Comparison of IRT True-Score and Equipercentile Observed-Score "Equatings"

Frederic M. Lord and Marilyn S. Wingersky
Educational Testing Service

Two methods of 'equating' tests are compared, one using true scores, the other using equipercentile equating of observed scores. The theory of equating is discussed. For the data studied, the two methods yield almost indistinguishable results.

Most item response theory (IRT) equating is currently attempted by the true-score equating procedure described in Lord (1980, chap. 13). Lord also described an IRT equipercentile observed-score procedure, which until now seems to have been little used in operational work, perhaps because it is more complicated and more expensive than the true-score procedure. This paper discusses theoretical considerations and reports an empirical research study comparing the results of applying these two procedures to real test data. Note that IRT plays only a subsidiary role in observed-score equipercentile equating; similar results could be expected for conventional equipercentile equating, assuming that the IRT model holds.

Kolen (1981) found the equipercentile observed-score IRT procedure to be one of the better of nine procedures compared in his empirical research study. However, his criterion was stability in cross-validation. Although stability is certainly desirable, stability is not a proper criterion for choosing the best equating method: Incorrect equating procedures may yield more stable results than correct procedures.

Sections 1 and 2 outline the true-score procedure and the observed-score equipercentile procedure, respectively. Section 3 discusses the theoretical advantages and disadvantages of each procedure. Section 4 describes the real test data used to provide a comparison of the two methods. Section 5 describes the procedures used for estimating item and ability parameters. Section 6 reports and summarizes the empirical results.

IRT models the probability of a correct response by an examinee to a test item as a monotonically increasing function of ability. The model used here is the three-parameter logistic model given by

$$P_i(\theta_a) = c_i + (1 - c_i)/\{1 + \exp[-1.7a_i(\theta_a - b_i)]\} \quad , \tag{1}$$

where $P_i(\theta_a)$ is the probability of examinee a answering item i correctly,
b_i is the difficulty of item i,
a_i is the discrimination index for item i,

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 8, No. 4, Fall 1984, pp. 453–461

0146-6216/84/040453-09$1.70

c_i is the lower asymptote for item i, and
θ_a is the ability of examinee a ($-\infty < \theta_a < \infty$).

$P_i(\theta_a)$ has a minimum of c_i and a maximum of 1. This model assumes that the test is unidimensional.

1. True-Score Equating

Since the expected score of examinee a on item i is $P_i(\theta_a)$, the examinee's expected number of correct answers is $\sum_i P_i(\theta_a)$. In classical test theory, this expectation is called the (number-correct) *true score*, $\xi_a \equiv \sum_i P_i(\theta_a)$. For the moment, there is no concern with the scores of particular examinees, so the subscript a is dropped. Here the true score for Test X containing n items is the mathematical variable

$$\xi \equiv \sum_{i=1}^{n} P_i(\theta) \quad , \tag{2}$$

a monotonic increasing function of θ. If Test Y contains m items and measures the same ability θ as Test X, the true score on Test Y is the mathematical variable

$$\eta \equiv \sum_{j=1}^{m} P_j(\theta) \quad . \tag{3}$$

The variables ξ, η, and θ are all measures of the same psychological trait; they differ only in the numerical scale on which the measurements are expressed. Thus, true scores $\xi = \xi_0$ and $\eta = \eta_0$ corresponding to any given $\theta = \theta_0$ represent identical levels of ability. Any examinee whose true score on Test X is ξ_0 must automatically have a true score on Test Y of exactly η_0, provided that the IRT model holds. The situation is the same as when it is said that 32° Farenheit has the same meaning as 0° Celsius, except that these temperature scales have a linear relationship, whereas the true-score scales have a nonlinear relationship. Thus, ξ_0 and η_0 are equated true scores; this is true in a much stronger sense than is usually implied by the term *equated*.

In IRT true-score equating, estimated item parameters are substituted into Equations 2 and 3 and a table of corresponding values of ξ and η is calculated. This constitutes the true-score equating table. This table is then applied in practice as if the true scores were observed number-correct scores. Since observed scores have different properties than true scores, this last step has no clear theoretical justification. It is done as a practical procedure, to be justified only by whatever usefulness and reasonableness can be empirically demonstrated for the results.

2. IRT Equipercentile Observed-Score Equating

If the assumptions of IRT hold (as is assumed throughout), the probability that an examinee of ability θ will have a number-correct score on a two-item test of $x = 0$ is Q_1Q_2, of $x = 2$ is P_1P_2, where $P_i \equiv P_i(\theta)$ and $Q_i \equiv 1 - P_i$; the probability that this examinee's score is $x = 1$ is $P_1Q_2 + Q_1P_2$. These three probabilities constitute the conditional frequency distribution $f_2(x|\theta)$.

If a third item is added to this test, the distribution of x is now:

$$f_3(x|\theta) = Q_3 f_2(x|\theta) + P_3 f_2(x-1|\theta) \quad (x = 0, 1, \ldots, 3) \quad , \tag{4}$$

where $f_r(x|\theta) = 0$ if $x < 0$ or $x > r$. Using this recursive procedure, a computer can readily determine $f_n(x|\theta)$, even for an n of several hundred items.

If the θ of each examinee is known, the (marginal) distribution of x for a group of N examinees is

$$\frac{1}{N} \sum_{a=1}^{N} f_n(x|\theta_a) \quad . \tag{5}$$

If an m-item Test Y yields number-correct score y and measures the same ability as Test X, then the (marginal) distribution of y for a group of M examinees is

$$\frac{1}{M}\sum_{b=1}^{M} f_m(y|\theta_b) \quad . \tag{6}$$

A monotonic transformation of the y scores can now be found from Equations 5 and 6 such that the distribution of the transformed y scores is the same as the distribution of the (untransformed) x scores, except for irregularities due to the fact that x and y can only assume integer values. This is done by finding, for each y score, the x that has the same percentile rank in Equation 5 that y has in Equation 6. The x so found is the desired transformed y score.

If the examinees who took Test Y have the same distribution of θ as the examinees who took Test X, then the resulting transformation of y is an "equipercentile equating" of the y scale to the x scale. *Within groups similar to the groups used to derive the transformation,* it has the valuable property that if a cutting score is chosen on the x scale and the same cutting score is used on the transformed y scale, the proportion of Test X examinees selected will be the same as the proportion of Test Y examinees selected. This property is essential if Test X and Test Y examinees are both to be treated equitably, so that an examinee cannot complain that he/she was unfairly measured by the choice of test administered.

When the groups taking Tests X and Y are known to have approximately the same distribution of θ (e.g., they are two random samples from the same population), there is no reason to use IRT equating. It is much simpler to do the equipercentile equating using the actual sample distributions of x and y, instead of Equations 5 and 6. The need for IRT arises when the ability distributions of the two groups may differ. In this case, IRT may allow estimation of the (marginal) frequency distributions of number-correct scores that would have resulted if all examinees had taken both tests, without practice or fatigue effects.

In order to do this, the item and ability parameters in Equations 5 and 6 must all be on the same scale. This is usually accomplished by administering a suitable "anchor test" to both groups of examinees. All answer-sheet responses for both groups are used in a single computer run that estimates all the IRT parameters on the same scale. These estimates are then used in Equations 5 and 6, substituting $N + M$ for N or M, to obtain the distributions of x and y for the combined group of $N + M$ examinees. Equipercentile equating of y to x is then carried out in the usual way. The use of distributions estimated by IRT has an additional advantage here: It smooths out the irregularities found in all sample distributions, which cause serious sampling errors in equipercentile equatings done with unsmoothed sample distributions.

3. Theoretical Perspectives

Practical workers, with the need for equating scores on two different test forms, have used over the years widely different methods (see Angoff, 1971) in an attempt to approximate the desired result. Each practical worker, needing a word to describe his/her results, asserts that he/she has produced an equating of y to x. Yet different methods and different groups do not produce identical "equatings."

Braun and Holland (1982, p. 14) stated, "There is some disagreement over what test equating is and the proper method for doing it." They then adopted the definition "Form-X and Form-Y are equated on [population] P" if the distribution of the transformed y scores in population P is the same as the distribution of the (untransformed) x scores.

This definition of the phrase "equated on population P" is beyond reproach. One problem, however, is that the qualifying phrase "on population P" is typically dropped by the practical worker who writes a research report or publishes an equating table in a test manual.

Unfortunately (as will be shown below), two tests that are equated on population P will typically

not be equated for various subpopulations that are included in P. Test scores that are equated for the population of college applicants may not be equated either for the population of female college applicants or for the population of male college applicants. The scores are still less likely to be equated for a subpopulation characterized by interest in science, or in music. For the subpopulation of Harvard applicants, the situation is worse.

If the proportion of applicants admitted to Harvard differs significantly depending on whether they were given Form X or Form Y of the test, it is clear that the "equating" was unsuccessful. Since similar inequalities are likely to characterize any equating on any specified population, it may be best not to say that the tests are "equated" at all, or to simply say that they are "approximately equated."

From a practical point of view, the approximate equating may be quite satisfactory for many subgroups. This is especially true if the two tests to be equated are already nearly parallel. If the two tests are distinctly different from each other in difficulty or reliability, however, it is unlikely that the equating will be adequate for a subpopulation having a mean and variance of ability that is sharply different from the mean and variance of the total population used to derive the equating transformation. Extensive practical data illustrating the many adequacies and some of the inadequacies of approximate equatings are given in the 30-volume *Anchor Test Study* (Loret, Seder, Bianchini, & Vale, 1974).

For a theoretical discussion of alternative equating methods, it is important not to begin with a definition of equating that is clearly inadequate for subpopulations of examinees. Given that the IRT model holds, IRT observed-score equating would, for example, be automatically endorsed by the Braun and Holland (1982) definition, since their definition mandates equipercentile equating. IRT true-score equating would be definitely rejected by their definition, since in general it will not lead to observed scores x and transformed observed scores y having the same frequency distribution, unless Tests X and Y are strictly parallel forms that are identical in difficulty, in reliability, and also in most other respects.

The definition to be used here for "equating" *without* the qualifying phrase "on a given population" is that for any group of examinees all at a given ability level (θ or ξ or η), the distribution of scores after equating is the same regardless of which test was administered. If the model holds, then such equating can be achieved for true scores to any desired degree of accuracy, provided that the sample size is large enough. For observed scores, such equating can only be approximated, regardless of sample size, unless Tests X and Y are strictly parallel tests.

The important virtue of IRT true-score equating is that if the IRT model holds, the true scores are clearly equated for all subpopulations of examinees. This results from the invariance of IRT parameters across populations of examinees, assumed by the IRT model. The clear flaw in any practical use of IRT true-score equating is that it equates true scores, not the actually observed fallible scores. Treating observed scores as if they were true scores cannot be justified on any theoretical grounds.

The virtue of IRT observed-score equating is that in a group like that used to derive the equating, any cutting score will accept the same percentage of examinees regardless of the test administered. The flaw is that this usually holds precisely only for that total group and not for other groups or subgroups.

This last statement is most clearly seen from a very extreme example. Suppose Forms X and Y have the same number of items and measure the same ability, θ, but differ in difficulty. If the equipercentile equating is carried out on a group of examinees who all are guessing at random on almost all the items, then the difference in difficulty between the two forms will not manifest itself and any equipercentile equating will approximate an identity transformation of score y (equated y scores are identical to raw y scores). If a slightly more competent group of examinees is used for the equipercentile equating, however, the difference in difficulty between forms will begin to become apparent and most y scores will be adjusted upwards or downwards accordingly. As the competence of the group used becomes higher and higher, the equating transformation found will differ more and more from the identity transformation found from the original extreme group.

As a second example of the theoretical invalidity of observed-score equating, suppose that Tests X and Y are of equal difficulty and that the true scores ξ and η have equal variance, but that y is much less reliable than x. Consider a subgroup of very talented examinees; to make the illustration clear, consider that in this subgroup all examinees have nearly identical θ values. Most of the variation in observed scores x and y is now due to errors of measurement. The equipercentile equating transformation found will thus approximate a straight line with slope

$$\frac{\text{standard deviation of the errors of measurement in } x}{\text{standard deviation of the errors of measurement in } y} \tag{7}$$

Since y is much less reliable than x, the slope will be much less than 1.0.

If, on the other hand, the equipercentile equating transformation is found from a group in which the true-score variance is large compared to the error variances, the transformation will tend to approximate a straight line with slope

$$\frac{\text{standard deviation of true scores on } x}{\text{standard deviation of true scores on } y}, \tag{8}$$

which is approximately 1.0. Intermediate situations will provide transformations with intermediate slopes. If the wrong equating is applied to any given subpopulation, then the population of examinees in the subpopulation accepted will depend on whether the examinees took Test X or Test Y—an inequitable result.

The theoretical position, then, is that each of the two methods described in Sections 1 and 2 (as well as all other available equating methods) has its own inadequacies. Since, in practice, some (approximate) equating method must be used, it will be informative to investigate empirically how the two methods of Section 2 compare in a specially contrived practical situation in which the correct equating is actually known in advance.

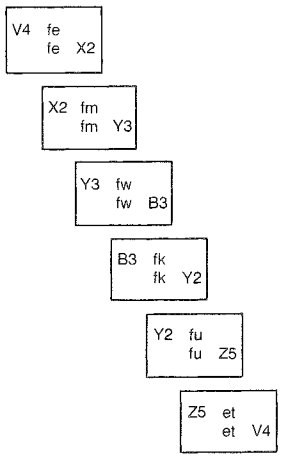
4. Data

The two equating methods were used to equate the chain of six Scholastic Aptitude Test (SAT) verbal tests described by Petersen, Cook, and Stocking (1983). The tests in this chain were selected so that the first test and the last test are the same. Different columns represent different tests, different rows represent different groups of examinees. Each test is equated to the next test in the chain using an anchor test. Figure 1 is a diagram of the chain. The capital letters represent the test form, the small letters represent the anchor test. Scores on Form V4 are equated to scores on Form X2 using Anchor Test fe. These equated scores on Form X2 are equated to scores on Form Y3 using Anchor Test fm. This gives an equating of Form V4 to Y3. In this manner, proceeding through the chain continues until the final equating of Z5 to V4, giving a table showing scores on the original V4 equated to the scores on V4 at the end of the chain. Any deviation from equality between the two sets of scores in this table could be attributed to scale drift or lack of model fit.

It is sometimes objected that a procedure that successfully equates a test to itself may be neither theoretically valid nor satisfactory in practice for equating one test to a different test. This is true. However, an investigation that equates a test to itself is still definitely useful, since any procedure that fails to yield approximately the required identity transformation can definitely be removed from the list of possible equating procedures.

Each form in the chain studied has 85 items except Form V4, which has 90 items. Each anchor test has 40 items. For each form there are two groups of examinees, each group taking a different anchor test. The two groups taking each form were random samples from the same population for all of the

Figure 1
Chain of Six SAT Verbal Equatings
(Upper Case Letters Designate Test Forms,
Lower Case Letters Designate Anchor Tests)



forms except Y3. For each parameter estimation run, a random sample of approximately 2,670 examinees was selected from the data obtained at the test administration of that form and anchor test. The maximum difference in mean score between two samples taking the same anchor test is .25 of the within-group standard deviation. (For further details about the data, see Petersen et al., 1983.)

5. Parameter Calibration

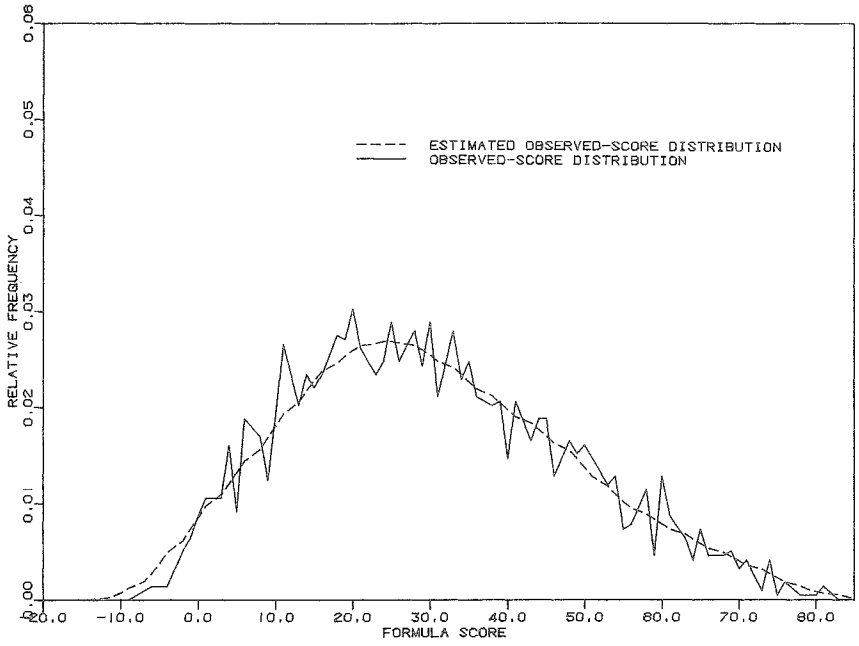
The item parameters and abilities were estimated by a modified version of the computer program LOGIST (Wood, Wingersky, & Lord, 1976) in six separate calibration runs. In Figure 1, each box (containing two forms and one anchor test) represents one LOGIST run. The item responses for items not taken by an examinee, such as the Form X2 items for examinees taking Form V4 in the first box, were treated as not-reached items.

All of the estimated parameters within each LOGIST run are on the same scale, and either method of equating can be used to equate the scores for the two tests. The anchor tests were not used directly in the equating but were used in LOGIST so that all the estimated parameters within a LOGIST run would be on the same scale.

6. Results

In using the IRT observed-score equating method, two estimated distributions of observed scores were equated so that the transformed y scores and the (untransformed) x scores would have the same distribution. Figure 2 demonstrates, at least for one test, that this estimated distribution of observed scores is a reasonable fit to the actual distribution of observed scores. The frequencies are plotted against formula scores, which are the number correct minus a fraction of the number incorrect; the fraction is the reciprocal of the number of choices minus one. Since the estimated observed-score distribution can only be obtained for number-correct scores, the transformation to formula scores assumes that there are no omits, that is,

Figure 2
Comparison of Distribution of Observed Scores and Estimated Distribution of Observed Scores—SAT Verbal—No Omits



that the number incorrect is the total number of items minus the number correct. In order to compare the two distributions, the observed-score distribution should be based on a group that has no omits. Consequently, a form of the SAT verbal test different from those in the chain was used for Figure 2 in order to get a sufficiently large enough sample for the frequency distribution and for the item calibration.

The agreement shown in Figure 2 is good except that the tails of the estimated distribution are too high. This discrepancy is presumably due to the use of estimated θ in place of true θ for the practical implementation of Equation 5. Since a similar discrepancy affects the estimated observed-score distributions of both Test X and Test Y, the effects of the discrepancies tend to cancel out in the equating process.

In this chain-equating study, each method of equating was applied separately to the whole chain of equatings, resulting in a line for each method equating Form V4 at the beginning of the chain to Form V4 at the end of the chain. These two lines are plotted in Figure 3 along with a 45° line. The solid line is the IRT true-score equating line; the dotted line, falling practically on top of the solid line, is the IRT observed-score equating line. To equate scores below "chance level," the method described in Lord (1980, pp. 210–211) was used for the IRT true-score line. For formula scores above 0, the maximum difference between the two equatings was .2; for scores below 0, the maximum difference was .8, which occurred at the chance level. If the equating methods were perfect and there was no scale drift, the equating line would be the dashed 45° line.

Figure 4 shows the two equating methods applied to one individual link in the chain. This particular link was selected because the IRT true-score equating line between these two forms had the greatest discontinuity in the slope at the chance level. The largest difference between the two lines occurred at

Figure 3
Comparison of True-Score Equating and Estimated Observed-Score Equating Over the Chain of Six Equatings

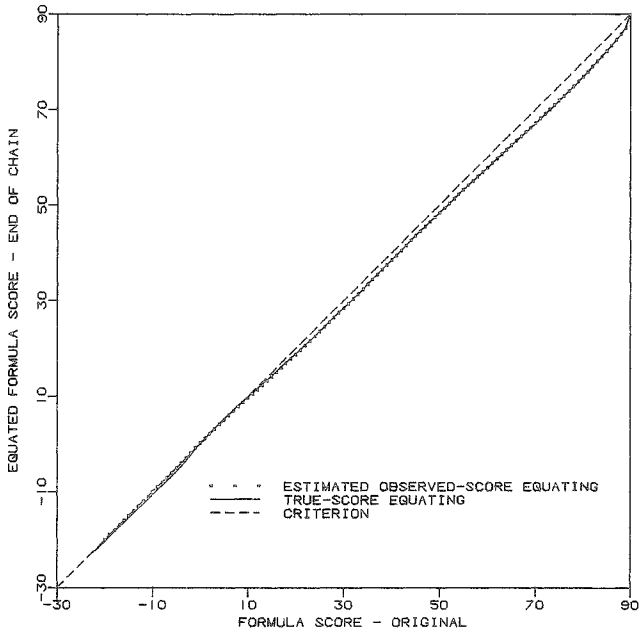
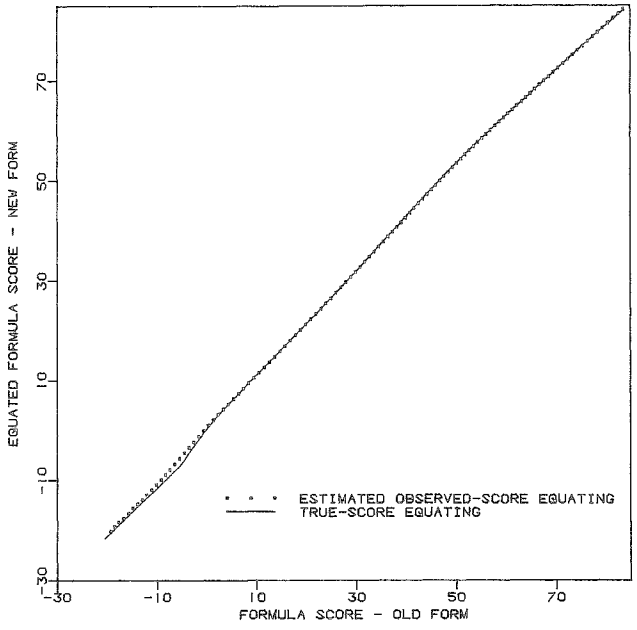


Figure 4
Comparison of True-Score Equating and Estimated Observed-Score Equating for the Single Link in Chain



the chance level and was 1.6. For scores above 0, the maximum difference between the two lines was .4.

Conclusions

Given that there is no clear theoretical justification for applying IRT true-score equating to observed scores and that the equipercentile equating of observed-score distributions is population dependent, the close agreement between the two equating lines is reassuring. In situations when anchor tests are not used, a similar close agreement would be expected between IRT true-score equating and conventional equipercentile equating of observed scores, provided that the IRT model holds and provided that the distributions used for the equipercentile equating are smoothed by an appropriate smoothing method. Conventional (non-IRT) equipercentile equating of observed scores is not recommended in situations when anchor tests are required.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington DC: American Council on Education.

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9–49). New York: Academic Press.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement, 18,* 1–11.

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

Loret, P. G., Seder, A., Bianchini, J. C., & Vale, C. A. (1974). *Anchor test study—Equivalence and norms tables for selected reading achievement tests (grades 4, 5, 6)*. Washington DC: U. S. Government Printing Office.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8,* 137–156.

Wood, R. L., Wingersky, M. S., & Lord, F. M. (1976). *LOGIST—A computer program for estimating examinee ability and item characteristic curve parameters* (ETS RM 76–6). Princeton NJ: Educational Testing Service.

Acknowledgment

This work was supported in part by contract N00014-80-C-0402, NR 150-453, between the Office of Naval Research and Educational Testing Service.

Author's Address

Send requests for reprints or further information to Frederic M. Lord, Educational Testing Service, Princeton NJ 08541, U.S.A.