An NCME Instructional Module on

IRT Equating Methods

Linda L. Cook and Daniel R. Eignor
Educational Testing Service

The purpose of this instructional module is to provide the basis for understanding the process of score equating through the use of item response theory (IRT). A context is provided for addressing the merits of IRT equating methods. The mechanics of IRT equating and the need to place parameter estimates from separate calibration runs on the same scale are discussed. Some procedures for placing parameter estimates on a common scale are presented. In addition, IRT true-score equating is discussed in some detail. A discussion of the practical advantages derived from IRT equating is offered at the end of the module.

The primary purpose for implementing testing under standardized conditions is to provide a means of measuring or evaluating a group of examinees' skills that is as fair and objective as possible. Test scores are often used for such purposes as the assessment of the abilities and/or skills of individuals who are competing for college admissions or seeking professional certification. This evaluation of test scores (when used in conjunction with other information) may lead to a decision to exclude a candidate from some academic program or to limit the ability of an examinee to practice the profession of his/her choice. In addition, important funding decisions and decisions regarding school curricula, etc., are sometimes made on the basis of the standardized test scores of groups of students.

Because of the importance often placed on the results of standardized testing, it is essential that the resulting test scores provide a fair and equitable evaluation of the skills or abilities that the test is purported to measure. Indeed, the provision of a fair and equitable means of psychological or educational assessment is one of the major reasons for the existence of standardized testing.

A number of situations may exist that result in a nonstandard testing experience and, consequently, an unfair evaluation of an examinee's skills or abilities. These situations include those related to the physical testing environment as well as the administration of the test and the actual content and psychometric characteristics of the test.

Linda L. Cook is Executive Director of the Admissions and Guidance Programs Area in the College Board Division of Educational Testing Service (ETS), Princeton, NJ 08541.

Daniel R. Eignor, also at ETS, is Principal Measurement Specialist and Associate Director of Statistical Analysis in the College Board Division. (The authors' names appear in alphabetical order.)

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.

One important way that large-scale testing programs attempt to ensure testing under standardized conditions is to try, as much as possible, to protect the security of their examinations. That is, every possible effort is made to ensure that some examinees will not be advantaged by preknowledge of the questions presented in an examination. In order to ensure that one or more examinees will not encounter test questions they may have previously seen, most testing programs develop many forms or versions of the same test. For example, the College Board Admissions Testing Program introduces a new form of the Scholastic Aptitude Test (SAT) at every national test date.

Although use of multiple forms of the same test ensures fairness to examinees in that no examinee is advantaged because he or she had preknowledge of the test questions, use of multiple forms of a test raises new fairness and equity issues, issues that are related to the similarity of the characteristics of the different test forms. Because it is virtually impossible for individuals constructing the multiple forms of the test to develop them in such a way that they are completely similar in reliability, difficulty, etc., the possibility exists that examinees may be advantaged simply because they took an easier or more reliable form of a test.

Test-equating procedures—both classical test theory procedures (i.e., traditional procedures; see Kolen, 1988) and item response theory (IRT) procedures—were developed in order to provide comparable scores on multiple forms of the same test, consequently avoiding some of the possible inequities that could occur if one examinee took a more difficult form of a test than that taken by another examinee.

It is important to emphasize the word *difficulty* when one refers to equating procedures; that is, it is important to understand that equating methods, both classical and IRT methods, were designed to take into account minor differences in form-to-form difficulty. These equating procedures were never intended to take into account large differences in form-to-form difficulty, reliability, or test content.

Unfortunately, many practitioners believe that IRT methods, because they fall into a category of equating methods referred to as true-score methods (as compared to observed-score methods), provide alternatives to classical test equating procedures in situations that require the equating of different tests (not alternate forms of the same test) that vary markedly in content, difficulty, or reliability. No equating procedure will function adequately in such situations.

According to Angoff (1984), there are four restrictions or requirements that must be met in order to say that two test forms have been equated: (a) the test forms to be equated should measure the same ability; (b) the resulting raw score to scale score conversion should be independent of the data used in deriving it and should be applicable in all similar situations; (c) scores on the two test forms should, after equating, be interchangeable in use; and (d) the equating should be symmetric, or the same, regardless of which test form is designated as the base. Angoff further points out that equating and the issue of unique conversions can really only be addressed when the test forms are parallel.

A number of comments can be made about these restrictions or requirements. First, while one of the restrictions requires that the two test forms measure the same ability, it is not specified that the ability be unidimensional. However, unidimensionality, or a close approximation to it, is a requirement of current IRT equating methods (see Hambleton, 1989). This means that somewhat tighter restrictions on the nature of the test data must be met for IRT equating applications. Second, the generalizability of the equating conversions from the data used for developing them may fall short in practice when classical procedures are used any time the groups taking the test forms to be equated are not random samples from the same population. The use of nonrandom samples is typical of practical equating situations. Third, the criterion of interchangeability of scores really only holds when the test forms are equally reliable.

IRT equating involves a) selecting a design, b) placing parameter estimates on a common scale, and c) equating test scores.

Lord's definition of equating (Lord, 1977) reflects in greater detail Angoff's third requirement, called the equity requirement:

> Transformed scores y^* and raw scores x can be called "equated" if and only if it is a matter of indifference to each examinee whether he is to take test X or test Y (p. 128).

Under this definition, (a) test forms measuring different abilities cannot be equated (comparable to Angoff's first restriction), (b) observed scores (those scores actually obtained by test takers) on unequally reliable test forms cannot be formally equated (this would violate Angoff's third restriction), and (c) observed scores on test forms of varying difficulty cannot be equated. Lord (1977) states,

> If tests X and Y are of different difficulties, the relation between their true scores is necessarily nonlinear, because of floor and ceiling effects. If two tests have a non-linear relation, it is implausible that they should be equally reliable for all subgroups of examinees. This leads to the awkward conclusion that, strictly speaking, observed scores on tests of different difficulty cannot be equated (p. 128).

While the above suggests that, in theory, observed-score equating is not possible except when test forms are of exactly equal difficulty, test equating is routinely carried out for test forms that are not strictly parallel. There will be problems in practice, however, any time the test forms to be equated are not close to the same level of difficulty and observed scores are used. For this reason, and also to satisfy Angoff's second restriction (that the conversions should be independent of the group used to obtain them), IRT equating methods, which are true-score based, have appeal for the solution of equating problems.

IRT Equating Process

Basic Principle of IRT Equating

Hambleton (1989) provided an introductory treatment of item response theory models, assumptions, and properties. One of the basic properties of item response theory discussed by Hambleton is the following: If the data used for the equating fit the assumptions of an IRT model and good estimates of the parameters for the model are obtained, it is possible to develop an estimate of an examinee's ability that is independent of the set of items (i.e., test form) to which the examinee responds. Consequently, it does not matter if an examinee takes an easy or a hard form of a test; the examinee's ability estimate developed from either test form will be identical, within measurement error, provided the parameter estimates for both forms have been placed on the same scale. Therefore, the differences in difficulty of the test forms taken by the examinee are no longer a problem. The examinee would receive the same ability estimate regardless of the particular form he or she takes. Further, if one is willing to use the underlying IRT ability scale for score reporting purposes, IRT reduces the need for equating test forms to the process of placing item parameter estimates on the same scale.

Unfortunately, large-scale testing programs are not always able to report scores using the IRT ability scale or metric. These programs usually continue to report scaled scores on the scale that was initially chosen for the particular test of interest, even though IRT may have subsequently been used for equating purposes. Fortunately, because ability estimates can be mathematically related to specific true scores on each of the two test forms, IRT equating of these true scores can be used and traditional scaled scores can be reported.

Kolen (1988) provided an introductory treatment of classical linear and equipercentile equating procedures. Similar to the equipercentile equating procedure discussed by Kolen, IRT equating methods can model either a linear or curvilinear relationship between scores on two forms of a test. If the relationship between scores on two forms of a test is linear, IRT methods will easily model this relationship. On the other hand, if the relationship between scores on the two forms is curvilinear, IRT equating methods can also effectively model this relationship.

Mechanics of IRT Equating

IRT equating can be viewed as a three-step process. Assuming that a suitable IRT model has been chosen (that is, a model that fits the data), the first step is to select a data collection design for equating. The second step involves placing item parameter estimates from separate calibration (parameter estimation) runs on the same scale. For some equating designs, however, this second step is not necessary. Also, when using certain computer programs, such as LOGIST (Wingersky, 1983), it is often the case that parameter estimation can be accomplished in a single calibration run. If a single calibration run is used, the item parameter estimates for the two test forms to be equated will automatically be on the same scale.

The third step involves using the relationship between abilities and true scores on the two test forms to be equated to establish the raw-to-scale relationship for the new form (the form requiring equating). As mentioned earlier, if a testing program can report scores on the ability metric, the equating has been accomplished through the calibration and subsequent placing of parameter estimates on the same scale, and the third step is not necessary. However, because many testing programs report scores on some established scale, such as the familiar 200 to 800 scale used by the SAT, the third step becomes necessary.

Step One: Selecting a Design

There are essentially three data collection designs used in IRT equating. These designs are the same as those used with traditional linear and equipercentile equating methods; they are the (a) single group, (b) random groups, and (c) anchor test designs. In the single group design, the same group takes both test forms to be equated. In the random groups design, two randomly selected groups, of equivalent ability, take different forms of the test. In the third design, two groups of examinees take different forms of a test; each form contains a common set of items (internal anchor) or a common anchor test (external anchor) is given with the forms. The groups in this third design do not need to be randomly equivalent, and usually they are not.

If classical equating methods are employed, the common items linking the two test forms in the third design are used to adjust for ability differences in the two groups. If IRT equating methods are used, the common items are used to place item parameter estimates on the same scale. The common items are used for item parameter scaling purposes regardless of whether the calibration is carried out in a single LOGIST run or in separate calibration runs.

A question often arises regarding how large a sample is necessary to carry out IRT or, for that matter, conventional equating. The response to this question is more straightforward for IRT equating applications than for classical equating applications. An adequate sample size for a particular IRT equating design depends on the number of examinees required to provide stable estimates of the parameters of the particular IRT model used to characterize the data. Larger sample sizes are required for the more complex models. Sample sizes currently used to calibrate data when the three-parameter logistic model is used to equate the SAT typically range from 2,500 to 3,000 examinees per item, although as few as 1,800 to 2,000 examinees per item have been viewed as acceptable. When compared to classical equating procedures used for the SAT, 3,000 examinees taking each item is larger than the sample needed if a linear equating method were to be used for the test and smaller than desirable if a curvilinear equating method, such as one of the equipercentile methods discussed by Angoff (1984), were to be used.

The properties of the common items used in an anchor test design have been considered and discussed for both classical and item response theory equating methods. (See Cook and Petersen, 1987, for a review of recent studies of the properties of common items.) It is safe to say that regardless of whether one chooses to use classical or IRT equating methods, choice of a common item set is very important. For both types of procedures, it is important for the common items to mirror, in content and statistical properties, the properties of the tests to be equated. In addition, a good rule of thumb is that the common-item anchor should be at least as long as 20% of the total test length.

The choice of design—single group, random group, or anchor test—is often dictated by the practical constraints of the testing program. For example, it is very unlikely that a test such as the SAT would be equated routinely using a single group design. This would require motivating a group of candidates to sit and take two forms of the test, most likely at the same test date. Even if candidates were motivated by financial remuneration to sit for two tests or were provided with some other incentive, it is unlikely that this type of data collection design would prove practical over the long run.

The second design, a random groups design, is unappealing for many large-scale testing programs because it involves giving an old form (i.e., previously administered form) of the test as well as the new form (the form requiring equating) at a regular test date. Many testing programs, including the Admissions Testing Program that administers the SAT, avoid this practice as much as possible because of the desire to ensure that every examinee takes a new form of the test, thus minimizing the possibility that a candidate may be advantaged due to preknowledge of the items on a test.

The design most typically used for the practical reasons just described is the anchor test design. However, this design is probably the most difficult to execute technically. The quality of the equating carried out using an anchor test design depends on the similarity of the groups taking the new and old forms of the test, the parallelism of the two tests to be equated, and the quality of the anchor test.

Step Two: Placing Parameter Estimates on a Common Scale

As a means of clarifying the need for a separate step to place parameter estimates on the same scale, consider the following situation: Suppose the same set of items is given to two groups of examinees, and the item parameters are estimated twice, once in each group. Suppose, too, that the model of choice is the two-parameter logistic (2PL) model. Because the item characteristic curves are supposedly independent of the groups used to derive them, the expectation would be that the two sets of item parameter estimates would be identical except for sampling error. However, this is not so. For the 2PL model, Stocking and Lord (1983) pointed out that the expression for the item characteristic curve is a function of $a_i(\theta_a - b_i)$, where the a_i, θ_a, and b_i are the item discrimination, ability, and item difficulty parameters of the model. However, for this model the origin and unit of measurement of the ability (and difficulty) metric are undetermined. This can best be seen by developing this algebraically. Consider the expression for the 2PL model discussed by Harris (1989):

$$P_i(\theta_a) = \frac{e^{Da_i(\theta_a - b_i)}}{1 + e^{Da_i(\theta_a - b_i)}}. \qquad (1)$$

Clearly $P_i(\theta_a)$ is a function of $a_i(\theta_a - b_i)$. Suppose now that θ_a is transformed by a linear transformation to give θ_a^*:

$$\theta_a^* = A\theta_a + B, \qquad (2)$$

where A is the slope of the linear transformation and B the intercept. b_i is similarly transformed to give b_i^*:

$$b_i^* = Ab_i + B, \qquad (3)$$

and a_i is transformed by multiplying by the reciprocal of the A parameter of the linear transformation:

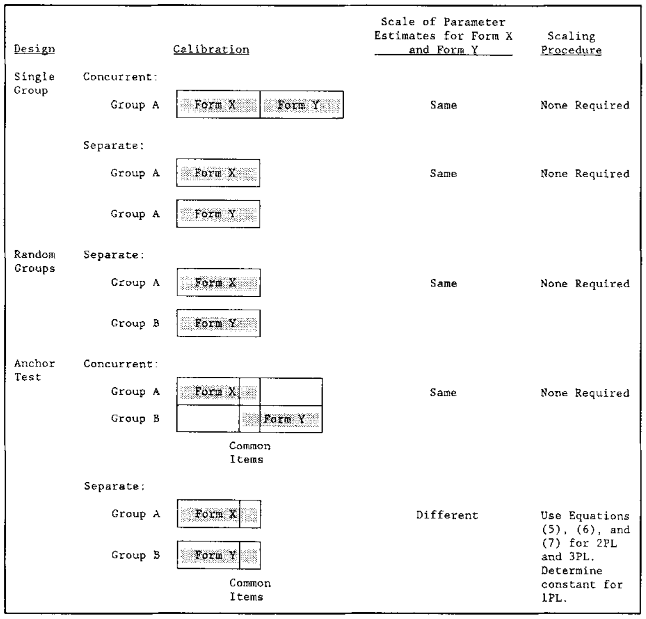
$$a_i^* = \frac{1}{A} a_i. \qquad (4)$$

How is the new item-response function $P_i(\theta_a^*)$ related to $P_i(\theta_a)$? $P_i(\theta_a^*)$ is a function of $a^*{}_i(\theta_a^* - b_i^*)$. However,

$$\begin{aligned} a_i^*(\theta_a^* - b_i^*) &= \frac{1}{A} a_i(A\theta_a + B - [Ab_i + B]) \\ &= \frac{1}{A} a_i(A\theta_a + B - Ab_i - B) \\ &= a_i \frac{1}{A}(A(\theta_a - b_i)) \\ &= a_i(\theta_a - b_i) \end{aligned}$$

Clearly $P_i(\theta_a^*) = P_i(\theta_a)$.

The above will be true for any linear transformation of the parameters—the scale of the ability metric is undetermined. There clearly is an indeterminacy in the model. The same



Design	Calibration	Scale of Parameter Estimates for Form X and Form Y	Scaling Procedure
Single Group	Concurrent: Group A: Form X Form Y	Same	None Required
	Separate: Group A: Form X Group A: Form Y	Same	None Required
Random Groups	Separate: Group A: Form X Group B: Form Y	Same	None Required
Anchor Test	Concurrent: Group A: Form X Group B: Form Y Common Items	Same	None Required
	Separate: Group A: Form X Group B: Form Y Common Items	Different	Use Equations (5), (6), and (7) for 2PL and 3PL. Determine constant for 1PL.

FIGURE 1. *Summary of calibration designs and resulting scale properties of IRT parameter estimates*

indeterminacy also exists for the three-parameter logistic (3PL) model.

In the LOGIST computer program, this problem of indeterminacy is solved for the 2PL and 3PL models by establishing an origin and unit of measurement for the ability (and difficulty) metric based on the ability of the group of examinees used to calibrate the items. Typically, the mean of the ability estimates ($\hat{\theta}$s) is set at zero, and the standard deviation of the $\hat{\theta}$s is set to one. Thus, as part of the calibration procedure, parameter estimates are placed on a scale that is defined by the mean and standard deviation of the ability distribution of the group responding to the items. Other computer programs fix the scale in different ways. An important point to note is that the relationship between scales derived from two different LOGIST calibrations will always be linear because they differ only in origin and unit of measurement. It also should be noted that when LOGIST parameter estimates are obtained in separate calibration runs, and the groups responding to the set of items in the two calibration runs are identical in ability, the item parameter estimates will be on the same scale and a transformation is not necessary. This would occur if a random groups equating design were used. Likewise, for the single group design, item parameter estimates will be on the same scale if two calibration runs were done because the same group will have responded to the set of items in each calibration run.

Because parameter estimates from two separate 2PL or 3PL calibration runs differ only in origin and unit of measurement (if they differ at all; they will not differ in the single group or random groups design if LOGIST is used), the transformation required to place two sets of parameter estimates on the same scale is a simple linear transformation. Transformation methods that have been developed for use with the 2PL and 3PL models attempt to estimate the parameters of this linear transformation (i.e., the slope and intercept) in a variety of ways. Stocking and Lord (1983) discuss a variety of these procedures.

The simplest 2PL or 3PL transformation method attempts to determine the slope and intercept parameters so that the transformed mean and standard deviation of the distribution of estimated item difficulties from the second calibration are equal to the mean and standard deviation of the estimated item difficulties from the first calibration. This is analogous to the linear score equating procedure discussed by Kolen (1988), except that item difficulties rather than observed test scores are equated. The linear parameters of the transformation are determined using only the items in common to the two calibration runs. An example of how to determine this kind of transformation follows.

Suppose that two separate LOGIST calibrations using the 3PL model have 20 items in common. The mean and standard deviation of the item difficulty estimates for the 20 items in the first calibration are .76 and 1.06 while, in the second calibration, they are .43 and .97, respectively. The parameters of the linear transformation can be determined as follows:

$$\frac{b_1 - .76}{1.06} = \frac{b_2 - .43}{.97}$$

or

$$b_1 = \frac{1.06}{.97}(b_2 - .43) + .76$$

$$b_1 = 1.09\, b_2 + .29$$

Hence, the slope, or A parameter of the linear relationship, is 1.09 and the intercept, or B parameter, is .29.

Once the A and B parameters of the transformation are determined, they are applied to all of the parameter estimates in the calibration run to be transformed. So, if $\hat{b}_i$ is the estimate of item difficulty obtained from the calibration of item i, and $\hat{b}_i^*$ denotes the transformed item difficulty estimate, then $\hat{b}_i^*$ is determined as follows:

$$\hat{b}_i^* = A\hat{b}_i + b, \tag{5}$$

where A and B are the parameters of the linear transformation. The same linear parameters are used to transform the item discrimination ($\hat{a}_i$) and ability estimates ($\hat{\theta}_a$), i.e.,

$$\hat{a}_i^* = \frac{\hat{a}_i}{A} \tag{6}$$

$$\hat{\theta}_a^* = A\hat{\theta}_a + B. \tag{7}$$

Because the pseudoguessing parameter estimate for item i ($\hat{c}_i$) in the 3PL model is read from the probability metric or Y-axis of the plotted item characteristic curve (rather than the ability metric or X-axis), no transformation is required.

As mentioned earlier, simple 2PL or 3PL transformation methods are based on determining the linear relationship between estimates of item difficulties obtained from two calibration runs. Theoretically, ability estimates or estimates of item discrimination parameters could also be used to establish the transformation, and in some situations are actually used. In practice, however, item difficulty estimates are most frequently used because they are the most stable of any of the parameter estimates.

Now, suppose that instead of the 2PL or 3PL model, an individual decides it is appropriate to use the one-parameter logistic (1PL), or Rasch, model. Harris (1989) has provided a discussion of the Rasch model. For this model, the expression for the item characteristic curve is a function of ($\theta_a - b_i$). Here, only the origin of the ability (and difficulty) metric is undetermined. In the LOGIST program, this problem of indeterminancy is solved by establishing the origin of the ability (and difficulty) metric based on the ability of the group of examinees used to calibrate the items, i.e., the mean of the $\hat{\theta}$s is set to zero. Other computer programs fix the scale in different ways. Regardless of computer program, however, the relationship between the scales derived from two different 1PL calibrations (using the same computer program) will differ by only a constant.

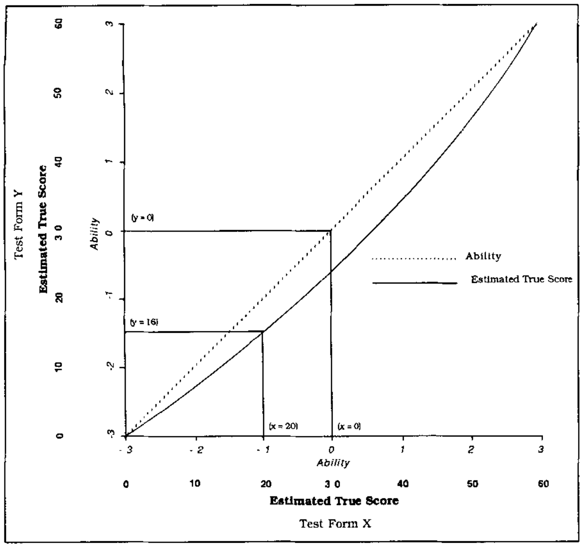


FIGURE 2. *Relationship between ability estimates and estimated true scores for Form X and Form Y*

Suppose that two separate LOGIST calibrations using the 1PL model have 20 items in common and that the mean of the item difficulty estimates for the 20 items in the first calibration is .76 and, in the second calibration, is .43. The constant of adjustment can then be determined as follows:

$$b_1 - .76 = b_2 - .43$$

or

$$b_1 = b_2 + .33$$

To adjust all item difficulty parameter estimates from the second calibration to be on the scale of the first calibration, simply add a constant, .33. Exactly the same constant would be added to the ability estimates from the second calibration to put them on the scale of the first.

The information provided in Figure 1 summarizes the previously described item calibration designs and the scale properties of the IRT parameter estimates obtained from these designs. Several points are worth noting: (a) if it is possible to calibrate item responses for two test forms simultaneously, parameter estimates will automatically be on the same scale; (b) for the single group and random groups data collection designs, parameter estimates for the two tests (identified as test form *X* and test form *Y* in Figure 1) will be on the same scale regardless of whether they are obtained in separate or simultaneous calibrations; and (c) for anchor test designs, a separate scaling step is required to place parameter estimates on the same scale if the parameter estimates are obtained in separate calibration runs.

Step Three: Equating Test Scores

Once parameter estimates for two forms of a test have been placed on the same scale, the ability estimate obtained for an individual will be the same within measurement error regardless of which form of the test he or she actually took. Therefore, if one can report ability estimates to examinees as their test scores, the equating is completed.

If a testing program is unable to report ability estimates to examinees, it is possible to translate any value of θ to corresponding estimated true scores on the two forms and use these estimated true scores as equated scores. If test form *X* and test form *Y* are both measures of the same ability θ, then estimated true scores on the two forms are related to θ by their test characteristic functions:

$$\hat{T}_x = \sum_{i=1}^{n_x} \hat{P}_i(\theta), \tag{8}$$

$$\hat{T}_y = \sum_{j=1}^{n_y} \hat{P}_j(\theta), \tag{9}$$

where

$\hat{T}_x$ = Test form *X* estimated true score,
$\hat{T}_y$ = Test form *Y* estimated true score,
and $\hat{P}_i$ (θ) and $\hat{P}_j(\theta)$ are the estimated item-response functions for items *i* (in test form *X*) and *j* (in test form *Y*), respectively.

Equations (8) and (9) can be used to transform any θ (not just $\hat{\theta}$s actually obtained from the administration of the test forms) to an estimated true score on the respective test forms. Because of this and because the item parameter estimates used in (8) and (9) are independent of the groups taking the tests, the conversion, or relationship between the true scores on the two forms, can also be said to be independent of the groups used to obtain it. Further, it should be noted that while the $\hat{\theta}$ s for individuals, estimated separately for two test forms, should be the same once transformed, the relationship between the estimated true scores will certainly be nonlinear if the forms differ in difficulty.

The plot shown in Figure 2 illustrates the relationship between ability estimates and estimated true scores for two test forms (test form *X* and test form *Y*) after the two forms have been equated using IRT methods. Differences in difficulty between the two forms are reflected in the curvilinear relationship between estimated true scores for forms *X* and *Y*. It can be seen from the plot that form *X* is the easier form of the test, with higher estimated true scores considered equivalent to lower estimated true scores on form *Y*. However, as can be seen in Figure 2, the relationship between ability estimates obtained on the two forms is linear, with slope of 1.0 and intercept of 0.0, indicating that an examinee's estimate of ability is the same regardless of the test form he or she was administered.

A practical problem with IRT true-score equating should be mentioned at this point. This problem has to do with the fact that the scores to be equated on two test forms are observed scores (the scores the candidates actually obtained) and not the estimated true scores derived from Equations (8) and (9). Studies suggest, however, that applying the true-score equating results to the actual observed-score data does not lead to unreasonable results (see Lord & Wingersky, 1983), although there is no theoretical justification for doing this. In most 2PL and 3PL model IRT equating applications, this is exactly what is done.

In practice, a fairly simple procedure is followed to provide reported scores based on an IRT equating procedure. First, abilities (i.e., θs) are transformed to estimated true scores on the two forms using Equations (8) and (9). Once the relationship between estimated true scores on the new and old forms of the test is determined, one simple further step is involved to establish the relationship between estimated true scores on the new test form and reported scores on the scale used by the particular testing program. This step involves taking the observed-score-to-reported-score transformation for the old form in the equating and using this transformation with the old form's estimated true scores. The relationship between new and old form estimated true scores is then used to derive an estimated true-score-to-reported-score transformation for the new form.

For illustrative purposes, Table 1 provides data to be used in an example of the steps that would be taken to determine scaled or reported scores for a new form of a test, on a scale that

Table 1

Three-Parameter Model IRT True-Score Equating Table for New Form's Estimated True Scores From 40 to 50

New Form's Estimated True Score	Theta	Old Form's Estimated True Score	Rounded Scaled Score
50	2.0089	48.7149	700
49	1.9170	47.6325	690
48	1.8300	46.5456	680
47	1.7467	45.4555	670
46	1.6665	44.3632	660
45	1.5887	43.2689	660
44	1.5129	42.1738	650
43	1.4385	41.0773	640
42	1.3652	39.9790	630
41	1.2928	38.8788	620
40	1.2210	37.7761	610

extends from 200 to 800, rounded to units of ten for reporting purposes, for estimated true scores between 40 and 50.

Focusing on, say, a new form's estimated true score of 45, one can see that this corresponds to an ability or theta value of 1.5887 (determined using Equation (8) in reverse form), which *in turn corresponds to an old form's estimated true score of* 43.2689 (determined using equation (9)). Then, 43.2689 is treated as if it were an observed score on the old form, and the old form's observed-score-to-reported-score transformation is then applied to 43.2689 to yield a reported score of 660. An individual who received an estimated true score of 45 on the new form and an individual who received an estimated true score of 43.2689 on the old form would both receive a rounded scaled score of 660.

It should be noted that the old form of the test is more difficult than the new form; for a given ability level, it yields a lower estimated true score than the new form. To demonstrate that the old form is more difficult, we can see what an estimated true score of, say, 44 would correspond to in scaled score terms for the new and old forms. If the old form is more difficult, a 44 should result in a higher scaled score. For the new form, the scaled score can be read directly from the table, 650. For the old form, an estimated true score of 44 is closest to the entry 44.3632 in Table 1, which rounds to 44 and which corresponds to a scaled score of 660. The old form does provide a higher scaled score for an estimated true score of 44.

Although in the example we have been working with estimated true scores, the true-score equating transformation shown in Table 1 would, in practice, be applied to the observed scores that the individuals obtained on the new form in order to create scaled or reported scores.

Practical Advantages of IRT Equating

In a previous section of this module, two theoretical advantages were offered for using IRT equating methods: (a) IRT equating may be the best method to use when tests of differing difficulties are given to nonrandom groups of examinees who differ in ability, and (b) IRT equating, because of the properties of IRT models, provides conversions that are independent of the group or groups used to obtain them.

Recent research on the application of IRT to the equating process has also brought to light a number of possible practical advantages that might be gained through the use of IRT equating. These advantages include the following:

1. IRT equating offers better equating than that offered by classical methods at the upper ends of score scales where important decisions are often made. As mentioned in the previous section, it is possible to equate estimated true scores for all values of θ. With classical methods, reasonable equating can take place for only those scores actually obtained by the test takers supplying data.
2. IRT equating affords greater flexibility in choosing previous forms of a test for equating purposes. Once a number of previous test forms have item parameter estimates placed on a common scale, it is possible to equate a new test form (once its parameter estimates have been placed on the same scale) to any or all of the old test forms.
3. Reequating is easier should it be decided to not score an item after the test is administered. Presently, when classical equating methods are used, if for some reason it is decided not to score an item, the shortened test must be rescored and reequated. If IRT equating of estimated true scores is used, the estimated true scores for the shortened test can be obtained by simply summing over the item characteristic curves for the remaining items in the test.
4. IRT equating offers the possibility of item-level preequating, or deriving the relationship between the test forms before they are administered operationally. This is possible when item-level pretest data are available and can be calibrated, and item parameter estimates can be placed on a common scale. The use of IRT for item-level preequating offers a unique contribution that cannot be obtained using classical methods and data collection designs.

Summary

In this module, the theoretical justifications and practical advantages of IRT true-score test equating procedures have been discussed. When tests differ in difficulty, or are unequally reliable, and are administered to samples that are not random samples from the same population, the formal requirements for equating outlined in this module are best met when true-score or IRT-based equating procedures are used. As pointed out, however, the practical difficulty in using IRT-based, true-score equating is that the scores to be equated on two tests are not true scores, but observed scores. Results from research studies suggest that applying the true-score equating results to the observed-score data does not lead to unreasonable results, although there is no theoretical justification for doing this. The justification is purely a practical one. For further discussions of this and other issues involving IRT equating, the reader is referred to Lord (1980), Cook and Eignor (1983), Skaggs and Lissitz (1986), Petersen, Kolen, and Hoover (1989), and Cook and Eignor (1989).

Self-Test

In the problems that follow, the numbers of items in the test forms are small to simplify numerical calculations. The numbers of items are not representative of the usual numbers of items in tests calibrated using LOGIST and equated using IRT equating methods.

1. Two 6-item forms of a test, Form *A* and Form *B*, have three items in common. Each form was calibrated

separately using LOGIST and the 3PL model. The table below presents the parameter estimates for the unique and common items in Forms *A* and *B*. As the first step in equating these two forms, use the procedure described in the text to place Form *A* item parameter estimates on the scale of the Form *B* item parameter estimates. Place the transformed item parameter estimates in the table. Use the following formula for computing standard deviations in doing necessary calculations:

$$S_x = \frac{1}{n}\sqrt{n\,\Sigma X^2 - (\Sigma X)^2}$$

			Form *A* Scale			Form *B* Scale		
		Item(i)	a_i	b_i	c_i	a_i	b_i	c_i
Form *A*		1	1.04	−.68	.22			
		2	1.38	.96	.16			
		3	.90	−.14	.08			
	Form *B*	4	1.62	1.22	.16	1.08	1.23	.17
		5	1.28	.06	.12	.85	−.51	.12
		6	1.84	−.32	.18	1.23	−1.08	.18
		7				.96	.24	.16
		8				1.60	1.32	.10
		9				.56	−.84	.07

2. Use the following table, containing values of $P_i(\theta)$ for each of the Form *A* and Form *B* items at θ values of −1, 0, and +1, to construct an (abbreviated) true-score equating table for θ values of −1, 0, and +1.

	$P_i(\theta)$ at θ =		
Item(i)	−1	0	+1
1	.7459	.8986	.9655
2	.2047	.3380	.6323
3	.4956	.7199	.8746
4	.1836	.2485	.7536
5	.4104	.7152	.9108
6	.6242	.9224	.9896
7	.2581	.4988	.8115
8	.1016	.1242	.3657
9	.4997	.7116	.8625

As an example of how the $P_i(\theta)$s were calculated, consider item 9 at θ = 1. The general formula for the 3PL model is

$$P_i(\theta) = c_i + \frac{(1 - c_i)}{1 + e^{-1.7a_i(\theta - b_i)}}$$

This becomes

$$P_9(1) = .07 + \frac{(1 - .07)}{1 + e^{-1.7(.56)(1-(-.84))}}$$

$$= .07 + \frac{.93}{1 + e^{-1.75168}}$$

$$= .07 + \frac{.93}{1 + .17348}$$

$$= .8625$$

See Harris (1989) for further examples.

3. Which form is more difficult—Form *A* or Form *B*?

4. Treat the (abbreviated) true-score equating table in Problem 2 as if it were applicable to observed scores. Suppose that on Form *B*, an observed score of 3 receives a scaled or reported score of 50, an observed score of 4 receives a scaled or reported score of 70, and an observed score of 5 receives a scaled or reported score of 90. Use linear interpolation (see example below) to determine what reported score a candidate should receive who got a Form *A* observed score of 4.

Example of linear interpolation

What reported score corresponds to a score of 3.2207 on Form *B*?

$$1\left[\begin{array}{ccc} & 4 & 70 \\ .2207 & \left[\begin{array}{cc} 3.2207 & \\ & \\ 3 & 50 \end{array}\right] & x \end{array}\right] 20$$

$$\frac{x}{20} = \frac{.2207}{1}$$

$$x = 4.414$$

So, the reported score corresponding to 3.2207 is 54.414 (i.e., 50 + 4.414).

Answers to Self-Test

1. (a) Using the *b* parameter estimates for the three common items, calculate the means and standard deviations:

Form *A*

$$\bar{b}_A = \frac{(1.22) + (.06) + (-.32)}{3}$$

$$\bar{b}_A = \frac{.96}{3}$$

$$\bar{b}_A = .32$$

$$S_{b_A} = \frac{1}{3}\sqrt{3[(1.22)^2 + (.06)^2 + (-.32)^2] - (.96)^2}$$

$$S_{b_A} = .6550$$

Form *B*

$$\bar{b}_B = \frac{(1.23) + (-.51) + (-1.08)}{3}$$

$$\bar{b}_B = \frac{-.36}{3}$$

$$\bar{b}_B = -.12$$

$$S_{b_B} = \frac{1}{3}\sqrt{3[(1.23)^2 + (-.51)^2 + (-1.08)^2] - (-.36)^2}$$

$$S_{b_B} = .9825$$

(b) Setting standard deviates equal:

$$\frac{b_B - (-.12)}{.9825} = \frac{b_A - (.32)}{.6550}$$

$$b_B = \frac{.9825}{.6550}(b_A) - .12 - \frac{.9825}{.6550}(-.32)$$

$$b_B = 1.50\, b_A - .60$$

The slope (*A*) of the linear transformation is 1.50, and the intercept (*B*) is −.60.

(c) Using the formulas

$$\hat{b}_{\mathrm{i}}^{*} = A\hat{b}_{\mathrm{i}} + B,$$

$$\hat{a}_{\mathrm{i}}^{*} = \frac{\hat{a}_{\mathrm{i}}}{A}, \qquad \text{and}$$

$$\hat{c}_{\mathrm{i}}^{*} = \hat{c}_{\mathrm{i}},$$

where $A = 1.50$ and $B = -.60$, the transformed item parameter estimates for the three unique Form *A* items are the following:

for *Item 1*: $\hat{a}_1^* = \dfrac{1.04}{1.50} = .69$

$$\hat{b}_1^* = 1.5\,(-.68) - .60 = -1.62$$

$$\hat{c}_1^* = .22$$

for *Item 2*: $\hat{a}_2^* = \dfrac{1.38}{1.50} = .92$

$$\hat{b}_2^* = 1.5\,(.96) - .60 = .84$$

$$\hat{c}_2^* = .16$$

for *Item 3*: $\hat{a}_3^* = \dfrac{.90}{1.50} = .60$

$$\hat{b}_3^* = 1.5\,(-.14) - .60 = -.81$$

$$\hat{c}_3^* = .08$$

The completed table of item parameter estimates would then look as follows:

		Form *A* Scale			Form *B* Scale		
	Item(i)	a_i	b_i	c_i	a_i	b_i	c_i
—Form *A*—	1	1.04	−.68	.22	.69	−1.62	.22
	2	1.38	.96	.16	.92	.84	.16
	3	.90	−.14	.08	.60	−.81	.08
—Form *B*—	4	1.62	1.22	.16	1.08	1.23	.17
	5	1.28	.06	.12	.85	−.51	.12
	6	1.84	−.32	.18	1.23	−1.08	.18
	7				.96	.24	.16
	8				1.60	1.32	.10
	9				.56	−.84	.07

2. Test Form *A* is made up of items 1 through 6. At θ = −1,

$$\begin{aligned}\hat{T}_{x_A} &= \sum_{i=1}^{6} \hat{P}_i(-1)\\ &= (.7459) + (.2047) + (.4956)\\ &\quad + (.1836) + (.4104) + (.6242)\\ &= 2.6644\end{aligned}$$

At θ = 0,

$$\begin{aligned}\hat{T}_{x_A} &= \sum_{i=1}^{6} \hat{P}_i(0)\\ &= (.8986) + (.3380) + (.7199)\\ &\quad + (.2485) + (.7152) + (.9224)\\ &= 3.8426\end{aligned}$$

At θ = 1,

$$\hat{T}_{x_A} = \sum_{i=1}^{6} \hat{P}_i(1) = 5.1264$$

Test Form *B* is made up of items 4 through 9. At θ = −1,

$$\begin{aligned}\hat{T}_{x_B} &= \sum_{i=4}^{9} \hat{P}_i(-1)\\ &= (.1836) + (.4104) + (.6242)\\ &\quad + (.2581) + (.1016) + (.4997)\\ &= 2.0776\end{aligned}$$

At θ = 0,

$$\hat{T}_{x_B} = \sum_{i=4}^{9} \hat{P}_i(0) = 3.2207$$

At θ = 1,

$$\hat{T}_{x_B} = \sum_{i=4}^{9} \hat{P}_i(1) = 4.6937$$

The (abbreviated) true-score equating table would then be as follows:

Form *A* Estimated True Score	θ	Form *B* Estimated True Score
2.6644	−1	2.0776
3.8426	0	3.2207
5.1264	1	4.6937

3. Form *B* is more difficult. For a given θ, it yields a lower estimated true score than the estimated true score for Form *A*.

4. The example, repeated here, was the first step in the necessary calculations. Remember that we are treating the true-score equating table as if it were applicable to observed scores.

$$1\left[\begin{array}{ccc} & 4 \qquad\quad 70 & \\ .2207 & \left[\begin{array}{cc} 3.2207 & \\ & \\ 3 & 50 \end{array}\right] & x \end{array}\right]20$$

$$\frac{x}{20} = \frac{.2207}{1}$$

$$x = 4.414$$

So the reported score corresponding to 3.2207 on Form *B* is 54.414 (i.e., 50 + 4.414).

The second step involves:

$$1\left[\begin{array}{ccc} & 5 \qquad\quad 90 & \\ .6937 & \left[\begin{array}{cc} 4.6937 & \\ & \\ 4 & 70 \end{array}\right] & x \end{array}\right]20$$

$$\frac{x}{20} = \frac{.6937}{1}$$

$$x = 13.874$$

So, the reported score corresponding to 4.6937 on Form *B* is 83.874 (i.e., 70 + 13.874).

A table can now be constructed as follows:

Form *A* Score	θ	Form *B* Score	Reported Score
2.6644	−1	2.0776	
		3.0	50
3.8426	0	3.2207	54.414
		4.0	70
5.1264	1	4.6937	83.874
		5.0	90

Finally, to find what reported score an observed score of 4 on Form *A* would receive, use linear interpolation again, but this time with Form *A* scores:

$$1.2838\left[\begin{array}{ll} & 5.1264 \quad 83.874 \\ .1574 & \left[\begin{array}{ll} 4.0 & \\ & \\ 3.8426 & 54.414 \end{array}\right] x \end{array}\right] 29.46$$

$$\frac{x}{29.46} = \frac{.1574}{1.2838}$$

$$x = 3.6119$$

So a 4 on Form *A* would receive a reported score of 58.0259 (i.e., 54.414 + 3.6119) or, rounded to the nearest integer, 58.

Note that this provides another indication that Form *B* is more difficult than Form *A*. An observed score of 4 on Form *A* receives a reported score of 58 while on Form *B* the same observed score receives a higher scaled score, 70. For the same observed score, an individual should receive a higher scaled score on the more difficult form.

Annotated References

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

A comprehensive treatment of classical score equating methods is provided as well as procedures for norming and defining score scales.

Cook, L. L., & Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 175–195). Vancouver, British Columbia: Educational Research Institute of British Columbia.

This introductory chapter on IRT equating includes a discussion of IRT equating concepts and related equating issues.

Cook, L. L., & Eignor, D. R. (1989). Using item response theory in test score equating. *International Journal of Educational Research, 13,* 161–173.

This article provides a review of IRT equating research carried out in the middle part of the 1980s, with research studies organized around four areas: (a) population invariance of equating results, (b) properties of linking items used in anchor test equatings, (c) use of IRT in the vertical scaling of tests, and (d) robustness of IRT equating to violations of underlying IRT model assumptions.

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.

This article is focused on a discussion of how classical and IRT equating methods are affected by (a) sampling error, (b) sample characteristics, and (c) characteristics of anchor test items.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 147–200). New York: Macmillan.

This chapter provides an introduction to the models, estimation methods, and applications of item response theory.

Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice, 8*(1), 35–41.

In this instructional module, the three commonly used logistic item response theory models are discussed and compared.

Kolen, M. J. (1988). Traditional equating methodology. *Educational Measurement: Issues and Practice,* 7(4), 29–36.

This instructional module is focused on classical linear and equipercentile score equating methods.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14,* 117–138.

This article appeared in a special edition of the *Journal of Educational Measurement* devoted to item response theory. The article is focused on a number of applications of IRT, one of which is test score equating.

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

This is the definitive textbook on item response theory and IRT applications. The book contains a chapter on the use of IRT for score equating purposes.

Lord, F. M., & Wingersky, M. S. (1983). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8,* 453–461.

In this article, IRT true-score equating results are compared to results using traditional equipercentile equating of observed scores. For the data studied, the two methods yield very similar results.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York: MacMillan.

An extensive treatment is provided of classical and item response theory equating methods and designs along with discussions of score scales and norms. The chapter describes developments that have taken place since the Angoff chapter was published.

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 56,* 495–529.

Most of the recent research on IRT equating methodology is reviewed in this article.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–210.

This article discusses a number of procedures for transforming parameter estimates from separate calibration runs to a common scale.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), *Application of item response theory* (pp. 45–56). Vancouver, British Columbia: Educational Research Institute of British Columbia.

In this chapter, the reader is introduced to the LOGIST computer program, which is a program used frequently to estimate the parameters for the three-parameter logistic item response model.