ED 208 023 TM 810 713

AUTHOR Cook, Linda L.; Eignor, Daniel R.
TITLE Score Equating and Item Response Theory: Some Practical Considerations.
PUB DATE Apr 81
NOTE 48p.; Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Equated Scores; *Latent Trait Theory; Mathematical Models; Test Construction; Testing Problems
IDENTIFIERS *One Parameter Model; Rasch Model; *Three Parameter Model

ABSTRACT

The purposes of this paper are five-fold to discuss: (1) when item response theory (IRT) equating methods should provide better results than traditional methods; (2) which IRT model, the three-parameter logistic or the one-parameter logistic (Rasch), is the most reasonable to use; (3) what unique contributions IRT methods can offer the equating process; (4) what work has been done that relates to the confidence that can be placed in the IRT equating results; and (5) what unresolved issues exist in the application of IRT to equating. Several issues are discussed to provide a background: formal definitions and requirements of equating; the basic principle of IRT equating; procedures for linking parameter estimates and deriving estimated true and observed score equatings using IRT; the practical advantages to be gained from using IRT equating; and the important distinction between test development and test analysis activities. (Author/BW)

# Score Equating and Item Response Theory:

# Some Practical Considerations[1]

Linda L. Cook[2]

Daniel R. Eignor

Educational Testing Service

[1] A paper presented at the annual meeting of AERA-NCME, Los Angeles, 1981.

[2] The authors names appear in alphabetical order.

# Score Equating and Item Response Theory:
# Some Practical Considerations

Linda L. Cook
Daniel R. Eignor
Educational Testing Service

## Introduction

Large scale testing programs are often involved in either of two situations that necessitate a consideration of the process of equating. In the first situation, a test has been constructed to measure a particular attribute, aptitude, or ability at some defined level of proficiency, and for a variety of reasons, most of them related to test security, multiple forms of the test are necessary. As well defined as a set of content and statistical specifications for a test may be, it is usually impossible to construct multiple forms of the test at exactly the same difficulty level. Since students taking different test forms are usually either competing with each other for certain desired outcomes or being judged as masters or non-masters of the test content vis a vis a cut-off point, it is critical that a method of equating or rendering comparable the scores, or the cut-off points, on multiple forms of a test be considered. When the forms to be equated test content at the same difficulty level, the process has been referred to in the literature and in practice as horizontal equating.

In the second situation, the testing program is interested in establishing a single scale that allows measurements to be compared for various levels of a defined attribute, aptitude, or ability; there may or may not be multiple forms of the test at the same level. For instance, many of the commercially marketed test batteries have tests developed for various grade levels (for example, third, fifth, and seventh grade). Because aggregate scores are often compared across levels (e.g., for program evaluation purposes), it is critical that scores obtained on the various levels of the test be equated, i.e., placed on a common underlying scale. This sort of equating, referred to as vertical equating, is designed to convert to one single scale the scores on multiple forms of a test each designed to measure a different level of the same attribute.

It should be noted that the intended product of both horizontal and vertical equating is obtained scores on multiple test forms that are on the same scale. In the case of horizontal equating, the forms to be equated have been constructed to be identical in difficulty level but differ for unintended reasons, while in vertical equating situations, the forms to be equated have been intentionally constructed to differ, often substantially, in both content and difficulty level. As Slinde and Linn (1977) point out, "It is no surprise that the problem of vertical equating is substantially more difficult and conceptually hazardous than that of horizontal equating."

A commonly accepted way of viewing equating is that scores on two different forms of a test may be considered equivalent if their frequency distributions for a particular group of examinees are identical. This type of equating, referred to as equipercentile equating (see Angoff, 1971), can be accomplished by setting equal raw scores on two forms of a test that have the same percentile rank for the group of examinees. Such a process leads to a consideration of the extent to which the test forms being equated differ in difficulty and the effect this has on the shape of the raw score distributions when the same group of examinees takes both test forms. If the test forms differ considerably in difficulty, the frequency distributions of the raw scores on the two forms will differ considerably in shape. If the distributions of raw scores on the two forms are forced to have the same shape (by equipercentile equating), then the raw score scale on one of the forms must be stretched and condensed to the extent that all moments of the distribution are transformed and the resulting relationship between raw scores on the two forms will be curvilinear. If, however, the tests are very similar in level of difficulty, the shapes of the two raw score distributions should differ only in the first two moments when administered to the same group of examinees. To effect a change in only the first two moments, thereby bringing the raw score distributions into coincidence, a linear transformation may be used. The equating is done by setting equal the standard deviates for scores on the two test forms, resulting in an equation which expresses the linear relationship between the raw scores on the two forms. Of course, equipercentile equating used in this situation will also result in a linear relationship between raw scores on the two test forms, i.e., equipercentile methods applied to two raw score distributions that differ only in their first and second moments will transform only these moments. Evident from this discussion is that equiperentile methods should be used for most vertical equating situations (i.e., the raw score distributions on the two forms differ in more than the first and second moments), whereas linear or equipercentile methods may be appropriate for horizontal applications. Jaeger (1981) has offered some procedures for choosing between linear and equipercentile methods in horizontal equating situations.

It should be noted that, in the above discussion, the same group of examinees were considered to have taken both test forms, thereby controlling for the possible differences in ability of the groups involved in the equating process. In reality, it is usually not the case that the same group takes both forms. Usually different groups or samples of examinees of potentially differing abilities take test forms of varying degrees of difficulty. A common item block or anchor test is administered as a portion of, or along with, each form as a measure of the difference in ability between the two groups. It is this situation of differences in test difficulties "contaminated" by differences in examinee abilities that has profound implications for the use of traditional equating methods, particularly when forms of quite different difficulties are given to groups or samples that are quite disparate in ability (the usual vertical equating situation). Slinde and Linn (1977) have discussed in some detail the use of traditional methods in vertical equating situations and the inherent problems.

The interest in item response theory (IRT) during the past decade has focused researchers' attention on the advantages, both theoretical and practical, that IRT might offer to the equating process. Recently, a number of research studies investigating the feasibility of using IRT equating have been performed. Also, a number of large scale testing programs are either presently using IRT equating methods or contemplating their use in the near future. Therefore it was deemed useful at this point in time to summarize both what has been learned thus far and what we still need to learn about the use of IRT equating methods.

The purposes of this paper are five-fold; to discuss 1) When IRT equating methods should provide better results than traditional methods, and when traditional methods should suffice, 2) In those instances when IRT methods should provide better results, which IRT model, the three-parameter logistic or the one-parameter logistic (Rasch), is the most reasonable to use, 3) What unique contributions can IRT methods offer the equating process, 4) What work has been done, at ETS and elsewhere, that relates to the confidence that can be placed in the IRT equating results, and 5) What unresolved issues exist in the application of item response theory to the problem of equating tests.

In order to accomplish these purposes, a number of background topics will first be discussed; these include 1) the formal definitions and requirements of equating (Angoff, 1971; Lord, 1977, 1980) and the implications of these definitions for the equating that is normally done, 2) the basic principle of IRT equating, and the theoretical advantages it offers over traditional methods, 3) basic procedures for linking parameter estimates and deriving estimated true and observed score equatings using IRT, 4) the practical advantages to be gained from using IRT equating rather than traditional equating in an operational testing program, and 5) the distinction made by Rentz and Bashaw (1977) between test development and test analysis activities, and why the distinction is important in discussions of equating.

## Background Information

### Formal Requirements for Equating

Angoff (1971) has delineated, in the context of conventional equating methods, the basic requirements of equating; Lord (1977, 1980) has restated and elaborated upon these requirements in a form that is both illuminating and amenable to a consideration of IRT methods. These requirements will be discussed because they have a good deal of influence on what we realistically should expect the equating process to be able to do. According to Angoff (1971), there are four restrictions or requirements to be met by the equating process: 1) the instruments in question should measure the same attribute, 2) the resulting conversion should be independent of the data used in deriving it and be applicable in all similar situations, 3) scores on the two forms should, after equating, be interchangeable in use, and 4) the equating should be symmetric, or the same regardless of which form is designated as the base. Angoff (1971) goes on to discuss that equating, and the issue of unique conversions, can only be addressed when the test forms are parallel, and cites the definition of parallelism given by Gulliksen (1950):

> "Two tests may be considered parallel forms if, after conversion to the same scale, their means, standard deviations, and correlations with any and all outside criteria are equal."

A number of comments can be made that should prove useful for the discussion that follows. One, while the first restriction requires that the two forms measure the same attribute, it is not stipulated that the attribute be unidimensional. While there are certain psychometricians, most notably Lumsden (1960, 1976), who question whether measurement is meaningful for non-unidimensional content domains, unidimensionality is nowhere specified in Angoff's equating requirements. Unidimensionality, or a close approximation to it, will be a formal requirement of IRT equating methods, meaning that somewhat tighter restrictions on the nature of the test data must be met for IRT applications. Two, the independence of the conversions from the data used for deriving them falls short in practice anytime the groups taking the forms are not randomly equivalent samples from the population for which the conversions are to be relevant. This is, in fact, the usual situation in equating, where frequently non-random groups, often differing in ability, take the forms to be equated. Three, as pointed out by Angoff (1971), the criterion of interchangeability of scores only holds when the forms are equally reliable. Angoff also discusses the process of score calibration, which can be used for test forms of differing reliability. The calibrated forms can still be referenced to the same scale, but (theoretically) not used interchangeably.

Lord (1977, 1980) has further clarified the above restrictions, and in doing so, has pointed out the theoretical advantages to be gained from using IRT instead of traditional equating methods. Lord's (1977) formal definition of equating 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."

Under this definition, 1) tests measuring different traits or abilities can't be equated (comparable to Angoff's first restriction), 2) raw or observed scores on unequally reliable tests can't be formally equated (Angoff's third restriction), but also 3) observed scores on tests 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."

Lord (1980) shows further that while the equity requirement can be met for perfectly reliable or infalliable test data (i.e., true scores), for observed score data the equity requirement can be met only if the two forms are truly parallel (i.e., equivalent item by item), in which case, equating would not be necessary in the first place.

While the above would seem to build the case that in theory observed score equating is not possible under any circumstances, in practice, this is not true. Lord (1980) has noted that in many practical situations, different forms of the same test have been developed to be sufficiently parallel that traditional procedures yield good results. There will be problems in practice, however, anytime the test forms to be equated are not of the same difficulty (i.e., vertical equating situations) and observed scores are to be used. It is for this reason, and also to satisfy Angoff's restriction two (the conversions should be independent of the groups used to obtain them), that IRT methods have great appeal for the solution of equating problems.

Basic Principle of IRT Equating

The basic underlying property of IRT that makes it useful for equating applications is as follows. If the data being considered for the equating fit the assumptions of an IRT model, it is possible to obtain an estimate of an examinee's ability that is independent of the subset of items (test form) that the examinee responds to. Hence, it does not matter if an examinee takes an easy or hard form of a test; his/her ability estimate obtained from both forms will be identical, within sampling error, once the parameter estimates are placed on the same scale. Therefore the differences in difficulty of the forms being taken is no longer a concern. Further, if one is willing to use the ability (θ) metric for score reporting purposes, IRT eliminates the need for equating test forms. All that remains to be addressed is the placing of parameter estimates, derived from independent calibrations, on the same scale. This linking process will be described in the next section.

For a variety of reasons, large scale testing programs are often unable to report scores using the ability metric, and instead most continue to report scaled scores in a traditional manner even though IRT has been used for equating purposes. (Tests specifically developed using IRT procedures don't usually suffer the same problem and often use a variety of direct transformations of the ability metric, see Wright, 1977.) At ETS, the reason for continuation of the use of traditional scaled scores is that the scales existed long before IRT equating was considered, and the scales have properties that are accepted and understood by examinees. Fortunately, because any value of θ can be mathematically related to estimated true scores on the two forms, a situation exists whereby IRT equating of these estimated true scores can be utilized and traditional scaled scores reported. Further, Lord (1980) points out that the three requirements of the equating process, equity, invariance across groups, and symmetry, which are not met when observed scores are equated, are met when true (perfectly reliable) scores are equated. Hence, test forms of decidedly different difficulties can be equated if true scores are used, and further, the groups no longer have to be random in order to derive an equating relationship that is invariant across groups (from the same population). This has prompted Lord (1977) to say that "...conventional equating methods are not strictly appropriate when non-parallel tests having a non-linear relationship are administered to non-equivalent groups."

While the equating of IRT-derived true scores would seem to solve a number of equating problems that have been discussed, it should be noted that in practice we work with true score estimates, not the true scores, which remain unknown values. Lord (1980) has pointed out:

> "However, an estimated true score does not have the properties of true scores; an estimated true score, after all, is just another kind of fallible observed score."

While the above is true, what is important to note is that observed scores and estimated true scores are somewhat different fallible scores, incorporating different kinds of error. Further, by selecting items that fit the IRT model and calibrating on large enough samples, we can insure that our true score estimates are sufficiently close to the actual true values so as to derive the important benefits of the equating; this is not so easily done with observed scores. In sum, while the estimated true score equating will not be perfect, it will offer much more in problem equating situations (i.e., test forms varying greatly in difficulty) then can be derived from conventional observed score equating.

## The IRT Equating Process

IRT equating can be viewed simplistically as a two step process. Assuming that an IRT model has been chosen, the first step involves choosing an equating design and then dealing with the problem of getting parameter estimates from separate calibration runs within this design on the same scale. (When using certain computer programs, such as LOGIST, it is often the case that all parameter estimation can be accomplished in a single calibration run.) The second step involves performing the actual equating; if a program can report scores on the ability metric, the equating has been accomplished. However, because many testing programs report scores on some other scale, which is a transformation of the raw score scale, the second step becomes necessary.

There are essentially three equating designs used in IRT equating, and these designs are analogous to the most frequently used conventional designs. These designs are referred to as the 1) single group, 2) random groups, and 3) anchor test design. In the single group design, the same group takes both test forms to be equated. Because the same group takes both forms, differences in test difficulty are not confounded by differences in group abilities, and because of this, conventional methods work quite well, provided the forms are not of grossly differing difficulties. In the random groups design, two randomly selected groups each take a different form of the test. If the groups are truly random groups (from the same population), they should be at equivalent ability levels, and once again, differences in test form difficulty will not be confounded by ability differences, and conventional methods should work well unless the forms are of grossly differing difficulties. In the third design, two different groups of examinees take two different forms of a test; each form either contains a common set of items or a common anchor test is given with the forms. This is perhaps the most frequently used design for both horizontal and vertical equating situations. The groups do not have to be random, and more often they are not; if conventional methods are used, the common items are used to adjust for ability differences in the two groups. Depending both on the differences in difficulty of the forms and on the nature of the samples, this adjustment may or may not be effective, and hence, for this design, IRT equating can be seen as a very attractive alternative.

As a means of clarifying the need for a separate step to place parameter estimates on the same scale, consider the following situation which, while not characteristic of a situation encountered in equating applications, is quite instructive. Suppose the same set of items is given to two different groups of examinees, and the parameters for these items are estimated twice, once in one group and then separately in the other. 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; this is not so. When item and ability parameters are estimated simultaneously in the three-parameter logistic model, to ensure convergence in the estimation procedure, ability parameter estimates are placed on a scale with an arbitrarily chosen mean and standard deviation. The mean ability is usually set to zero and the standard deviation one, and the item parameter estimates (only difficulty and discrimination) are adjusted accordingly. If the two groups differ in ability level, the item parameter estimates will also differ. There will, however, be a linear relationship between item difficulties (or the θ's, which are on the same metric) estimated in the two groups, and this relationship can be used to place all parameter estimates on the same scale.

It should be clearly understood that when all items are administered to a single group of examinees and the parameters are estimated simultaneously, the item parameters are on a common scale. When this is not the case, i.e., when different sets of items are administered to the same group of examinees and calibrated separately, when the same set of items are given to different groups of examinees, or when different sets of items are administered to different groups of examinees, the item parameter estimates for the three-parameter logistic model are not on a common scale and must be adjusted. This adjustment is possibly only for the following three situations: 1) different sets of items are administered to the same group of examinees (common people are available), 2) the same set of items are administered to different groups of examinees (common items are available), or 3) some items that are the same (anchor test) and some items that are different are administered to different groups of examinees (again common items are available, but only a subset of the total). Situations one and three are characteristic of those encountered in practical IRT equating applications using the single group and anchor test design. Situation two might be encountered when comparing parameter estimates from pretest data with parameter estimates from operational form data. Appendix A of this paper describes in greater detail the procedures used for placing item parameter estimates on the same scale for the above three situations using the three-parameter and also the one-parameter logistic model. Also contained in this Appendix is an outline which delineates the placing of the parameter estimates on the same scale for the three equating designs discussed above.

As mentioned earlier, 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. This procedure is described in detail in Appendix B. It is also possible to use the estimated true scores to generate a frequency distribution of estimated number right observed scores on the two test forms. These scores may then be equated using traditional equipercentile methods. It should be noted that while the θ's estimated separately for two test forms share a linear relationship even if the forms are quite different in difficulty, the relationship between the estimated true scores will certainly be non-linear if the forms differ in difficulty. The same will be true of the relationship evidenced in the equating of the estimated observed score frequency distributions.

Because of the special nature of the Rasch model, it is possible to use the ability estimates obtained from a parameter estimation program to directly equate the actual observed scores. Like the other methods, this method is also not without its problems. The procedure is described in more detail in a section of Appendix B, as are the problems involved with the procedure.

## Practical Advantages of Using IRT Equating

Besides the theoretical advantage offered earlier for using IRT equating methods, i.e., it is the only reasonable method to use when tests or test forms of differing difficulty are given to non-random groups of differing abilities, there are also a number of practical advantages to be gained through the use of IRT. These include:

1. Improved equating, including better equating at the end of the scale where important decisions are often made. As mentioned before, it is possible to equate estimated true scores for all values of $\theta$, not just those actually obtained from the data.
2. Greater test security through less dependence on items in common with a single old form. If old forms of tests have calibrated items on the same scale, the common item block can come from multiple old forms.
3. Easier re-equating should items be revised or deleted. Presently, when traditional equating methods are used, if there are revisions or deletions of a substantial nature, the revised form must be readministered for equating purposes. If IRT equating of estimated true scores is used, the estimated true score for the revised test can be gotten by simply summing over the $\hat{P}_i(\theta)$ for those items left in the revised form.
4. The possible reduction of bias or scale drift which may occur in equating situations when traditional methods are used over time, most notably when the equating samples from the old and new forms are not random samples (from the same population). This will be discussed further in a later section of this paper.
5. The possibility of pre-equating, or deriving the relationship between the test forms before they are administered operationally. This is possible only when pre-test data is available. The use of IRT for pre-equating offers a unique contribution that can't be derived using traditional methods.

## Test Construction and Test Analysis

In discussing the problem of model-data fit for the Rasch model, Rentz and Bashaw (1975; 1977) delineated the differences between test construction and test analysis activities, a distinction that will prove most useful in clarifying when IRT equating methods are more advantageous than traditional methods. In test construction activities, the IRT model, in conjunction with content specifications, is used as a guide for selecting items on the test. Poorly fitting items to the model can be discarded, and items of moderately poor fit can be modified. Rentz and Bashaw (1977) state: "Thus, for this application, indications of model-data fit are necessary for items, the presumption being that the final collection of items will include only those that meet whatever criteria for fit might be established." For purposes of a discussion of equating, in this context, test construction would mean that the test to be equated and the base test have IRT parameter estimates for items that fit the model well or moderately well before equating is even considered.

In the test analysis situation, the final test form is fixed and badly fitting items can't be discarded. "Rather, the objective in this case is to derive whatever benefits the model is robust enough to provide, under potentially less-than-ideal item fit conditions." (Rentz and Bashaw, 1977). For equating purposes, test analysis activities would refer to fitting an IRT model to already existing new and base test data so that equating can be facilitated through the use of IRT methods. It would seem reasonable, however, to consider fitting an IRT model for equating purposes only if the IRT method offered something over any of the non-IRT equating procedures. If conventional procedures are deemed adequate, and nothing additional can be derived from IRT procedures, then going to the expense of an IRT equating and dealing with the problems of non-fitting items can be justified only in the weakest sense by the fact that it can serve as a check on the conventional equating.

## Discussion Section

When should IRT equating methods provide better results than traditional methods, and when should traditional methods suffice?

In answering this question, three distinctions are useful. These are 1) Whether the equating is being done in a test construction or test analysis mode, 2) Whether the test or test forms to be equated differ greatly in difficulty (this is the usual horizontal-vertical equating distinction, although it is possible to have test forms at the same level which differ greatly in difficulty), and 3) What is the nature of the samples taking the tests or test forms. Are they random groups from the same population; if they are non-random, do they differ greatly in the ability being measured?

If the test forms to be equated have been specifically designed or constructed using IRT test development procedures, then IRT methods should be used for equating. It would prove impractical to throw away useful parameter information and equate using traditional methods. While it is true that the traditional methods will work well if the tests do not differ greatly in difficulty and the groups in ability, IRT procedures "protect" from the problems encountered when this is not the case. The IRT equating methods should work tolerably well across all combinations of differences in test difficulty and group ability. Choice of specific IRT model for equating will be dictated by the choice of the model used in the actual test construction process.

If the test forms have been assembled using standard test development procedures (i.e., the test analysis mode), then IRT equating should be considered only in those instances where traditional methods do not work well. These instances include 1) vertical equating situations, where tests differing in difficulty are given to groups of differing abilities or 2) horizontal equating situations where test forms of differing difficulty are given to non-random groups that may differ in ability (the usual anchor test design). Further, if the test forms do not differ greatly in difficulty but the groups are non-random groups from the same population, conventional methods, while working tolerably well, will not insure that the equating results are generalizable to other groups for whom the forms are appropriate. IRT equating methods, used in this instance, will insure generalizability.

In an attempt to clarify those instances in which IRT equating should provide better results than traditional methods, entries have been placed in the following two-way table:

| | | Equating | |
|---|---|---|---|
| | | Horizontal | Vertical |
| Activity | Test Construction | IRT | IRT |
| | Test Analysis | IRT or Conventional[1] | IRT |

As substantiation for the above generalizations, a number of research studies can be cited. Lord (1975), in comparing traditional and IRT equating for the three basic equating designs, found good correspondence between traditional and IRT equatings for tests not differing widely in difficulty when using the

[1] The choice of method should be determined through a consideration of the differences in difficulty of the test forms, the differences in ability of the groups, and the necessity for generalizable equating results.

single group and random groups designs, where differences in ability level are not an issue. Lord (1975) did find, however, substantial differences between conventional and IRT equating for tests differing in difficulty given to non-equivalent groups when using an anchor test approach. Marco, Petersen, and Stewart (1979) also found that IRT methods were superior to traditional methods when tests of differing difficulty were equated using an anchor test approach. A number of researchers (Beard and Pettie, 1979; Golub-Smith, 1980; Rentz and Bashaw, 1975, 1977) have confirmed the fact that traditional and IRT equatings correspond well when a horizontal equating of test forms is done, even when the test forms were not specifically developed to fit a particular IRT model. These researchers have been working with the Rasch model, and while the results are encouraging in terms of suggesting the Rasch model is robust in equating situations, from a practical standpoint, the fact that the methods behave similarly suggests continued use of conventional methods unless some additional benefits accrue from the IRT equating.

When IRT methods provide better results, which IRT model should be used?

Substantial recent research sheds some light on which IRT model to use when performing vertical equating in test analysis situations. Slinde and Linn (1978, 1979), Loyd and Hoover (1980), and Kolen (1981) have demonstrated, using either direct equating or indirect techniques, that the Rasch model is probably inappropriate for the vertical equating of tests not specifically designed to fit the model. Gustafsson (1979a, 1979b) has pointed out one reason for the failure of the Rasch model in this situation. When guessing behavior is present in the item responses for tests being vertically equated, a negative correlation results between traditional item difficulty and item discrimination indices. Since item difficulties are bound to differ for the forms, the negative correlation forces the discriminations to vary also, thereby bringing to test the equal item discrimination assumption of the Rasch model. While the results of the study by Loyd and Hoover (1980) also demonstrate a problem with the Rasch model for vertical equating situations, these authors are concerned that because the nature of the content specifications for the test changes appreciably with level, there may be a problem of unidimensionality across levels that is causing the failure of the Rasch model. The issues raised by Gustafsson and Loyd and Hoover have implications as to whether the three-parameter logistic model should be better than the Rasch model for vertical equating. If, as pointed out by Gustafsson (1979b), the item discriminations vary across forms due to the existence of guessing, the three-parameter logistic model, which can handle variation in item discriminations and also guessing, should prove useful. If however, the problem is one of dimensionality, as Loyd and Hoover (1980) point out, no unidimensional IRT model can solve the problem. Further, while certain studies (Kolen, 1981; Marco, Petersen, and Stewart, 1979) point to a superiority of the three-parameter logistic model in vertical equating situations, there is always the problem of deciding on a criterion upon which to judge which method is superior. The results at present do seem to suggest, however, that the three-parameter logistic model offers ~~a more~~ viable alternative for the vertical equating of approximately unidimensional tests.

In the horizontal equating of test forms in test analysis situations, IRT methods should be considered when the test forms differ somewhat in difficulty and the groups are non-random and non-equivalent in nature, which usually occurs with anchor test designs. The results of the Marco, Petersen, and Stewart study (1979) suggest that, for test forms that differ in difficulty developed from the same set of content specifications, the three-parameter logistic model is superior for equating purposes. Kolen (1981) has pointed out, however, (as did Marco et al) that the criterion for judging the superiority of equating methods in their study may have been biased against certain of the methods.

For the horizontal and vertical equating of test forms that have been specifically constructed to fit an IRT model, the choice of model for equating follows automatically from the choice of model in the test construction process. Little has been specifically written, however, about which IRT model should prove superior in test construction activities for horizontal and vertical equating situations. The comments that follow are gleaned from the research done on the vertical and horizontal equating of tests in a test analysis mode, with the hope that these results generalize to test construction activities. It would appear that for test forms developed from the same set of test specifications, either the Rasch or three-parameter logistic model can be used in the test construction process. Of course, the added assumptions of the Rasch model, equal item discriminations and no guessing, must be dealt with, but if the developer has reasonable flexibility to choose fitting items and still meet the original (or slightly revised) content specifications, the Rasch model is viable. In fact, it would be to the developer's best interest to use the Rasch model whenever possible because of the measurement consequences that result. When tests or test forms are being developed to purposely test at different levels however, the nature of the content specifications must also change somewhat across levels (see Slinde and Linn, 1977); and because of this fact, it will be a much more difficult task to prepare items that measure the content specifications, are at a difficulty level appropriate for the level being tested, and at the same time, are equally discriminating across all levels. It should be noted that if this is not possible, certain researchers (Lumsden, 1978; Wood, 1978) would say that there is a dimensionality problem. According to Lumsden (1978), "Test scaling models are self-contradictory if they assert both unidimensionality and different slopes in the item characteristic curves." A similar conclusion may result, however, from purely content considerations. Is it reasonable to expect the assumption of unidimensionality to underlie a set of test forms designed to test individuals at grossly different levels of ability? In sum, the issue in vertical test construction situations may not ultimately be whether the three-parameter logistic model is more viable than the Rasch model, but whether any IRT model is appropriate. This of course is an equally reasonable question to pose for vertical equating in test analysis situations.

What unique contributions can IRT methods offer the equating process?

There are at least three situations in which IRT methods can make a unique contribution to the process of test equating; that is, an equating can be accomplished that would have been either impossible or of minimal utility when using conventional methods.

The first of these situations involves the pre-equating of test forms. Pre-equating refers to the process of establishing equating conversions between a new form and a base form or forms prior to the time the new form is administered. The process depends on the adequate pretesting of a pool of items from which the new test form will be built, the calibration of these items using IRT methods, and the utilization of a linking scheme to place the IRT parameters from the pretested items all on the same scale and also on the same scale as the old form(s). The process of pre-equating is presently under investigation at ETS because at least three very important outcomes accrue from the process. One, IRT-based pre-equating is unaffected by the possible future problem of revealing common item equating sections under disclosure legislation because there would be no need for these sections in the first place. Two, since equating using IRT pre-equating methods is possible prior to the actual administration of the test, new test forms can be introduced at low volume administrations; a particular problem if conventional methods had to be used. Three, pre-equating removes the equating process from the score reporting cycle (the period from the time the test is administered to the time scores are reported), thereby minimizing the chance of equating errors and at the same time freeing up time for other psychometric activities.

A second unique contribution of IRT to the test equating process involves equating tests that do not contain common items and, at present, can't be pre-equated. As an example, consider the following. Each October, two forms of the Preliminary Scholastic Aptitude Test/National Merit Scholarship Qualifying Test (PSAT/NMSQT) are administered, and for security reasons, the two forms contain no common items. As a result, the two forms are not equated to each other, but are both equated to the same two old SAT test forms. Comparability of scores across the two forms is thus established indirectly through a mutual relationship with the SAT forms. It would obviously be more desirable to effect a direct form-to-form equating rather than depend on the indirect equating presently used. If the data collected at the two administrations can be arranged as in Figure 1, it is possible, using LOGIST, to estimate all item and ability parameters in a single computer run. Hence, item parameters for both PSAT/NMSQT forms will be on the same scale, thus providing a direct equating of the ability

| Group | PSAT/NMSQT Form 1 Unique Items n=20 | PSAT/NMSQT Form 1 - SAT First Old Form Common Items n=22 | PSAT/NMSQT Form 1 - SAT Second Old Form Common Items n=23 | PSAT/NMSQT Form 2 Unique Items n=19 | PSAT/NMSQT Form 2 - SAT First Old Form Common Items n=23 | PSAT/NMSQT Form 2 - SAT Second Old Form Common Items n=23 | SAT First Old Form Unique Items n=40 | SAT Second Old Form Unique Items n=39 |
|---|---|---|---|---|---|---|---|---|
| PSAT/ NMSQT Form 1 | X | X | X | Not Reached | Not Reached | Not Reached | Not Reached | Not Reached |
| PSAT/ NMSQT Form 2 | Not Reached | Not Reached | Not Reached | X | X | X | Not Reached | Not Reached |
| SAT First Old Form | Not Reached | X | Not Reached | Not Reached | X | Not Reached | X | Not Reached |
| SAT Second Old Form | Not Reached | Not Reached | X | Not Reached | Not Reached | X | Not Reached | X |

Figure 1: Calibration Plan for Direct IRT Equating of PSAT/NMSQT Form 1 Verbal Section to PSAT/NMSQT Form 2 Verbal Section. The entire matrix represents a single calibration run. Crosses indicate items that examinee groups were actually exposed to. Each PSAT/NMSQT and SAT sample contains approximately 2,000 cases.

estimates. An equating of estimated true scores or estimated observed score frequency distributions automatically follows. The results of doing the above have been reported by Cook, Dunbar, and Eignor (1981).

The final unique contribution of IRT to the equating process involves the equating of a test comprised of items from a locally developed item bank to a standardized norm-referenced test that has national norms data. Any test made up of items from the item bank may then be used in conjunction with the norms data, provided the items from the bank have been calibrated and placed on the same underlying scale. The items comprising the test can then be matched to the measurement need (for instance, pretest or posttest) and the norms data can be used for evaluation of pupil growth. Holmes (1980) has investigated the above procedure for use in Title I evaluations, using the one-parameter logistic or Rasch model. The procedure, based upon what was documented in the Holmes report, is as follows:

1. A local item bank testing relevant content taught in a district or system is developed. An IRT model is fit to the items (the content domain must be reasonably unidimensional), based on pre-test data, and all the parameter estimates are placed on the same scale.
2. A norm-referenced test which tests comparable content and has representative national norms is selected.
3. A test built from the local item bank and the norm-referenced test are administered to the same group of examinees.
4. All items from both tests are calibrated together. For a particular item bank test score, the equivalent ability estimate can be determined. In turn, this estimate and the item parameters for the norm-referenced test allow the determination of the equivalent normed test score.

This is done for the range of item bank test scores, which in total comprises an IRT (estimated true score) equating of scores on both tests.

5. Each equated normed test score has a percentile rank associated with it that can be converted into a Normal Curve Equivalent (NCE) score required for Title I evaluation purposes. These percentile ranks can be determined through interpolation of the raw score to percentile norms table provided with the normed test.
6. The equated item bank test scores are translated into item bank ability estimates using the item parameter estimates already in existence for all the items from the pre-test data.
7. The end result is a one-to-one correspondence between total item bank ability estimates and NCE units, to be used for evaluation purposes.
8. Any possible subset of items from the bank selected for a particular purpose results in measurement on the common ability metric which can be related to the NCE units. Tests that measure relevant local content and are peaked to provide maximum information for the examinee group can then be developed and administered with the resulting measurement of growth on the mandated NCE scale.

The unique aspect of this process is not the equating of a locally developed test to a nationally normed test (this could be done using conventional methods), but the equating of the local bank ability scale to the norm-referenced test. Without having done this, each locally developed test would have to be equated, rather than the equating being done only once.

It should be noted that a major concern expressed by Holmes in the project report was fit of the data to the Rasch model. While we also share a similar concern, expressed further in a later section of the paper, nothing precludes the use the three-parameter model in Holmes' study. The equating done was not the actual raw to raw equating through estimated abilities that can be done only with the Rasch model, but instead, estimated true score equating, which can be done with any of the models.

What has been done that relates to the confidence that can be placed in IRT equating results?

The problem involved in evaluating the results of any IRT equating concerns the criterion measure. Since nobody ever knows what the true equating may be, i.e., the best criterion against which to judge the results of the actual equating, other criterion measures have often been devised; these vary in degree of complexity and in assumptions made. In situations where conventional equating methods are known to function well or have been in existence for some time, the results of the conventional method(s) forms a criterion against which the IRT equating may be evaluated (see Lord, 1975; Beard and Pettie, 1979; Rentz and Bashaw, 1977; Golub-Smith, 1980; Marco, 1977; Woods and Wiley, 1977, 1978). In other situations, the test itself may form a criterion; that is, the test is equated to itself (see Lord, 1975, 1977; Marco, Petersen, and Stewart, 1979). To the extent that the equating results coincide with expectation, one has confidence in the method. In other situations, one can use stability of equating rather than accuracy of equating as a criterion measure for evaluative purposes. Kolen (1981) cross-validated his equating results with random samples of individuals. More specifically, he formed frequency distributions for his random cross-validation samples and then compared his equated score frequency distributions with these; a mean squared difference between scores with identical percentile ranks was used for evaluative purposes. Loyd and Hoover (1980) formed a somewhat different criterion against which to evalute the results of their study, which involved the use of the Rasch model in vertical equating of forms given to examinee groups of differing abilities. They equated the same forms using groups of comparable abilities. A comparison of the two equatings then allows one to ascertain whether the results obtained were greater than those expected from simple sampling differences in parameter estimates obtained for groups of comparable abilities.

Another way to gain confidence in IRT or conventional equating results is through a consideration of the scale drift that occurs when multiple forms of a test are equated over time. Scale drift will have occured if the results of equating Form A to Form D is not the same as that obtained by equating Form A to Form D through intervening Forms B and C. One would have confidence in the equating method that resulted in the least scale drift. A problem with the above example is that there is no good way of knowing which equating method was best for directly equating Form A to Form D. An excellent way of dealing with this problem is through the use of a circular closed chain, as depicted in Figure 2. Form V4, which has previously been put on scale, can be equated to itself through the five intervening forms. Any discrepancy between the transformation obtained from the circular chain of equatings and the initial V4 scale could be attributed to scale drift. One would then have confidence in the equating method that resulted in the least discrepancy between the initial scale of V4 and the scale resulting from the chain of equatings. A study comparing scale drift for IRT and conventional equating methods applied to aptitude test data has been done by Petersen, Cook, and Stocking (1981). A similar study using achievement test data is presently

$$\begin{tikzcd}
V4^{1} \arrow[r] & fe^{2} \arrow[r] & X2 \arrow[r] & fm \arrow[r] & Y3 \arrow[r] & fw \arrow[d] \\
et \arrow[u] & Z5 \arrow[l] & fu \arrow[l] & Y2 \arrow[l] & fk \arrow[l] & B3 \arrow[l]
\end{tikzcd}$$

Figure 2: Verbal Aptitude Test Equating Chain Taken from Petersen, Cook, Stocking Study (1981).

[1] Denotes operational verbal test form.

[2] Denotes common item equating section.

being conducted at ETS.

In sum, a number of ways have been devised for evaluating IRT and conventional equating results. These methods can be viewed as practical solutions to the problem that one never knows what the true or best equating criterion is in a particular situation.

What are the unresolved issues that relate to IRT equating?

There are two varieties of unresolved issues involving IRT equating. One set of issues has to do with the mechanics of IRT equating, and these may be called direct equating issues. The other set of issues has to do with the use of IRT in the test construction process, and how this then relates to IRT equating. These are more indirect issues, such as dimensionality, but they do influence what can be reasonably expected from an IRT equating. These indirect issues will be touched upon briefly, and then the more direct equating issues discussed in some detail.

Most of the IRT test construction work has been done using the Rasch model. Advocates of using the Rasch model in test construction situations stress that the most important criterion in deciding upon items for a test is goodness of fit of the items to the model (Rentz and Rentz, 1978). Recently, two levels of concern have been voiced reflecting this focus on goodness of fit. Wood (1978) and Whitely (1977) are concerned that this focus will necessarily restrict measurement to domains that, while unidimensional, do not necessarily measure what we really want to measure. Gustafsson (1979a), on the other hand, is concerned that the usually applied Rasch goodness of fit tests are not sensitive to multidimensionality among the items, and advocates the application of a number of other tests sensitive to violations of unidimensionality. Finally, Wood (1978) has fitted random data to the Rasch model and was not stopped by the usual goodness of fit tests. While not wanting to enter further into a debate about the use of goodness of fit tests to construct unidimensional tests, we shall note from the above that the use of IRT equating in test construction activities may not be as straightforward as suggested. If the constructed tests are not unidimensional, then the issue becomes exactly the same as that addressed in the test analysis mode--namely, how robust is IRT to violations of assumptions in equating situations. Hence, unless the process of test construction leads to a unidimensional domain of meaningful content, IRT equating procedures must be considered in a different light, no longer as a natural outcome of the test development process.

There are a number of more direct unresolved issues that will be addressed next. Many of these issues have come to the front in the IRT equating work that is ongoing at ETS. When using the three-parameter logistic model for equating, two specific issues have come up. One has to do with the type of score to be equated when ability estimates cannot be used for reporting purposes. This is particularly a problem for testing programs that have a long history of use of a particular scale and forms placed on that scale through conventional observed formula score equating. When IRT equating is done, should the relationship between estimated number right true scores, estimated true formula scores, or estimated number right observed score frequency distributions on the new and base forms be used to place the new form on scale? Ideally, the relationship between estimated observed formula score frequency distributions should be used, but this relationship is unobtainable using IRT methods. The second issue has to do with which calibration design is best for linking parameter estimates for which sort of data. As explained in Appendix A, there are essentially three methods of getting parameter estimates on the same scale using LOGIST with anchor test designs. Method one, called concurrent calibration, involves running all the data in one LOGIST run, treating data for a particular group on the form not taken as not reached. (Figure 1 represents a concurrent calibration run.) Method two involves fixing the difficulties for the common items in the second calibration run at the values estimated in the first calibration run. Method three involves estimating the parameters separately in two calibration runs and then using the relationship between the difficulty parameters for the common items to place all parameter estimates on the same scale. Experimentation at ETS with these methods seems to suggest that no one method is uniformally best, but that the choice of method seems to vary with the data set.

Another issue presently of interest has to do with the one-parameter logistic model, where essentially two separate IRT equating procedures can be used. One procedure, used by Rentz and Bashaw (1975) and Loyd and Hoover (1980), is based on the direct relationship between Rasch model observed scores and ability estimates. Observed scores on test forms corresponding to the same ability estimate are considered equated. The other procedure, used by Kolen (1981), corresponds to that usually used for the three-parameter logistic model, where there is no direct relationship between observed score and ability. For any particular ability, knowledge of the item parameter estimates for each form allows generation of estimated true scores, which can be considered equated (see Appendix B). From these estimated true scores, frequency distributions of estimated observed scores may be generated and equated using conventional equipercentile methods. While the first procedure mentioned above is straightforward, there is a problem if for a particular ability level, corresponding integer raw scores do not exist on the two forms. With the other procedure, the problem of missing data does not exist because the estimated true score relationship can be determined for any ability level, not just those ability estimates derived from the data. Of interest is which procedure would be best to use in which situation.

There are a number of other issues of a more general nature that will be briefly mentioned. One has to do with the demonstration of unidimensionality for tests being vertically equated. For a variety of reasons, the assumption of unidimensionality can be violated for tests that are intentionally built to vary in difficulty, and procedures need to be considered that address this concern. Another issue has to do with determining which types of test data IRT equating procedures will work best with and what types are problematic. While this can be viewed as a dimensionality issue, a robustness issue, or both, there is more to it. It is conceivable that IRT equating will be of differential utility for a variety of tests, all of which do not greatly violate the assumption of unidimensionality. It would be useful to know for which kinds of tests IRT equating works best and for which it gives the poorest results. Finally, an issue presently of interest at ETS is how to determine or establish a base scale when using IRT procedures. As mentioned earlier, for a variety of tests, ETS is locked into using a previously established scale. The issue of a new scale would present itself if either a new program were being introduced or a decision were made to change content specifications on existing tests to the extent that equating was no longer possible and perpetuation of the existing scale unreasonable. Should scores then be reported on the ability metric, some linear transformation of that scale, the estimated true score scale, or the estimated observed score scale? Of interest is the generation of arguments in favor of each scale so that an informed decision can be made.

## Conclusions

The purpose of this paper was to address, using available research, some practical issues relating to IRT equating procedures. The outcome of the paper is most likely that we have brought up more issues yet to be resolved than we have clarified existing issues. This is undoubtedly due to what is presently known about IRT equating procedures. Hopefully as more IRT equating research is done the questions posed in this paper will come to be resolved.

## REFERENCES

Angoff, W.H. Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Beard, J.G., and Pettie, A.L. A comparison of linear and Rasch equating results for basic skills assessment tests. Paper presented at the annual meeting of AERA, San Francisco, 1979.

Cook, L.L., Dunbar, S.A., and Eignor, D.R. IRT equating: A flexible alternative to conventional methods for solving practical testing problems. Paper presented at the annual meeting of AERA, Los Angeles, 1981.

Golub-Smith, M. The application of Rasch model equating techniques to the problem of interpreting logitudinal performance on minimum competency tests. Paper presented at the annual meeting of AERA, Boston, 1980.

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Gustafsson, J.-E. Testing and obtaining fit to the Rasch model. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1979. (a)

Gustafsson, J.-E. The Rasch model in vertical equating of tests: A critique of Slinde and Linn. Journal of Educational Measurement, 1979, 16,153-158. (b)

Holmes, S.E. ESEA Title I linking project: Final report. Salem Oregon: Oregon Department of Education, 1980.

Jaeger, R.M. Some exploratory indices for selection of a test equating method. Journal of Educational Measurement, 1981, 18, 23-38.

Kolen, M.J. Comparisons of traditional and item response theory methods for equating tests. Journal of Educational Measurement. 1981, 18, 1-11.

Lord, F.M. A survey of equating methods based on item characteristic theory. Research Bulletin 75-13. Princeton, NJ: Educational Testing Service, 1975.

Lord, F.M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

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

Loyd, B.H. & Hoover, H.D. Vertical equating using the Rasch model. Journal of Educational Measurement, 1980, 17, 179-194.

Lumsden, J. The construction of unidimensional tests. Psychological Bulletin, 1961, 58, 122-131.

Lumsden, J. Test theory. Annual Review of Psychology, 1976, 27, 251-280.

Lumsden, J. Tests are perfectly reliable. British Journal of Mathematical and Statistical Psychology, 1978, 31, 19-26.

Marco, G.L. Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 1977, 14, 139-160.

Marco, G.L., Petersen, N.S. & Stewart, E.E. A test of the adequacy of curvilinear score equating models. Paper presented at the 1979 Computerized Adaptive Testing Conference, Minneapolis, 1979.

Petersen, N.S., Cook, L.L., and Stocking, M.S. Scale drift: A comparative study of IRT versus linear equating methods. Paper presented at the annual meeting of AERA, Los Angeles, 1981.

Rentz, R.R., and Bashaw, W.L. Equating reading tests with the Rasch model, volume I final report, volume II technical reference tables. Athens, GA: University of Georgia, Educational Research Laboratory, 1975. (ERIC Document Reproduction Nos. ED 127 330 through ED 127 331).

Rentz, R.R. & Bashaw, W.L. The national reference scale for reading: An application of the Rasch model. Journal of Educational Measurement, 1977, 14, 161-180.

Rentz, R.R., and Rentz, C.C. Does the Rasch model really work? A discussion for practitioners. NCME Measurement in Education, 1979, 10(2).

Slinde, J.A. & Linn, R.L. Vertically equated tests: Fact or phantom? Journal of Educational Measurement, 1977, 14, 23-32.

Slinde, J.A. & Linn, R.L. An exploration of the adequacy of the Rasch model for the problem of vertical equating. Journal of Educational Measurement, 1978, 15, 23-35.

Slinde, J.A. & Linn, R.L. A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty. Journal of Educational Measurement, 1979, 16, 159-165.

Whitely, S.E. Models, meanings and misunderstandings: Some issues in applying Rasch's theory. Journal of Educational Measurement, 1977, 14, 227-235.

Wood, R. Fitting the Rasch model--A heady tale. British Journal of Mathematical and Statistical Psychology, 1978, 31, 27-32.

Woods, E.M. & Wiley, D.E. An application of item characteristic curve equating to single form tests. Paper presented at the annual meeting of the Psychometric Society, Chapel Hill, NC, 1977.

Woods, E.M. & Wiley, D.E. An application of item characteristic curve equating to item sampling packages on multi-form tests. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

Wright, B.D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

## Appendix A

### Scaling Parameters -- Three-Parameter Logistic Model

Situation 1. Two different sets of items (Form X and Form Y) are given to the same group of examinees

a. Calculate: $M_{\theta_X}$, $SD_{\theta_X}$, $M_{\theta_Y}$, $SD_{\theta_Y}$

where X and Y designate Forms X and Y and M and SD represent the means and standard deviations of θ's (ability parameters) estimated by the two test forms.

b. If the assumptions of the model are met, the θ's will have the following linear relationship:

$$\theta_Y = A\theta_X + B \quad (1)$$

$$\text{where } A = \frac{SD_{\theta_Y}}{SD_{\theta_X}}$$

$$\text{and } B = M_{\theta_Y} - AM_{\theta_X}$$

c. The item parameters are adjusted as follows;

$$c_g^* = c_g \quad (2)$$

$$a_g^* = a_g A \quad (3)$$

$$b_g^* = \frac{b_g - B}{A} \quad (4)$$

Situation 2. The same set of items is given to two different groups of examinees (Group A and Group B)

a. Calculate: $M_{b_A}$, $SD_{b_A}$, $M_{b_B}$, $SD_{b_B}$

where M and SD represent means and standard deviations, the subscripts A and B represent groups and b represents the item difficulty parameter.

b. If the assumptions of the model are met, the b's will have the following linear relationship

$$b_B = Ab_A + B \qquad (5)$$

where, $A = \frac{SD_{b_B}}{SD_{b_A}}$

and $B = M_{b_B} - AM_{b_A}$

c. The item discrimination ($a_g$) and psuedo guessing parameters ($c_g$) as well as ability estimates ($\theta_a$) are adjusted as follows:

$$c^*_g = c_g \qquad (6)$$

$$a^*_g = a_g A \qquad (7)$$

$$\theta^*_a = \frac{(\theta_a - B)}{A} \qquad (8)$$

Situation 3. Some items that are the same and some items that are different are administered to different groups of examinees (Group A and Group B)

a. Expressions 5-8 can be used in this situation. Linear parameters (A and B) determined from the common items given to the two groups of examinees are used to adjust all item parameter and ability estimates obtained for one of the forms to the scale of the second form.

b. The following is an alternative method that may be used in this situation.[1]

i. Estimate parameters for Form Y and the common items using data obtained when the form was given to Group B

ii. Estimate parameters for Form X and the common items using data obtained when the form was given to Group A holding the b values for the common items fixed at estimated values obtained from Group B

iii. This procedure ensures that Form X item parameters and ability estimates will be on the Form Y scale.

### Scaling Parameters -- One-Parameter Logistic Model

Situation 1. Two different groups of items (Form X and Form Y) are given to the same group of examinees

[1] Not all computer programs have the capabilities of accepting parameter estimates from a previous run. LOGIST, the computer program used at ETS, does have this capability.

a. Calculate: $M_{\theta_X}$, $M_{\theta_Y}$

b. Calculate the linking constant, $k = M_{\theta_Y} - M_{\theta_X}$

c. Adjust all ability parameters estimated by Form X as follows:

$$\theta^*_{a_X} = \theta_{a_X} + k \qquad (9)$$

d. Adjust all Form X difficulty parameters as follows:

$$b^*_{g_X} = b_{g_X} + k \qquad (10)$$

Situation 2. Two different groups of items (Form X and Form Y) along with a common set of items are given to two different groups of examinees (Group A and Group B)

a. Calculate: $M_{b_A}$, $M_{b_B}$

where $M_{b_A}$ and $M_{b_B}$ refer to the mean easiness of the common items given to the respective groups

b. Calculate the linking constant, $k = M_{b_B} - M_{b_A}$

c. Adjust the form X item easiness parameters as follows:

$$b^*_{g_X} = b_{g_X} + k \qquad (11)$$

d. Adjust all ability parameters estimated by Form X as follows:

$$\theta^*_{a_X} = \theta_{a_X} + k \qquad (12)$$

Situation 3. The same set of items is given to two different groups of examinees (Group A and Group B)

a. In this case, if one calculates $M_{b_A}$ and $M_{b_B}$ based on all the items, they will be equal within sampling error. Hence, there is no linking constant — all parameter estimates are on the same scale without adjustment.

## Equating Designs

A. Single Group Design

1. Two test forms are given to the same group of examinees

a. Conventional methods work very well in this situation

b. Simplest approach would be to estimate all item and ability parameters in a single computer run

i. All item parameters and ability estimates will be on the same scale

ii. Estimated θ's obtained from the two forms will be identical except for sampling error. If one is willing to report ability estimates to examinees, no further effort is necessary.

c. Item parameter and ability estimates could be obtained in two separate computer runs

i. This would necessitate placing item and ability parameter estimates on the same scale. Procedures given for Situation 1 could be used for this purpose.

B. Random Groups Design

1. Two randomly selected groups each take a different form of the same test

a. Conventional methods work fairly well in this situation

b. Assumption is that two groups are equivalent in ability

c. Could analyze the data in two separate computer runs and use the procedure described in Situation 1 to place item and ability parameters on the same scale.

d. The following procedure could also be used

i. Analyze the data in two separate computer runs

ii. Obtain a distribution of $\hat{\theta}$'s for each data set, e.g. Form X given to Group A, Form Y given to Group B

iii. Equate the $\hat{\theta}$'s obtained from the two runs by ordinary equipercentile methods

C. Anchor Test Design

1. Two groups of examinees take two different forms of a test, but each form contains a common set of items

   a. Simplest way to accomplish the equating is to estimate all item and ability parameters together in a single computer run

   b. The two forms of the test to be equated are considered to be one long test consisting of items comprising Form X and Form Y

   c. All of the examinees in both groups (Group A and Group B) are assumed to have taken all of the items in both test forms; where there are no responses, the items are assumed to be not reached.[1]

   d. The item parameter estimates for both forms will be on the same scale and ability estimates obtained from either form will be equivalent

   e. Suppose item and ability parameters have been obtained separately for the test forms administered to their respective groups

      i. Procedures given for Situation 3 could be used to place all item and ability parameters on the same scale

      ii. An alternate procedure would be to estimate the item and ability parameters for Form Y given to Group B Following this, the item and ability parameters for Form X given to Group A would be estimated fixing the item difficulty parameters for the common set of items contained in Form X at the values previously obtained from the Form Y estimation procedure.[1]

[1] The computer program LOGIST used at ETS for parameter estimation has the capability for dealing with these options; other programs do not. Hence, parameter estimates derived from these programs must be put on scale using the method described previously in Situation 3.

## Appendix B

## Alternatives to Equating Ability Estimates

### A. Equating Estimated True Scores

When reporting $\hat{\theta}$'s is not a viable alternative for a testing program, it is possible to use the relationship between θ and true score to obtain equated estimated number right true scores

1. If Form X and Form Y are both measures of the same ability, θ, then their estimated number right true scores can be calculated as follows:

$$\hat{T}_X = \sum_{i=1}^{n_X} \hat{P}_i(\theta) \qquad (1)$$

$$\hat{T}_Y = \sum_{j=1}^{n_Y} \hat{P}_j(\theta) \qquad (2)$$

where,

$\hat{T}_X$ = Form X estimated true score for a given θ

$\hat{T}_Y$ = Form Y estimated true score for a given θ

and $\hat{P}_i(\theta)$, $\hat{P}_j(\theta)$ are the item response functions for items i, i=1...$n_X$ (in Form X) and j, j=1...$n_Y$ (in Form Y) respectively, using parameter estimates

2. Using expressions 1 and 2, it is possible to find an estimated number right true score on Form X that is equivalent to an estimated number right true score on Form Y for any given θ.

It is also possible to use the parameter estimates to obtain equated estimated true formula scores

1. Estimated true formula scores are calculated in a manner similar to that used to obtain estimated true scores:

$$\hat{R}_X = \sum_{i=1}^{n_X} \hat{P}_i(\theta) - [\sum_{i=1}^{n_X} \hat{Q}_i(\theta)]/A-1 \qquad (3)$$

$$\hat{R}_Y = \sum_{j=1}^{n_Y} \hat{P}_j(\theta) - [\sum_{j=1}^{n_Y} \hat{Q}_i(\theta)]/A-1 \qquad (4)$$

where,

$A$ = the number of choices per item

$\hat{R}_X$ = Form X estimated true formula score for a given $\theta$

$\hat{R}_Y$ = Form Y estimated true formula score for a given $\theta$

$\hat{P}_i(\theta)$, $\hat{P}_j(\theta)$ are as defined previously

and $\hat{Q}_i(\theta)$, $\hat{Q}_j(\theta)$ are equal to $1-\hat{P}_i(\theta)$, $1-\hat{P}_j(\theta)$, respectively

2. As was the case for estimated number right true scores, it is possible to find an estimated true formula score on Form X that is equivalent to an estimated true formula score on Form Y for any given $\theta$. It should be noted that in both instances, the equations implicitly assume; however, that every individual responds to all items; i.e., there are no omissions or not reached items.

## B. Equating Estimated Number Right Observed Score Frequency Distributions

A third possibility is to generate estimated number right observed score distributions for Form X and Form Y and to equate these observed score distributions using ordinary equipercentile equating methods

1. The frequency distribution of <u>number-right</u> observed scores for a given $\theta$, $f(x|\theta)$ is a generalized binomial distribution (Kendall and Stuart, 1969, Section 5.10). This distribution can be generated by the generating function.

$$\prod_{i=1}^{n} (P_i + Q_i) \qquad (5)$$

2. Using the parameter estimates in $P_i$ and $Q_i$, the estimated total group or marginal distribution of number right observed scores will be

$$f(x) = \frac{1}{N} \sum_{a=1}^{n} f(x|\theta_a) \qquad (6)$$

where, a indexes examinees

## C. Which type of Score is Most Appropriate

1. Estimated true scores have the following disadvantages:

a. The possible range for true scores is only from $T = \sum_{i=1}^{n} c_i$ ( the pseudo chance level) to $T = n.$[1] Many

[1] This is, of course, a problem when the three-parameter logistic model is used.

testing programs report scores below this level and therefore require an equating process that will provide conversions for the lower level scores

b. A procedure exists for providing these conversions

i. Determine the mean and standard deviation of scores below chance level on Form X

$$M_X = \frac{A}{A-1} \sum_{i=1}^{n_X} c_i - \frac{n_X}{A-1} \qquad (7)$$

$$S_X^2 = \left(\frac{A}{A-1}\right)^2 \left[ \sum_{i=1}^{n_X} c_i - \sum_{i=1}^{n_X} c_i^{\,2} \right] \qquad (8)$$

where,

$M_X$ = the mean of Form X scores below chance level,  
$S_X^2$ = the variance of Form X scores below chance level,  
A = the number of choices per item, and  
$c_i$ = the psuedo guessing parameter for item i

ii. Equations 7 and 8 are repeated to determine $M_Y$ ( the mean of Form Y scores below chance level) and $S_Y^2$ ( the variance of Form Y scores below chance level).

iii. Linear parameters for equating Form X scores below chance level to Form Y scores below chance level are determined as follows:

$$A = \frac{\sqrt{S_Y^2}}{\sqrt{S_X^2}} \qquad (9)$$

$$B = M_Y - A\, M_X \qquad (10)$$

iv. This procedure may also be used to determine the linear parameters for equating number right true scores below chance level. In this case,

$$M_X = \sum_{i=1}^{n_X} c_i \text{ and } S_X^2 = \sum_{i=1}^{n_X} c_i - \sum_{i=1}^{n_X} c_i^{\,2}$$

2. The equipercentile equating of estimated number right observed scores also has a disadvantage in that conversions obtained from this type of equating may not be applicable, in the strictest sense, to observed formula scores. (Note that it

is not possible to generate an estimated observed formula score frequency distribution.)

D. Using Rasch θ Estimates to Equate Actual Observed Scores

Because of the special nature of the Rasch model (total score is a sufficient statistic for estimating ability, a monotonic relationship exists between raw score and estimated ability), it is possible to use the results of IRT parameter estimation to directly equate actual number right observed scores. This is the procedure used by Rentz and Bashaw (1975) in performing the raw score to raw score equatings when applying the Rasch model to the Anchor Test Study data. It was also used by Loyd and Hoover (1980). The steps listed below are synthesized from Rentz and Bashaw's (1975) procedure:

1. It should first be noted that a conversion or scoring table is standard output from a Rasch parameter estimation program. This table lists for every obtained raw score on the test the corresponding ability estimate $\hat{\theta}$.

2. For the two tests to be equated (X and Y), there will be two conversion tables, one relating x to $\hat{\theta}_x$ and the other y to $\hat{\theta}_y$. Suppose further that one of the parameter scaling methods has been used to obtain $\hat{\theta}_x{}^*$, which is now on the same scale as $\hat{\theta}_y$. (Y is the base test.)

3. For each possible score $y_j$, find the score $x_i$ such that $\hat{\theta}_{y_j} - \hat{\theta}_{x_i}^*$ is a minimum.

4. The score $x_i$ that minimizes $\hat{\theta}_{y_j} - \hat{\theta}_{x_i}^*$ is the equivalent score of $y_j$.

There is a problem with the above procedure which results in what Rentz and Bashaw (1975) refer to as "assignment error". Assignment error occurs when it is necessary to assign an examinee a raw score on the equated test that is most equivalent to a raw score on the base test. Suppose, for instance, that on the base test a raw score of 10 corresponded to a $\hat{\theta}$ of 2.0, and on the equated test a raw score of 9 corresponded to a $\hat{\theta}$ of 1.9 and a raw score of 10 to a $\hat{\theta}$ of 2.2. In this case, a raw score of 9 on the equated test would be taken to be equivalent to a raw score of 10 on the base test because the $\hat{\theta}$ of 1.9 is closest to 2.0. The assignment error would be the difference in these $\hat{\theta}$'s. Obviously, because of the discrete nature of raw scores, longer tests, having more raw scores and $\hat{\theta}$'s, will result in fewer assignment errors. However, the same sort of problem would occur if there were missing data (raw scores). This speaks to the need for an adequate sample of examinees whose abilities will cover the range of possible raw scores on the two test forms if this sort of equating is to be considered.