Model-Free Evaluation of Equating and Scaling

D. R. Divgi
The University of Iowa

Standardized tests are equated and scaled in order that scores on different tests can be compared. If one test yields higher expected scaled scores than another, the scale is biased against those who take the latter test. The amount of bias, defined as the difference between expected values, depends on ability. This paper presents two methods for estimating this relationship and the bias in the scale, using a predictor as the measure of ability. The resulting evaluation is absolute in the sense that the scale is judged according to its own properties and not by comparison with an arbitrarily designated criterion scale. Moreover, there is no need to assume a particular theoretical model to be correct. An application of the methods showed that the Rasch model is not suitable for vertical equating of multiple-choice tests.

Any large-scale program of standardized testing involves a number of tests that measure the same skill. In order to allow comparisons of scores on different tests, the tests are equated and raw scores are transformed onto a common scale. Although many different methods are available for equating and scaling (Angoff, 1971; Lord, 1977), there is no generally accepted method for evaluating the results. Slinde and Linn (1977) have pointed out large discrepancies between score equivalences provided by test publishers and by the Anchor Test Study. Marco, Petersen, and Stewart (1979) judged the results of an equating procedure by comparing them with results from two "criterion equatings." Their tables show that the two criterion equatings were often substantially different, despite being obtained from the same sample. Obviously, both cannot be "correct" or even "best" in any sense. This supports Lord's (1977) statement that "the results obtained by conventional methods cannot be justified as a criterion" (p. 132). In the absence of a commonly accepted standard, there is no logical basis for comparative evaluations.

Equating Bias

An absolute evaluation of an equating procedure can be sought by comparing its results with the ideal. According to Lord (1977), "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 [she] is to take test X or test Y" (p. 128). If the tests differ in difficulty, the easy test is more reliable at low abilities and the difficult test at high abilities. Therefore, although the (curvilinear) relationship between true scores can be determined, observed scores cannot be equated (Lord, 1977, p. 128).

If Tests X and Y cannot be equally reliable at each ability level, it may be asked that the scale at least be unbiased. In other words, an ex-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 5, No. 2, Spring 1981, pp. 203–208
© Copyright 1981 Applied Psychological Measurement Inc.
0146-6216/81/020203-06$1.30

aminee's expected score on the common scale should not depend on the test taken. Wright (1968) has used this as the criterion of "item-free measurement." Differences between means are affected by nonlinear transformations. The score generally used to make decisions is not y^* or x but the transformed value on the common scale. With the same equivalence relations between raw scores, different transformations yield scales that differ not only in numerical properties but even in meaning (Echternacht, 1977; Gardner, 1962). Differences that look equal on one scale may look vastly different on another (see Table 1 in Slinde & Linn, 1977, for differences between comparisons based on scaled scores and on grade equivalents). Therefore, any evaluation ought to be based on scores on the common scale and not on raw scores. If two scales are provided for the same test battery, each should be evaluated separately.

If the tests are almost equally difficult, the transformations of x and y into scaled scores will be very similar, and little bias will be found. It is in vertical equating that problems arise. Therefore, assume that Test X is appreciably more difficult than Test Y. In addition, purely to simplify the following argument, let both tests contain the same number of items n. All scales used in practice are such that a particular value of x corresponds to a higher scaled score than the same value of y. At the same time, a given examinee tends to get higher raw scores on Test Y than on Test X. These two facts tend to cancel out and make expected scaled scores equal at medium ability levels (provided the scale is well constructed). Persons with very low or very high abilities get almost equal raw scores on both tests, the former by guessing and the latter due to ceiling effects. In such cases, Test X will yield higher scaled scores than Test Y, and the scale is biased against those who take Test Y (for an illustration see Divgi, 1980). Thus, bias is a function of ability.

Method

The dependence of bias on ability can be calculated theoretically by using a latent trait model. Once a model is selected, distributions of x and y can be calculated at any ability, θ, and hence the mean, the error variance, and the information function for the corresponding scaled scores. Divgi (1980) used the three-parameter logistic model because the one- and two-parameter models make additional assumptions that are unlikely to hold for real multiple-choice tests. Wright (1977a) has argued, however, that the parameters of the three-parameter model cannot be estimated. Thus, no universally accepted model is available.

If the use of a model is to be avoided, bias can be estimated using real data in which scores on Tests X and Y are available for each person. As true abilities are unknown, it is, of course, impossible to form a group of persons with equal abilities. An observed quantity that is correlated with ability must be used, which will be denoted by z in the present paper. In a group of persons with the same value of z, different individuals will have different abilities, and the mean bias will be the average over this distribution of ability. This averaging makes bias look less variable than it is; any observed minimum will be larger than the true value, and any observed maximum will be smaller. Such blurring increases with the conditional variance of ability at given z. Therefore, z should be chosen to correlate as strongly as possible with what the tests measure. Thus, if Tests X and Y are reading tests, z might be a mathematics or a general aptitude score. If two or more predictors are available, it is advisable to form z by multiple regression of $(x + y)$ on the complete set of predictors. This will yield smaller conditional variance than the use of any single predictor. (Scores x and y should not be used to form ability groups. Spurious effects are likely if the same measurement errors affect formation of subgroups and subsequent analysis within these subgroups. See Gustafsson, 1979.) A single value or a small range of values of z can be used to define a group that is fairly homogeneous in ability. Mean bias is calculated separately in each group. Its relationship to mean z in the group can be displayed in a table or a graph. If the scale is unbiased, expected bias vanishes at every ability. Then, mean bias in each group will be zero except for random error.

The primary drawback of the above method is that it requires a very large sample. Each group should be large enough to yield a reasonably small standard error of the mean. At the same time, it is desirable that the variation of ability and hence of z be as small as possible within any single group. This means that the number of groups should be large. The need for a very large sample can be avoided by using a regression approach. Let $S_X(x)$ and $S_Y(y)$ be scaled scores corresponding to raw scores x and y, respectively. The scaled scores can be plotted against z. If the scale is biased, there will be systematic differences between the two scatterplots. However, the scatter will make the differences difficult to see and to quantify. It is far more convenient to examine bias directly. Therefore, calculate $S_X(x) - S_Y(y)$ for each person. Fit ,it with a polynomial of z. Start with a quadratic function and add more terms to the polynomial until improvement in fit becomes nonsignificant by statistical or personal criteria. A quadratic fit was quite satisfactory for the Rasch scale used in the present study.

Illustration

The methods described above were used to study vertical equating with the Rasch model. (For details see, e.g., Rentz & Bashaw, 1977; Wright, 1977a.) According to the Rasch model, the probability that a person with ability θ will answer item g correctly is

$$P_g(\theta) = 1/[1 + \exp(b_g - \theta)] \qquad [1]$$

where b_g is the difficulty parameter of item g. Therefore, $P_g(\theta) \rightarrow 0$ when $\theta \rightarrow -\infty$. Thus, the probability of a correct answer by pure guesswork is assumed to be zero. Moreover, the slope $dP_g(\theta)/d\theta$ at $\theta = b_g$ is assumed to be the same for every item, i.e., all items are assumed equal in discriminating power. For a person with number-correct score x, the maximum likelihood estimate $\hat{\theta}$ is the value that satisfies

$$\sum_{g \in X} P_g(\hat{\theta}) = x \; . \qquad [2]$$

The estimated ability, or a linear transformation of it, is used as the Rasch scaled score (Rentz & Bashaw, 1977). Equation 2 cannot be solved if x equals zero or if the number of items equals n. However, a test publisher must provide a scaled score corresponding to every possible raw score. In this study the scaled score at $x = n$ was obtained by quadratic extrapolation of values at $x = n-3$, $n-2$, and $n-1$. Similarly, the scaled scores at $x = 3$, 2, and 1 were extrapolated to $x = 0$.

Data

Item data were taken from the Reading test of the Survey battery of the 1978 Metropolitan Achievement Tests, Intermediate Level, Form J (Prescott, Balow, Hogan, & Farr, 1978). Rasch calibration was carried out with the UCON procedure with a correction factor for inconsistency (Wright & Douglas, 1977). All 60 items were calibrated together, with a sample of 2,000 examinees. The goodness-of-fit test recommended by Rentz and Rentz (1979) identified only six items as nonfitting. These were retained during later analyses. The test was divided into Difficult and Easy subtests of 30 items each, which were used as Tests X and Y, respectively. Their mean difficulties differed by 1.53 logits (i.e., log-odds units). This difference is large but not unreasonable. It is smaller than the largest difference between successive levels reported by Rentz and Bashaw (1977). Estimated Rasch ability was used as the scaled score. The "predictor" z was the predicted total Reading score obtained by multiple regression on Mathematics, Language, Science, Social Science, and Otis-Lennon raw scores. The squared multiple correlation was .78.

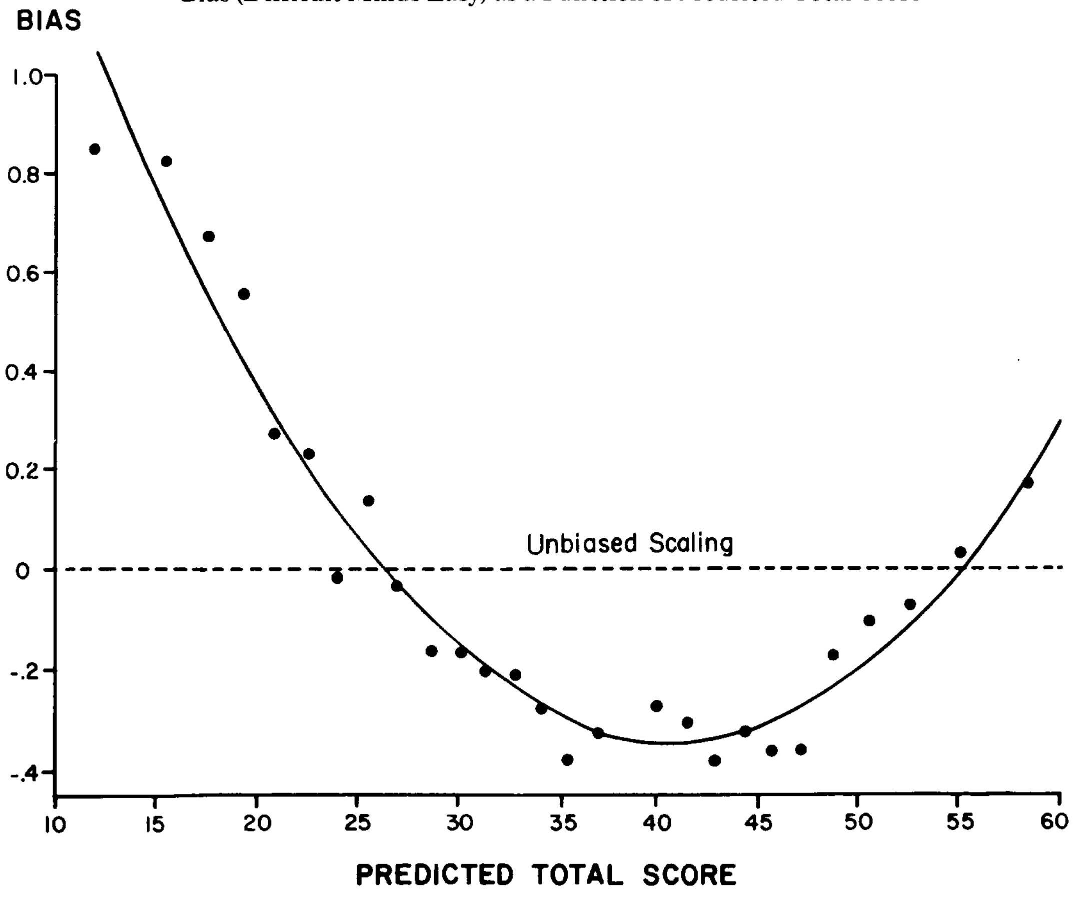
The subtest scores x, y, and predictor score z were calculated for all available cases in the Fall Standardization data set (N = 5,512). The records were sorted in ascending order of z. Then, starting from smallest z, each group of 200 successive cases was considered to be an ability group. Thus, there were 27 groups of 200 cases each and one of 112 cases.

Results

The mean differences between scaled scores (Difficult minus Easy) in these groups are shown in Figure 1. Standard errors of means ranged between .063 and .079 logit. The mean and standard deviation over the entire sample were −.037 and 1.06 logits, respectively. The difference $S_X(x)$–$S_Y(y)$ was regressed on z and z^2. The F ratio for regression was 354. The fitted curve is shown in Figure 1. The results show that at low and at very high abilities the Rasch scale favors those who take the difficult test; at medium abilities it favors those who take the easy test.

If the Rasch model provided "item-free measurement" (Wright, 1968), the difference would have been zero (apart from random error) at all values of z. The F ratio is too large for the difference between the model and reality to be due to chance alone. The minimum of the fitted curve is −.35 logit and its value at the mean predictor score in the first group is 1.04 logits. The positive differences at low abilities show that the examinees scored higher on the difficult test than the Rasch model predicted, i.e., that they increased their scores by guessing. The negative values at medium abilities arise from trying to fit the Rasch curve to item characteristic curves that do not fit the Rasch model.

Figure 1
Bias (Difficult Minus Easy) as a Function of Predicted Total Score



The interpretation of Figure 1 was supported by results from the following simple model. Assume all items to have the same guessing parameter c so that Equation 1 is replaced by

$$P_g(\theta) = c+(1-c)/[1+\exp(b_g-\theta)]. \quad [3]$$

All difficult items have the same difficulty b_D and all easy items have difficulty b_E. Items are calibrated using a population in which everyone has ability $\theta = 0$ and using a very long test containing equal numbers of easy and difficult items. Then, the variance of proportion-correct scores is very small; hence, ability estimates for all examinees are practically equal. Mean scores p_E and p_D on easy and difficult items are obtained by using b_E and b_D in Equation 3 with $\theta = 0$. Then, the difference between difficulty estimates is

$$\hat{b}_D-\hat{b}_E = \ln[p_E(1-p_D)/p_D(1-p_E)]. \quad [4]$$

The Easy and Difficult subtests consist purely of easy and difficult items, respectively, and contain equal numbers of items, n. The Rasch ability estimate $\hat{\theta}_D$ corresponding to score x on the Difficult subtest is $\hat{b}_D + \ln[x/(n-x)]$, with quadratic extrapolation used to define $\hat{\theta}_D$ at $x = 0$ and at n; similarly, for $\hat{\theta}_E$. At any given ability, θ, the distribution of raw scores on a subtest is given by the binomial distribution. Therefore, it is easy to calculate the bias $E(\hat{\theta}_D|\theta) - E(\hat{\theta}_E|\theta)$. Calculations were carried out with $c = .2$, $b_D-b_E = 1.5$ (whence $\hat{b}_D-\hat{b}_E = 1.24$), and $n = 30$. The bias was .90 at $\theta = -3.0$, zero near $\theta = -.25$ and 3.0, and $-.36$ at $\theta = 1.75$. This relationship between bias and ability is very similar to that in Figure 1, thus confirming the interpretation of the empirical results.

Discussion

The two procedures described here are useful because they do not require the assumption that a particular theoretical model is correct. The primary drawback is that both test scores must be available for each person, which is rarely the case.[1] The methods can be modified for a situation where Tests X and Y are taken by different random samples, but such designs are also rare. Therefore, the procedures are useful mainly in studies of methods for equating and scaling, where long tests are divided into subtests, which are then equated (e.g., Marco et al., 1979). In contrast to the approach of Marco et al., the results are absolute in the sense that each method is judged according to its own consequences and not by comparison with another method.

Of the two procedures, the regression approach is by far the more convenient. The least squares fit of parameters assumes that the residual variance is the same at all values of the predictor. Departures from this assumption were minor in the present study. Over the 28 groups, within-group standard deviations of $S_X(x) - S_Y(y)$ ranged from .89 to 1.12 logits. Figure 1 shows satisfactory agreement between the quadratic curve fitted by unweighted least squares and group means. In any case, weighted least squares can be used if necessary. The use of a polynomial is not a restrictive assumption, since no a priori limit is placed on the degree of the polynomial.

Although an evaluation of the Rasch model was not the primary topic of this study, the results provide strong evidence against using the Rasch model for vertical equating. Wright (1968) split a reading test into difficult and easy parts, calculated the difference between ability estimates for each person, and divided it by its (estimated) standard deviation. He found the sample mean of this standardized difference to be almost zero and hence claimed that he had demonstrated "item-free person measurement" with the Rasch model. Actually, what he had shown to be item-free was not measurement on a single examinee but the mean over a large sample containing a wide range of ability. Figure 1 shows that individual ability estimates based on

[1] An important exception is the Anchor Test Study, whose data were used by Rentz and Bashaw (1977). An evaluation of Rentz and Bashaw's National Reference Scale using the present approach would be very interesting.

difficult and easy tests differ systematically and that these differences can be quite large.[2] Slinde and Linn (1979) demonstrated differences between difficult and easy tests and attributed them to effects of guessing. Their design was complicated because they formed high-, middle-, and low-ability groups and estimated item parameters separately in each group. The results in Figure 1 are much more detailed and are based on a simpler design. Yet they lead to the same conclusion as that of Slinde and Linn (1979). The Rasch model is not useful for vertical equating of multiple-choice tests.

References

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

Divgi, D. R. *Evaluation of scales for standardized tests.* Paper presented at the annual meeting of the American Educational Research Association, April 1980.

Echternacht, G. Grade equivalent scores. *Measurement in Education,* 1977, *8*(2), 1-4.

Gardner, E. F. Normative standard scores. *Educational and Psychological Measurement,* 1962, *22,* 7-14.

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.

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

Marco, G. L., Petersen, N. S., & Stewart, E. E. *A test of the adequacy of curvilinear score equating models.* In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference.* Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory, 1980, 167-196.

Prescott, G. A., Balow, I. H., Hogan, T. P., & Farr, R. C. *The Metropolitan Achievement Tests.* New York: The Psychological Corporation, 1978.

Rentz, R. R., & Bashaw, W. L. The National Reference Scale for Reading: An application of the Rasch model. *Journal of Educational Measurement,* 1977, *14,* 161-179.

Rentz, R. R., & Rentz, C. C. Does the Rasch model really work? *Measurement in Education,* 1979, *10*(2), 1-8.

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. 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.

Wright, B. D. Sample-free test calibration and person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problems.* Princeton NJ: Educational Testing Service, 1968.

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

Wright, B. D. Misunderstanding the Rasch model. *Journal of Educational Measurement,* 1977, *14,* 219-225. (b)

Wright, B. D., & Douglas, G. A. Best procedures for sample-free item analysis. *Applied Psychological Measurement,* 1977, *1,* 281-295.

Acknowledgments

I am grateful to the Psychological Corporation for providing data for the study.

Author's Address

Send requests for reprints or further information to D. R. Divgi, 353 Lindquist Center, University of Iowa, Iowa City IA 52242.

[2] As a standard for comparison it may be noted that acceptable error in ability estimation due to *random* errors in difficulty estimates is .1 logit (Wright, 1977b, p. 224). A systematic error of the type seen in Figure 1 is more serious than a random error of the same magnitude because systematic errors will accumulate over successive equatings.