the relationship would become more regular with larger sample sizes. Studies by Fairbank (1987) and by Kolen (1984) indicate that smoothing has the potential to improve the estimation of the equipercentile relationship. In Figure 3, smoothing had only a minor effect. Smoothing typically has more of an effect with smaller sample sizes. Linear equating results are also shown in Figure 3 as dotted line. The linear equating results are similar to the equipercentile equating in the middle range of scores, but differences occur at the extremes.

Comparison among methods. If the forms to be equated have equal standard deviations, then mean equating and linear equating will produce the same results. If the distributions have the same shape (skewness, kurtosis, etc.), then the linear and equipercentile methods produce the same results. Equipercentile equating typically requires larger sample sizes than does linear or mean equating, and is substantially more complex computationally than the linear or mean methods, especially for the common item nonequivalent groups design.

The Synthetic Group in the Nonequivalent Groups Design

A single group or randomly equivalent groups of examinees are needed to define mean, linear, and equipercentile conversions. In the nonequivalent groups design, however, there are two distinct groups of examinees (i.e., Group A takes Form 1 and Group B takes Form 2), and these two groups typically differ somewhat in achievement level. To conduct equating under this design, Braun and Holland (1982) introduced the concept of the *synthetic group.* The synthetic group is defined as a weighted combination of Groups A and B. The conversion is defined for this synthetic group.

Different definitions of the synthetic group have been suggested (Angoff, 1987; Kolen & Brennan, 1987). One definition involves weighting Groups A and B equally. Another is to weight Groups A and B by sample size. Still another is to define the synthetic group as the group administered the new form. Kolen and Brennan (1987) suggest that, in practice, the particular definition of the synthetic group used has a minimal effect on the resulting conversion but is important to consider for conceptual and practical reasons.

After the synthetic group is defined, the performance of this group on Form 1 and Form 2 needs to be estimated. A major problem here is that no data are collected on Form 2 for Group A or on Form 1 for Group B, yet we need to estimate how these groups would perform on these forms. A variety of solutions have been suggested, each requiring strong statistical assumptions. One of these methods, the Tucker method, assumes, among other things, that the linear regression of Form 1 or Form 2 scores on common item scores is the same for Group A and Group B. Another method, the Levine, an equally reliable method, assumes, among other things, that the correlations between true scores on Form 1, Form 2, and the common items all are 1.00. Each of these methods requires strong statistical assumptions that can be expected to be met, at best, only approximately in practice.

To illustrate how the process of defining and forming the synthetic group might be accomplished, refer again to the data shown in Table 2. To conduct equating using the nonequivalent groups design, it is necessary to estimate how Group A examinees would have performed had they been administered Form 2 and how Group B examinees would have performed had they been administered Form 1. A statistical method, like the Tucker method, would be used to obtain these estimates. Based on the common items, Group B is higher achieving than Group A. Suppose that the Tucker method indicated that Group B scored 4 points higher, on average, on each of the forms than Group A. Then the estimated Group B mean on Form 1 would be 76 (72 + 4 = 76) and the estimated Group A mean on Form 2 would be 73 (77 − 4 = 73). If a synthetic group is formed by equally weighting Group A and Group B, then the Form 1 mean for the synthetic group would be the average of the Group A and Group B means on Form 1, which equals (72 + 76)/2 = 74. Similarly, the Form 2 mean for the synthetic group would be (73 + 77)/2 = 75. Which form is easier, on average, for the synthetic group? Because, in the synthetic group, the mean for Form 1 is one point lower (74 − 75 = −1) than the mean for Form 2, Form 1 is one point more difficult, on average, than Form 2. Under this design, equating using linear methods involves considering standard deviations in addition to means, and equipercentile equating introduces additional complexities. All of these methods require consideration of the composition of the synthetic group.

Equating Error

Different types of equating error influence the interpretation of results from the application of equating methods. Equating designs and equating methods should be chosen to lead to as little equating error as possible, given practical constraints. In some practical circumstances the amount of equating error may be so large that it is better not to even attempt to equate.

Random equating error is present whenever samples from populations of examinees are used to estimate parameters such as means, standard deviations, and percentile ranks. Random error can be reduced by using larger samples of examinees and by the choice of equating design. Random error can be especially troublesome when practical considerations dictate the use of small samples of examinees.

Systematic equating error results from violations of the assumptions and conditions of the particular equating methodology used. In the single group design, failure to control for fatigue and practice effects can be a major source of systematic error. In the random groups design, systematic error will result if the spiraling process is ineffective in achieving group comparability. Systematic equating error is especially problematic in the nonequivalent groups design. Systematic error will result if the assumptions underlying the method (e.g., Tucker or Levine) used are not met. These assumptions can be especially difficult to meet if the groups differ substantially, or if the common items are not representative of the total test form in content and statistical characteristics. In addition, systematic error will likely result if the common items function differently from one administration to another. For example, common items sometimes function differently if their position in the old and new form is not the same. Or, in some professional certification and licensure examinations (e.g., medicine), changes in the body of knowledge can change the difficulty of an item or even the keyed answer. For any equating design, systematic error can result if the new form and old form differ in content, difficulty, and reliability.

When a large number of test forms are involved in the scaling/equating process, both random and systematic error tend to accumulate. Although random error can be quantified fairly readily using the standard error of equating, systematic error is much more dificult to estimate. In conducting equating and in setting up equating plans, it is necessary to attempt to minimize both kinds of error.

Self-Test

1. A scholarship test is administered twice per year and different forms are administered on each test date. Currently, the top 1% of the examinees on each test date earn scholarships, and the test forms are not equated.
   a. If the test forms were equated, would this affect who was awarded a scholarship? Why or why not?
   b. Suppose the top 1% who took the test during the year (rather than at each test date) were awarded scholarships. In this case, could equating affect who passed? Why or why not?
2. Refer to the example in Table 1. If Form 4 were found to be uniformly one point more difficult than Form 3, what scaled score would correspond to a Form 4 raw score of 29?
3. The following data resulted from the administration of two forms of a test using a random groups design:

Form 1	Form 2
$\overline{X}_1 = 30$	$\overline{X}_2 = 32$
$S_1 = 5$	$S_2 = 4$

   a. Using mean equating, what Form 1 score corresponds to a Form 2 score of 36?
   b. Using linear equating, what Form 1 score corresponds to a Form 2 score of 36?
4. Based on the graphs in Figure 2, what Form 2 score corresponds to a Form 1 score of 15 using equipercentile equating?
5. Refer to the data shown in Table 3.
   a. Which group would appear to be higher achieving on a set of common items composed only of Type I items?
   b. Which group would appear to be higher achieving on a set of common items composed only of Type II items?
6. One state passes a law that all items that contribute to an examinee's score on a test must be released to that examinee, on request, following the test date. Assuming that the test is to be secure, which of the equating designs can be used to equate forms of the test? Briefly indicate how equating would be accomplished using this (these) design(s).

Answers to Self-Test

1. a. Because the top 1% of the examinees on a particular test date will be the same regardless of whether or not an equating process is used, equating would not affect who was awarded a scholarship.
   b. It is necessary to consider examinees who were administered two forms as one group in order to identify the top 1% of the examinees during the whole year. If the forms on the two test dates were unequally difficult, then the use of equating could result in scholarships being awarded to different examinees than just using the raw score on the form each examinee happened to be administered. When successful equating is feasible, it generally provides for a more equitable basis for awarding scholarships, because equating adjusts for the differences in the difficulty of test forms.
2. Because Form 4 is one point more difficult than Form 3, a score of 29 on Form 4 would be indicative of the same level of achievement as a raw score of 30 on Form 3. From Table 1, a Form 3 score of 30 corresponds to a Form 2 score of 29, which corresponds to a Form 1 score of 28, which corresponds to a scaled score of 14. Therefore, a Form 4 raw score of 29 converts to a scaled score of 14.
3. a. From Equation 1,

$$X_1 = X_2 - \overline{X}_2 + \overline{X}_1 = X_2 - 32 + 30 = X_2 - 2.$$

   Thus, using mean equating, a Form 2 score of 36 corresponds to a Form 1 score of 36 − 2 = 34.
   b. From Equation 2,

$$X_1 = \frac{5}{4}(36) + 30 - \frac{5}{4}(32) = 35.$$

   Thus, using linear equating, a Form 2 score of 36 corresponds to a Form 1 score of 35.
4. A Form 1 score of 15 has a percentile rank of approximately 39 as does a Form 2 score of approximately 18. Thus, a Form 1 score of 15 indicates approximately the same level of achievement as a Form 2 score of 18.

5. a. Group B would appear to be higher achieving because they correctly answer 80% of the Type I items as compared to 70% for Group A.
   b. Group A would appear to be higher achieving because they correctly answer 80% of the Type II items as compared to 70% for Group B.
6. Because the test is to be secure, items that are going to be used as scored items in subsequent administrations cannot be released to examinees. A common items design with an *external* set of common items would be the easiest design to implement in these circumstances. On a particular administration, each examinee would receive a test form containing the scored items, a set of unscored items that had been administered along with a previous form, and possibly another set of unscored items to be used as a common item section in subsequent equatings. Thus, all items that contribute to an examinee's score would be new items that would never need to be reused. The single group and random groups designs also could be implemented using a special study. For example, using the random groups design a number of forms could be spiraled in a state or states that did not have test disclosure legislation, and these forms then used later in the state with the legislation. In this case, no common items would be needed.

Annotated References

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

Presents a discussion of equating and related issues and standards for conducting equating.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores.* Princeton, NJ: Educational Testing Service. Originally appeared in R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.), 508–600. Washington, DC: American Council on Education, 1971.

A comprehensive treatment of equating as well as norming and score scales that presents major developments through 1971. This chapter contains discussions of issues from both conceptual and statistical perspectives.

Angoff, W. H. (1987). Technical and practical issues in equating: A discussion of four papers. *Applied Psychological Measurement, 11,* 291–306.

An extensive discussion of the Cook and Petersen, Fairbank, Kolen, and Brennan, and Brennan and Kolen papers in a special issue of this journal. Presents a conceptual perspective that is sometimes different from that presented by the authors of the individual papers.

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.

Describes a study that showed that equating relationships for alternate test forms are very similar for different examinee groups.

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.

A detailed mathematical analysis of traditional equating methods. Much of the volume that contains this chapter is also recommended.

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

A number of practical issues are discussed in this paper including equating error, equating adequacy, content specifications changes, cutting score issues, and security breaches.

Cook, L. L., & Eignor, D. R. (in press). Equating using item response theory methods. *Educational measurement: Issues and practice.*

An instructional module that focuses on item response theory equating methods.

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.

Review of the literature on how equating methods are affected by sampling error, sample characteristics, and characteristics of common items. Includes discussions of smoothing in equipercentile equating and the effects of the characteristics of common items on equating results.

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

An extensive study of analytic methods for smoothing in equipercentile equating. The results indicate that some of the methods studied had the potential to improve equating.

Harris, D. J., & Kolen, M. J. (1986). Effect of examinee group on equating relationships. *Applied Psychological Measurement, 10,* 35–43.

An illustrative study showing that equating relationships are similar for different examinee groups when conducting random groups equating of alternate test forms.

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.

A study that shows the problems that arise in nonequivalent groups equating when the content of the common items is not representative of the content of the total test.

Kolen, M. J. (1984). Effectiveness of analytic smoothing in equipercentile equating. *Journal of Educational Statistics, 9,* 25–44.

A description of a study of an analytic smoothing method in equipercentile equating and a study of the effectiveness of the method.

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

Provides a formulation and comparison of the Tucker and Levine equally reliable equating methods that explicitly considers the synthetic group, discusses issues in defining the synthetic group, and provides formulas for decomposing observed differences in means and variances into population differences and form differences.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (in press). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: Macmillan.

Extensive treatment of traditional and item response theory equating methods and designs along with discussions of score scales and norms.

Skaggs, G., & Lissitz, R. W. (1987). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 56,* 495–529.

Reviews much of the recent empirical research on equating methodology.

Teaching Aids Are Available

A set of teaching aids, designed by Michael J. Kolen to complement his ITEMS module, "Traditional Equating Methodology," is available at cost from NCME. These teaching aids contain masters that can be used to create transparencies of figures and tabular material contained in this module. As long as they are available, they can be obtained by sending $5.00 to: **Teaching Aids, ITEMS Module #6, NCME, 1230 17th St., NW, Washington, DC 20036.**

Teaching Aids to Accompany the ITEMS Module

Traditional Equating Methodology

Michael J. Kolen

The American College Testing Program

Created November, 1988

NOTE: The following is a handout which can be used to create transparencies. The author has found these useful in presenting the material contained in the module.

Sources of Equating Error

1. Random - Problematic with small samples

2. Systematic - Due to violations of statistical assumptions

   a. Differences in examinee groups

   b. Differences in test content

   c. Differences in the functioning of common items

Motivation for Equating

1. Student X takes a college admissions test for the second time and earns a higher score than on the first testing. Why?

   a. Her achievement level increased.

   b. Her achievement level is the same, but she already saw the items once and learned the answers to them.

   c. She took an easier set of test questions the second time.

Purpose of Equating

By means of a statistical adjustment of test scores, the purpose of equating is to adjust for differences in difficulty among forms of a test so that the forms can be used interchangeably. After successful equating, examinees can be expected to earn the same score regardless of the test form administered.

NOTES: (1) To be considered as "forms of a test," forms must be developed from the same content and statistical specifications. (Forms built to different content or difficulty specifications cannot be equated.)

(2) Equating is needed because, in practice, it is impossible to construct test forms that are precisely equal in difficulty.

Three Related Problems

1. Equating - Converting scores on alternate forms of a test to the same scale.

2. Vertical Scaling - Placing tests of different difficulties but closely related content on the same scale.

3. Concordance - Placing different tests that are to be used for a particular purpose on the same scale.

Three Ways to Report Scores for Multiple Form Testing Programs

1. Report raw scores for every form, regardless of the difficulty of the form.

   Problems: Examinees administered an easier form are advantaged. Examinees taking a more difficult form are disadvantaged. Trends in examinee ability over time are confounded with test form difficulty.

2. Convert scores such that the examinee distribution (e.g., mean and standard deviation) is always the same.

   Problems: Examinees tested with a lower achieving examinee group will be advantaged. Examinees tested with a higher achieving group will be disadvantaged. Trends in examinee ability cannot be addressed using reported scores.

3. Use equating and report equated scores. Equating, if successful, adjusts for differences in the difficulty of test forms and accounts for differences in examinee groups, so that successfully equated scores are not affected by the problems that beset 1 and 2 above.

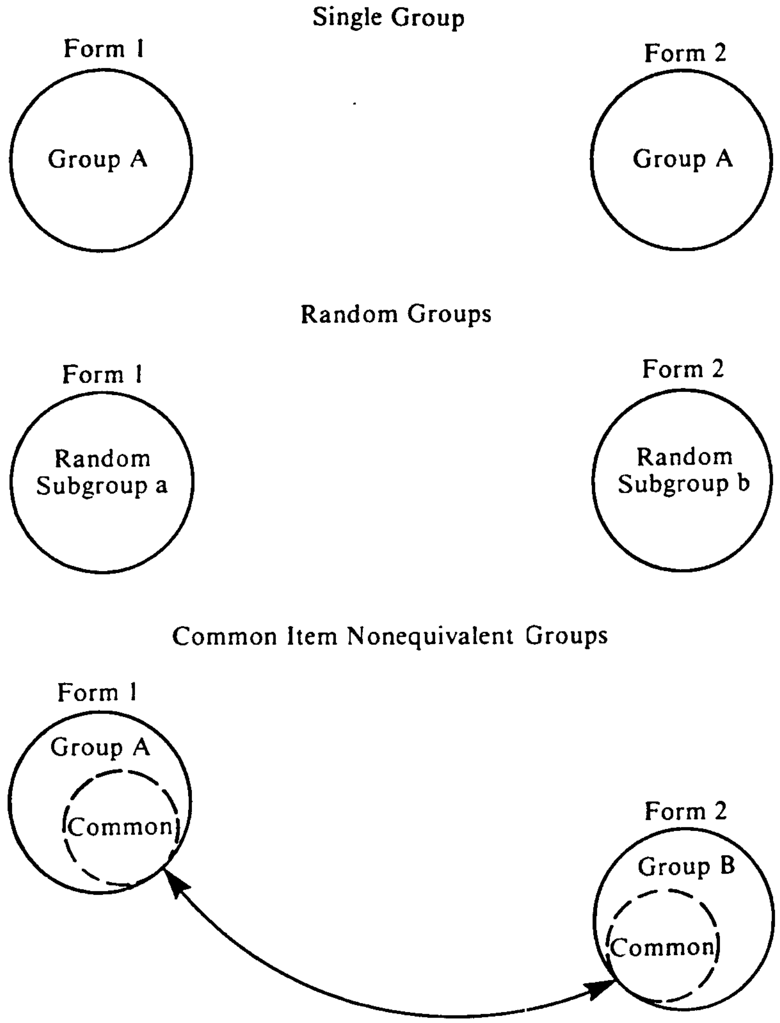
   Problem: It is not always possible to conduct adequate equating.

Illustration of the Scaling/Equating Process

Hypothetical Conversion Tables for Three Test Forms

Form 1		Form 2	Form 1	Form 3	Form 2
Raw	Scaled	Raw	Raw	Raw	Raw
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
30	15	30	29	30	29
29	15	29	28	29	28
28	14	28	27	28	27
27	14	27	26	27	26
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

Three Data Collection Designs



Some Issues in Common Item Equating

Hypothetical Means for Two Forms of a 100-Item Test With 20 Common Items

Group	Form 1	Form 2	Common Items
A	72	--	13 (65%)
B	--	77	15 (75%)

Questions

1. Which group is higher achieving?

2. Which form is easier?

Need for Content Representativeness

Average Percent Correct on Two Item Types for Two Groups

Item Type	Group A	Group B
I	70%	80%
II	80%	70%

Suppose total test contains half Type I and half Type II items. Then

$$\bar{X}_A = .5(70\%) + .5(80\%) = 75\%$$

$$\bar{X}_B = .5(80\%) + .5(70\%) = 75\%$$

Suppose common item set contained 3/4 Type I and 1/4 Type II items. Then

$$\bar{X}_A = .75(70\%) + .25(80\%) = 72.5\%$$

$$\bar{X}_B = .75(80\%) + .25(70\%) = 77.5\%$$

Traditional Equating Methods and Some of Their Characteristics

Equating Method	Form of Function	Statistics the Same for the Two Forms	How Results are Communicated
Mean	Y = X + B	Mean	As a translation constant (B) and a rounding rule
Linear	Y = AX + B	Mean, standard deviation	As a slope (A), intercept (B), and a rounding rule
Equipercentile	Complicated	Mean, standard deviation, distributional shape	Requires a conversion table
IRT	Complicated	None	Requires a conversion table

Mean Equating Example

$\overline{X}_1 = 72$ $\overline{X}_2 = 77$

How can scores on form 2 be transformed to the scale of form 1?

$$X_1 - \overline{X}_1 = X_2 - \overline{X}_2$$

$$X_1 = X_2 - \overline{X}_2 + \overline{X}_1$$

$$X_1 = X_2 - 77 + 72$$

$$X_1 = X_2 - 5$$

What is the form 1 equivalent of a form 2 score of 70?

$$X_1 = 70 - 5 = 65.$$

Linear Equating Example

$\overline{X}_1 = 72 \qquad \overline{X}_2 = 77$

$S_1 = 9 \qquad S_2 = 10$

How can scores from form 2 be transformed to the scale of form 1?

$$X_1 = AX_2 + B$$

$$A = \frac{S_1}{S_2} \qquad B = \overline{X}_1 - A\overline{X}_2$$

For the example

$$A = \frac{9}{10} = .9 \qquad B = 72 - .9(77) = 2.7$$

$$X_1 = .9X_2 + 2.7$$

What is the form 1 equivalent of a form 2 score of 70?

$$X_1 = .9(70) + 2.7 = 65.7$$

Equipercentile Equating

Steps

a. Construct relative cumulative frequency distribution for each form

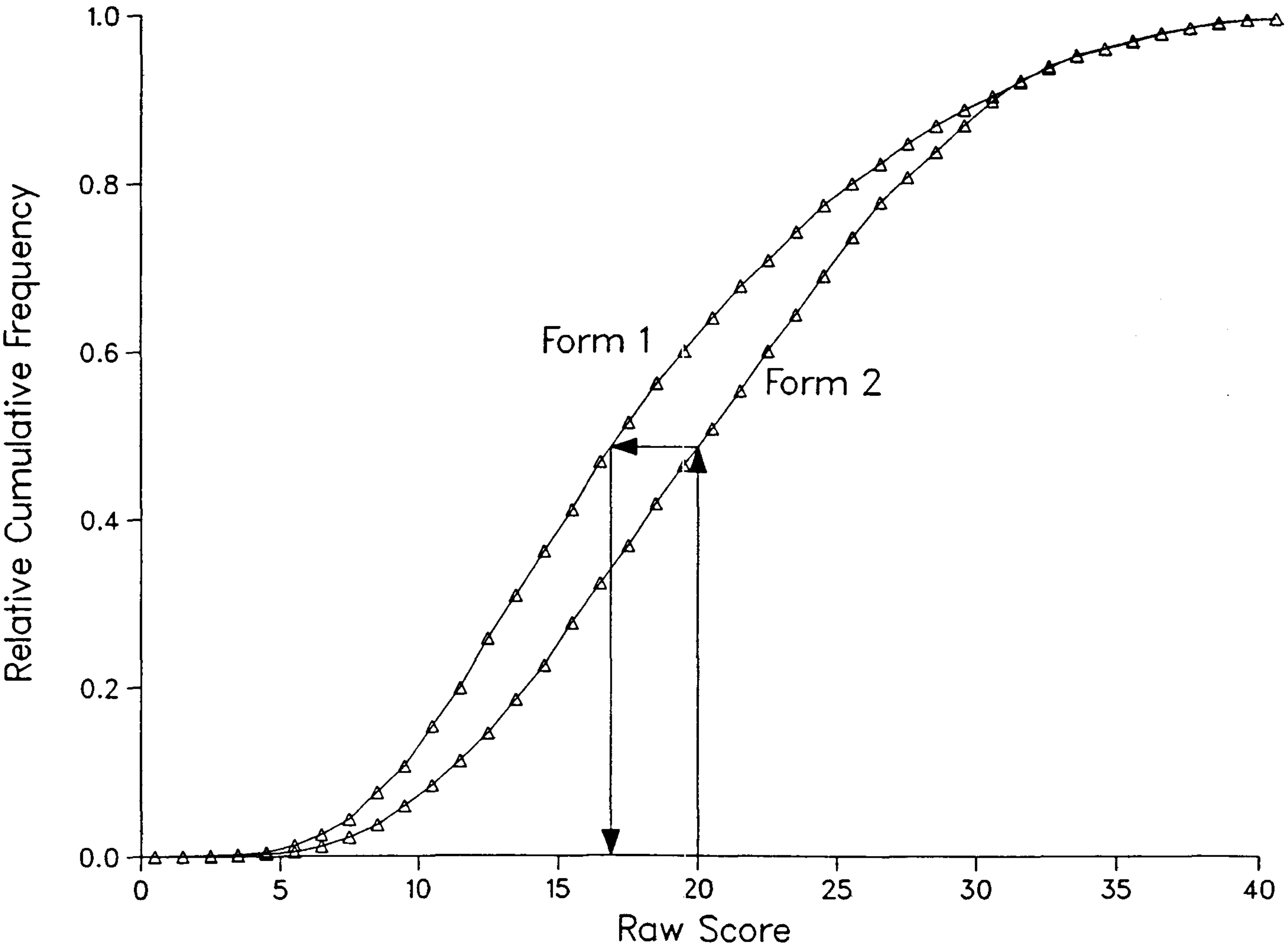
b. Find equipercentile equivalents

Smoothing is used to reduce error in estimating equipercentile equivalents. Smoothing can be done analytically or by hand.

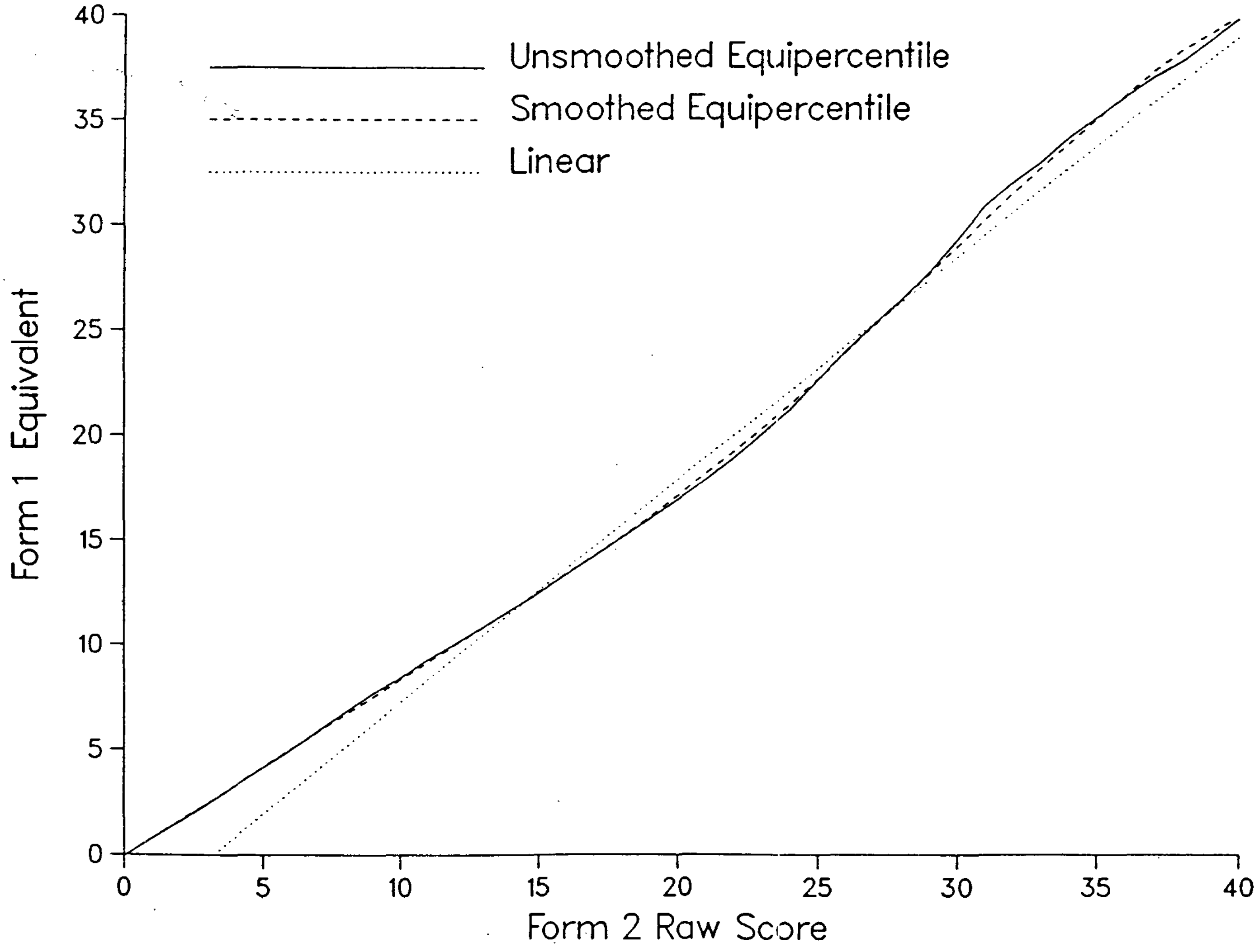
a. Postsmoothing--Smooth the equipercentile equivalents

b. Presmoothing--Smooth the distributions.

The danger in smoothing is that relationships can be distorted. For this reason, more than one method or degree of smoothing should be used, the results should be examined graphically, and statistics (e.g., means, standard deviations) should be examined.



Relative cumulative frequencies (percentile ranks/100) for two ACT Assessment Mathematics forms



Equating relationships between two ACT Assessment Mathematics forms

Some Equating Assumptions

1. <u>Random Groups Design</u>. Groups taking the alternate forms are really random.

2. <u>Nonequivalent Groups Design</u>.

   a. Regressions of total test on common items are the same for the examinees taking the old and new forms (Tucker linear),

   b. True scores are perfectly related on the two forms and common items for both groups (Levine linear),

   c. The same ability is being measured by the two forms and common items for both groups (IRT),

   d. Common items behave in the same way in the old and new forms.

Sources of Equating Error

1. Random - Problematic with small samples

2. Systematic - Due to violations of statistical assumptions

   a. Differences in examinee groups

   b. Differences in test content

   c. Differences in the functioning of common items

Motivation for Equating

1. Student X takes a college admissions test for the second time and earns a higher score than on the first testing. Why?

   a. Her achievement level increased.

   b. Her achievement level is the same, but she already saw the items once and learned the answers to them.

   c. She took an easier set of test questions the second time.

Purpose of Equating

By means of a statistical adjustment of test scores, the purpose of equating is to adjust for differences in difficulty among forms of a test so that the forms can be used interchangeably. After successful equating, examinees can be expected to earn the same score regardless of the test form administered.

NOTES: (1) To be considered as "forms of a test," forms must be developed from the same content and statistical specifications. (Forms built to different content or difficulty specifications cannot be equated.)

(2) Equating is needed because, in practice, it is impossible to construct test forms that are precisely equal in difficulty.

Three Related Problems

1. Equating - Converting scores on alternate forms of a test to the same scale.

2. Vertical Scaling - Placing tests of different difficulties but closely related content on the same scale.

3. Concordance - Placing different tests that are to be used for a particular purpose on the same scale.

Three Ways to Report Scores for Multiple Form Testing Programs

1. Report raw scores for every form, regardless of the difficulty of the form.

   Problems: Examinees administered an easier form are advantaged. Examinees taking a more difficult form are disadvantaged. Trends in examinee ability over time are confounded with test form difficulty.

2. Convert scores such that the examinee distribution (e.g., mean and standard deviation) is always the same.

   Problems: Examinees tested with a lower achieving examinee group will be advantaged. Examinees tested with a higher achieving group will be disadvantaged. Trends in examinee ability cannot be addressed using reported scores.

3. Use equating and report equated scores. Equating, if successful, adjusts for differences in the difficulty of test forms and accounts for differences in examinee groups, so that successfully equated scores are not affected by the problems that beset 1 and 2 above.

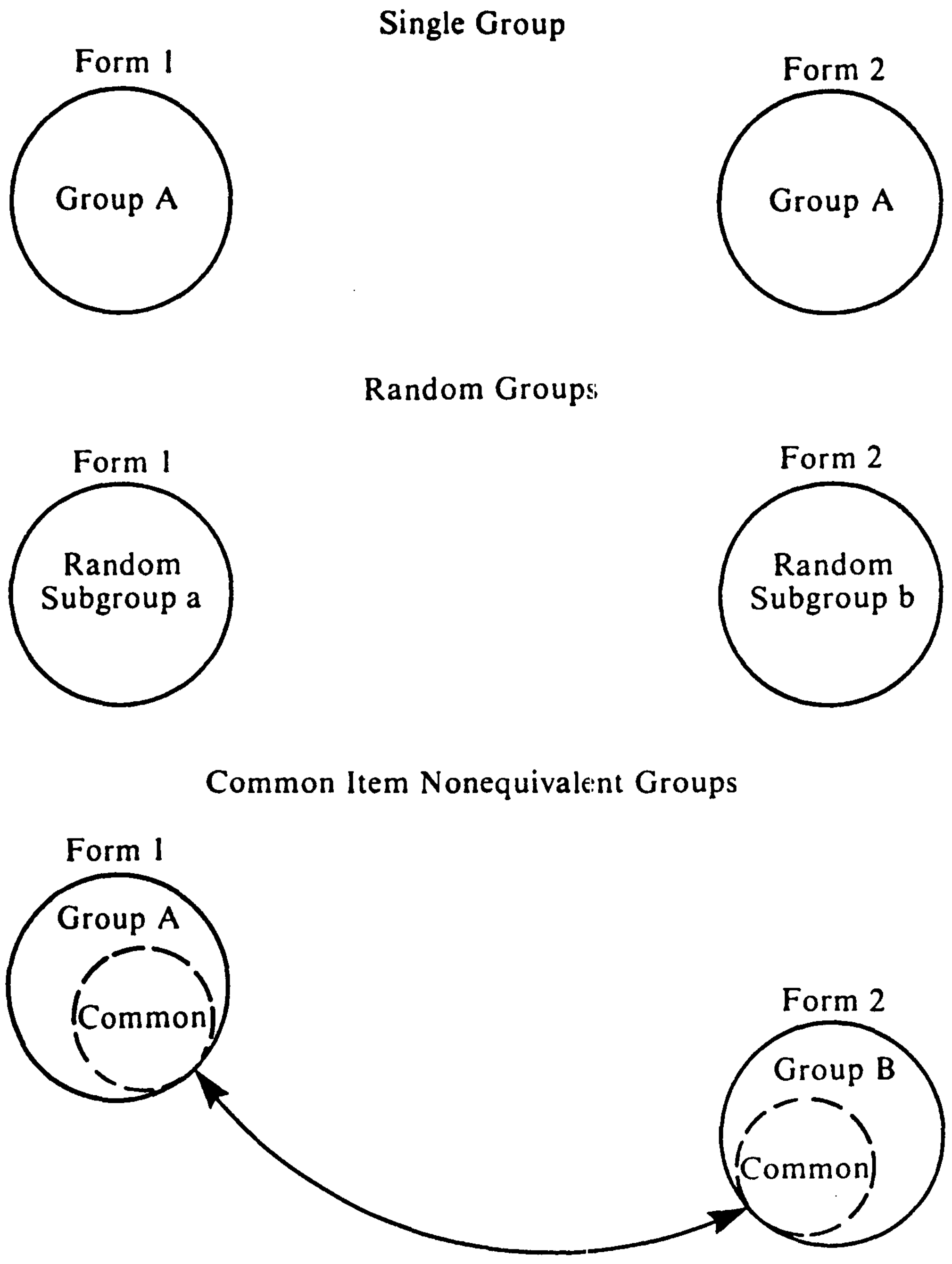
   Problem: It is not always possible to conduct adequate equating.

Illustration of the Scaling/Equating Process

Hypothetical Conversion Tables for Three Test Forms

Form 1		Form 2	Form 1	Form 3	Form 2
Raw	Scaled	Raw	Raw	Raw	Raw
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
30	15	30	29	30	29
29	15	29	28	29	28
28	14	28	27	28	27
27	14	27	26	27	26
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

Three Data Collection Designs



Some Issues in Common Item Equating

Hypothetical Means for Two Forms of a 100-Item Test With 20 Common Items

Group	Form 1	Form 2	Common Items
A	72	--	13 (65%)
B	--	77	15 (75%)

Questions

1. Which group is higher achieving?

2. Which form is easier?

Need for Content Representativeness

Average Percent Correct on Two Item Types for Two Groups

Item Type	Group A	Group B
I	70%	80%
II	80%	70%

Suppose total test contains half Type I and half Type II items. Then

$$\overline{X}_A = .5(70\%) + .5(80\%) = 75\%$$

$$\overline{X}_B = .5(80\%) + .5(70\%) = 75\%$$

Suppose common item set contained 3/4 Type I and 1/4 Type II items. Then

$$\overline{X}_A = .75(70\%) + .25(80\%) = 72.5\%$$

$$\overline{X}_B = .75(80\%) + .25(70\%) = 77.5\%$$

Traditional Equating Methods and Some of Their Characteristics

Equating Method	Form of Function	Statistics the Same for the Two Forms	How Results are Communicated
Mean	Y = X + B	Mean	As a translation constant (B) and a rounding rule
Linear	Y = AX + B	Mean, standard deviation	As a slope (A), intercept (B), and a rounding rule
Equipercentile	Complicated	Mean, standard deviation, distributional shape	Requires a conversion table
IRT	Complicated	None	Requires a conversion table

Mean Equating Example

$\overline{X}_1 = 72$ $\overline{X}_2 = 77$

How can scores on form 2 be transformed to the scale of form 1?

$$X_1 - \overline{X}_1 = X_2 - \overline{X}_2$$

$$X_1 = X_2 - \overline{X}_2 + \overline{X}_1$$

$$X_1 = X_2 - 77 + 72$$

$$X_1 = X_2 - 5$$

What is the form 1 equivalent of a form 2 score of 70?

$$X_1 = 70 - 5 = 65.$$

Linear Equating Example

$\overline{X}_1 = 72 \qquad \overline{X}_2 = 77$

$S_1 = 9 \qquad S_2 = 10$

How can scores from form 2 be transformed to the scale of form 1?

$$X_1 = AX_2 + B$$

$$A = \frac{S_1}{S_2} \qquad B = \overline{X}_1 - A\overline{X}_2$$

For the example

$$A = \frac{9}{10} = .9 \qquad B = 72 - .9(77) = 2.7$$

$$X_1 = .9X_2 + 2.7$$

What is the form 1 equivalent of a form 2 score of 70?

$$X_1 = .9(70) + 2.7 = 65.7$$

Equipercentile Equating

Steps

a. Construct relative cumulative frequency distribution for each form

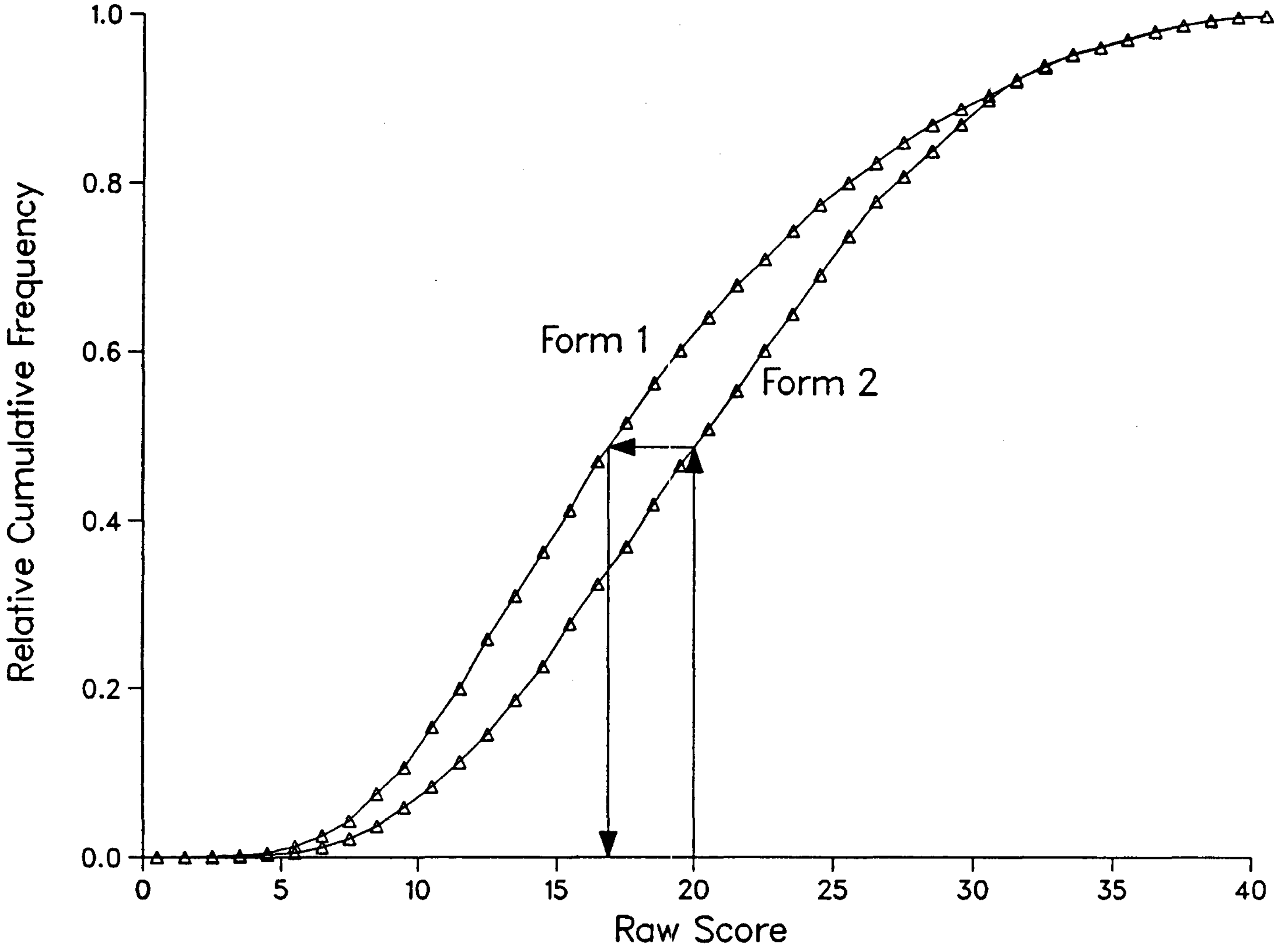
b. Find equipercentile equivalents

Smoothing is used to reduce error in estimating equipercentile equivalents. Smoothing can be done analytically or by hand.

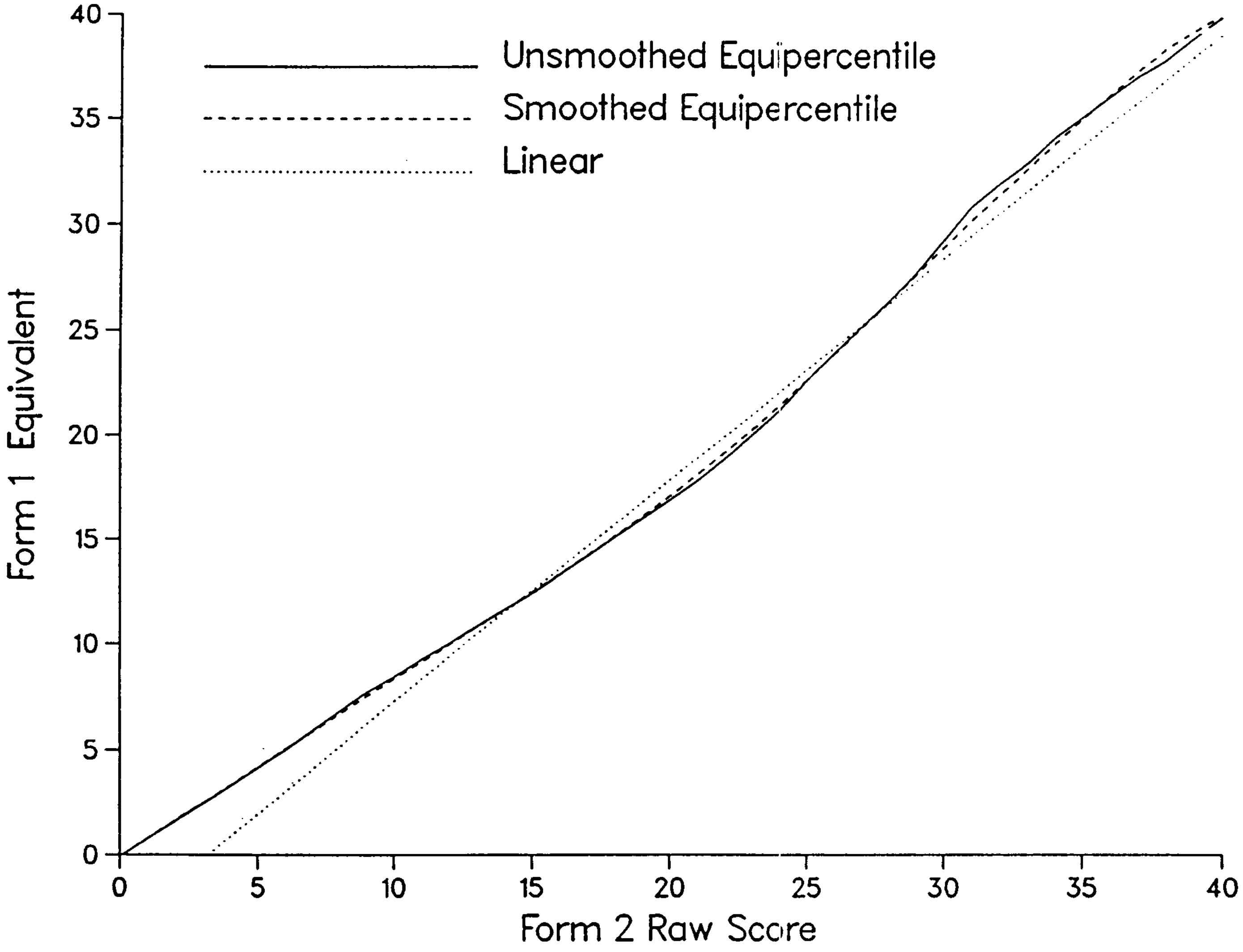
a. Postsmoothing--Smooth the equipercentile equivalents

b. Presmoothing--Smooth the distributions.

The danger in smoothing is that relationships can be distorted. For this reason, more than one method or degree of smoothing should be used, the results should be examined graphically, and statistics (e.g., means, standard deviations) should be examined.



Relative cumulative frequencies (percentile ranks/100)
for two ACT Assessment Mathematics forms



Equating relationships between two ACT Assessment Mathematics forms

Some Equating Assumptions

1. Random Groups Design. Groups taking the alternate forms are really random.

2. Nonequivalent Groups Design.

   a. Regressions of total test on common items are the same for the examinees taking the old and new forms (Tucker linear),

   b. True scores are perfectly related on the two forms and common items for both groups (Levine linear),

   c. The same ability is being measured by the two forms and common items for both groups (IRT),

   d. Common items behave in the same way in the old and new forms.