The Use of Presmoothing and Postsmoothing to Increase the Precision of Equipercentile Equating

Benjamin A. Fairbank, Jr.
Performance Metrics, Inc.

The effectiveness of smoothing in reducing sample-dependent errors in equipercentile equating of short ability or achievement tests is examined. Fourteen smoothers were examined, 7 applied to the distributions of scores before equating and 7 applied to the resulting equipercentile points. The data for the study included both results of simulations and results obtained in the operational administration of a large testing program. Negative hypergeometric presmoothing was more effective than the other presmoothers. Among the postsmoothers, both orthogonal regression and cubic splines were effective, especially the latter. The use of smoothing methods must be considered in light of their costs (increases in average signed deviations) and benefits (decreases in root mean square deviations). For many purposes, the benefits of smoothing with the negative hypergeometric may outweigh its costs.

Test equating is the process of finding which scores on two or more similar tests correspond to the same level of ability (or other trait) in a population of examinees. In principle, when two tests have been equated, either can be used with equal confidence to measure ability. The tests under investigation in this study were four-option, multiple-choice tests that are scored on the basis of the number of correct responses.

Test equating may be implemented in a wide variety of ways. Some of these methods are of recent origin and are technically sophisticated; others have been in use for several decades (see Holland & Rubin, 1982). This study addressed only equipercentile test equating as applied to two equivalent groups (Angoff, 1971). Lord (1980) demonstrated that two tests cannot be equated unless they are either perfectly reliable (an impossibility) or are strictly parallel, in which case they would not need to be equated. In practice, however, it is possible to equate highly similar tests, sometimes called "roughly parallel" tests, by the equipercentile method in such a way that the errors of equating are very small in comparison with other errors associated with testing (e.g., the errors of measurement arising as a consequence of the unreliability of tests, particularly the inherent lower reliability of short tests). In any case, although there may be some purposes to which it would be misleading to put equated scores, Lord (1980) pointed out that if scores are equated by the equipercentile method, then, when equated cutting scores are used, the different equated forms will result in the selection of the same proportion of examinees on all forms of the test, except for errors related either to sampling in the equating process or to the particular examinees tested operationally.

As with any procedure having the goal of estimating population characteristics based on data obtained from a sample, there are always sample-dependent errors present in test equating. If an equipercentile equating were to be done twice with similar samples from the same population, the re-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 11, No. 3, September 1987, pp. 245–262

0146-6216/87/030245-18$2.15

sults would differ. The extent of such differences has been estimated by Lord (1982) and their magnitudes appear as the standard errors of equipercentile equating. As expected, the size of the errors decreases linearly with the square root of the sample size. It is thus operationally impractical to reduce error beyond a certain amount by increasing sample size. As a consequence, practitioners of equipercentile test equating have sought other ways to reduce equating error. They have frequently used the methods of smoothing (Angoff, 1971; Cureton & Tukey, 1951; Divgi, 1983; Kolen, 1984; Lindsay & Prichard, 1971).

Smoothing

Two general classes of smoothing methods are defined here. Presmoothing is defined as the process of smoothing the observed score frequency distributions prior to equating. Postsmoothing is defined as the process of smoothing the equipercentile points after equating. The common intent of both smoothing methods is to remove small sample-dependent fluctuations from the nonsmoothed equatings so that the small-sample equatings will more nearly approximate the asymptotic equatings, or those which would result from the use of samples so large that the sample-dependent errors approach zero. The extent to which the various methods achieve this common intent is investigated by this research.

Presmoothing

Presmoothing methods are based on the concept that an observed data point in a sequence of points shows the combined effect of an underlying systematic relation among the points and sample-specific fluctuation or error of observation. If each point were replaced by a value jointly determined by the point replaced and the vicinal points, then the influence of the error of observation should be reduced, and the influence of the underlying regular function should be maintained. Seven presmoothing methods were used in this study:

1. Three-point moving medians
2. Five-point moving medians
3. Three-point moving weighted averages
4. Five-point moving weighted averages
5. Five-point moving weighted averages with root transformation
6. 4253H Twice
7. Negative hypergeometric

Six of these seven presmoothing methods are general-purpose methods which were developed for the smoothing of sequences of observations such as time series data (Tukey, 1977; Velleman, 1980; Velleman & Hoaglin, 1981). Detailed technical descriptions of the methods are available in the references cited; short descriptions are provided here. Moving medians and moving averages were used for presmoothing, as were a combined or compound presmoother and a presmoothing method based on a particular model of test scores, the negative hypergeometric distribution.

Of the seven methods for presmoothing the score distributions, three are described by Tukey (1977). In the first method, frequency distributions are smoothed by moving medians of span 3. Smoothing by moving medians of span 3 involves replacing each observed frequency with the median of three frequencies: that of the score of interest, that of the frequency associated with the next lower score, and that associated with the next higher score. The end values of the distribution, those corresponding to scores of 0 and perfect scores, cannot be smoothed effectively by moving medians. Moving medians of span 5 are found analogously, except that each frequency is replaced with a value which is the median of the frequency of interest, the two preceding frequencies, and the two following frequencies. The end points are not smoothed, but the next-to-end points are replaced by the smoothed values found by smoothing by medians of span 3.

Presmoothing by three-point moving weighted averages is analogous to three-point moving median smoothing, but instead of replacing each point in the raw frequency distribution with its median, it is replaced with a value that is calculated by taking the sum of twice the point being smoothed, the previous point, and the following point, then dividing the result by 4. This is equivalent to using weights of 1, 2, and 1. Again, the end values are

not smoothed. Five-point moving weighted averages are found by taking the raw frequencies five at a time and replacing each frequency with a weighted average of the frequency and the four surrounding values. The weighting function is one recommended by Angoff (1971); it weights the five points by the factors −3, 12, 17, 12, −3, and divides the resulting sum by 35. The recommended weights limit the effect of the smoothing process on the linear, quadratic, and cubic components of a curve. The end frequencies are not smoothed, but the next-to-end frequencies are smoothed by the three-point moving weighted average using weights of 1, 2, 1.

The five-point moving weighted average with root transformation is identical to the five-point moving weighted average, except that before the smoothing is applied, all of the frequency values are transformed by taking their square roots. The square roots are then smoothed. Following the smoothing, the inverse transformation, a squaring, is applied. The use of the square root transformation has the effect of decreasing the influence of larger values relative to the effect of the same smoother without the square root transformation. As a result, if a frequency is higher than surrounding frequencies, it is more effectively reduced with the root transformation. Conversely, if a frequency is lower than surrounding frequencies, it is more effectively raised to the surrounding values when the root transformation is not used. At the range of frequencies reported here, however, the differences are very slight.

The sixth smoother is a combination of smoothers proposed by Velleman (1980). Designated as 4253H Twice, it requires the successive application of four different smoothers, including moving medians of spans 4, 5, and 3, then finding the differences between the smoothed and unsmoothed distributions, the smoothing of that sequence of differences by the same compound method, and, finally, adding the smoothed differences back into the smoothed distribution. (For details, see Tukey, 1977; Velleman & Hoaglin, 1981.)

The final presmoothing method (Keats & Lord, 1962; Lord & Novick, 1968, pp. 515–520) is one devised explicitly for smoothing or fitting frequency distributions of test scores. The distribution is the negative hypergeometric, the appropriateness of which is derived from a binomial error model of test scores. The model assumes several technical conditions, one of which is equivalent to the assumption that all of the items on the test whose score distribution is being fit are equally difficult. That condition is known to be false in the case of virtually all operational tests, but the fit of the negative hypergeometric is still good enough to make it promising for further study (Keats & Lord, 1962).

Postsmoothing

Equipercentile equating starts with tables which show the frequency of each score in the samples tested for each of two tests, and ends in a table which associates with each score on one test a score on the other test. An integer score on one test is usually found to correspond to a non-integer score on the other test; the non-integer score may be estimated by linear interpolation. A plot of the score pairs shows a monotonically nondecreasing function whose form depends on characteristics of the sample and characteristics of the two tests being equated.

Postsmoothing is the process of passing a straight line or a curve among the points which define the equipercentile relationship. The equated scores are then determined by the resulting function. Postsmoothing methods have traditionally required the practitioner to judge where to pass a curve through a set of points (Angoff, 1971). In place of the use of a draftsman's French curve or analogous drawing aid, a number of analytic postsmoothing methods have been developed. Seven such methods were investigated here:

1. Linear regression
2. Quadratic regression
3. Cubic regression
4. Orthogonal regression
5. Logistic ogive
6. Cubic splines
7. Five-point moving weighted averages

The simplest equation which may be fit to the points resulting from an equipercentile equating is a straight line. This study investigated two different

straight lines: that defined by conventional least squares and that defined by orthogonal regression. The conventional least-squares procedure minimizes the sum of the squared vertical deviations from the line. In effect, the scores on the experimental test are considered to be known without error, and the line which best fits the equipercentile equivalents on the reference test is found. In orthogonal regression (Madansky, 1959), the quantity minimized is not the sum of the squared deviations parallel to the y-axis, but rather the sum of the squared deviations when those deviations are taken in a direction perpendicular to the regression line.

Orthogonal regression is appropriate when the variables represented on both axes are subject to measurement error, and neither can properly be considered the dependent or independent variable. This is frequently the case in test equating, for two reasons. First, such an equating can be used to convert scores from either test to the other. It is thus dissimilar to a least-squares regression equation in which the regression of y on x is rarely the same as that of x on y. Second, there are usually similar amounts of error associated with the reference and the experimental test.

The first two postsmoothing methods, then, are straight lines fit by conventional regression and by orthogonal regression. When conventional regression is used, the independent variable is the set of scores on the experimental test ranging from the lowest observed score to the highest observed score. The dependent variable is made up of the equipercentile points. Only under certain circumstances is it possible to fit resulting points well with a straight line. A straight line is appropriate if the two tests have the same skewness and kurtosis. The positioning and slope of the straight line will compensate for differences in means and standard deviations in the two tests. If there is a curvilinear component to the relationship defined by the equipercentile equating, then it must be fit by a curvilinear function. Quadratic and cubic functions have been used to fit such curves.

This investigation considered quadratic and cubic best-fitting (criterion of minimum least-squares deviations) smoothing curves. Quadratic curves can fit points whose best-fitting line is concave either upward or downward, whereas cubic equations can fit curves with an inflection point, so that part of the curve is concave upward and part of it is concave downward. The third and fourth postsmoothing methods, then, were quadratic and cubic regression functions, fit by the method of least squares and modified by the requirement of monotonicity.

In some equatings it is observed that the equipercentile equating function is relatively flat at both of its ends and steeper in the middle. Such a shape can be fit by a cubic curve, but it can also be fit by a logistic ogive, a curve defined by

$$Y = A + \frac{B - A}{1 + \exp[-C(X - D)]} \quad (1)$$

where A, B, C, and D are fitted constants. The points resulting from equipercentile equating were fit by a logistic ogive, the fifth postsmoothing method.

All of the smoothing methods mentioned above have the disadvantage that they impose a function of a given form on the data, even if it is not appropriate. Such a Procrustean requirement is contrary to the rationale of smoothing, especially when the shape of the function is not appropriate to the points to which it is to be fit. The sixth and seventh postsmoothing methods do not define the shape of the function in advance of the fitting. The sixth function fit to the points was not a continuous function, but rather a smoothing of the discrete resulting points. The smoothing function replaces each point with a point which is the weighted average of the point being replaced and the four surrounding points. The method is that of five-point moving weighted averages, as described earlier. The equating requires interpolation between the resultant points.

The final postsmoothing function was used by Kolen (1984), who obtained good results by fitting cubic smoothing spline functions to the points resulting from the equipercentile equating. A smoothing spline differs from an interpolating spline in that the latter is constrained to pass through exactly known points, while the former is conceived of as passing among approximately known points. As used by Kolen, a cubic smoothing spline

for N points (in the present case, an equipercentile equating of two N-item tests) is a set of $N-1$ cubic functions, each of which takes as its domain the interval from the Ith point to the $(I+1)$th point on the x-axis. The range and specific form of the function are determined by the data in the interval. The cubic functions come together with the same function value and slope (or derivative) at each of the interior $N-2$ points, which are called ducks or knots in the language of spline fitting. The resulting curve can be of almost any differentiable shape.

Objectives

The aim of the present effort was to evaluate the effects of different methods of presmoothing and postsmoothing on the accuracy of test equating. The study was exploratory in nature, designed to determine which methods hold the most promise for operational use.

Method

This investigation used three different approaches to determine the effectiveness of each of the 14 smoothing methods. The first approach used simulated tests and examinees; the second and third used data from tests administered to examinees under operational conditions. The advantage of simulated tests and examinees is that all quantitative aspects of the tests and examinees are completely specified, and it is possible to know in advance the results of theoretically errorless equatings or those equatings which are unaffected by sample-dependent errors. Operational data have the advantage that they are obtained under conditions typical of the ones under which smoothing methods would be used. The data contain all of the departures from theory that may be found in operational test settings.

The first of the three methods of evaluation involved comparing each of the smoothed equatings with a known errorless equating. The known errorless equating was based on a method that yielded results typical of an equating using an infinitely large sample. The method requires deriving a distribution of expected score frequencies, the distribution being that which would result from administering the test to a sample so large that the observed proportions at each score were observed essentially without error. The results of the simulated test administrations were compared to that criterion equating.

The second method was a similar comparison of sample and criterion equatings, but in place of data based on simulations and on an errorless equating, the comparison used operationally obtained data and an equating based on an unusually large sample size. The third method used the statistical jackknife (Mosteller & Tukey, 1977) to estimate the size of standard errors of smoothed and unsmoothed equatings using operationally obtained data and simulated data. As a methodological cross-check, the errors for unsmoothed equating were also compared to standard errors computed by means of Lord's (1982) formula.

Simulations

The objective of the simulations was to provide data that modeled those which might result from administration of tests similar to the subtests of the Armed Services Vocational Aptitude Battery (ASVAB; United States Military Entrance Processing Command, 1984). The range of test lengths investigated covered the range of subtest lengths in the operational ASVAB. Three test lengths were used: 15 items, 30 items, and 50 items. For each test length, two very similar tests were created in simulation. The tests were not strictly parallel. They were, however, as similar to each other as are ASVAB subtests within a single subject area in ASVAB 8, 9, and 10 (Ree, Mullins, Mathews, & Massey, 1982).

A sample of 2,000 randomly selected simulated examinees (simulees) was administered one test, while a second sample of 2,000 was administered the other test. This process was repeated for a total of 100 simulated administrations for each test length. The same two simulated tests were used, but the sample of simulees was drawn anew for each simulated administration. Different simulated samples were used for each of the test lengths. Detailed descriptions of the methods of simulation and the

characteristics of the resulting tests may be found in Fairbank (1985).

The simulations were implemented within the framework of item response theory (IRT). Each aspect of the simulated tests and of the simulees was specified in IRT terms in such a way as to model operational subtests in the ASVAB testing program. (See United States Military Entrance Processing Command, 1984, for a description of the ASVAB program.) Simulated items were generated at random so that the items' distributions of a, b, and c parameters approximately matched those reported by Ree et al. (1982) for the operational subtests.

The method used to simulate the administration of a test is similar to a method developed by Ree (1980) for use in a simulation implemented in another context. Such a simulation results in response vectors which include correct responses due to the joint influences of ability and guessing, just as operational data show both such influences. When all 2,000 simulees had "responded" to all items in a test, the test was scored and analyzed to determine the mean and standard deviation of scores, the item difficulties, the item biserial correlation coefficients, and other statistics. The resulting test statistics and distributions were compared with the results of the subtests which the simulated tests were designed to match. After several iterated adjustments of the simulation parameters, the technical aspects of the resulting simulated tests resembled the technical aspects of the ASVAB tests very closely.

Technical and statistical details of the tests, including their test characteristic curves and test information curves, are presented in Fairbank (1985). Each of the six simulated examinations was "taken" by 100 groups of 2,000 simulees. Either of two methods was used to administer a test in simulation. The first method is that described by Ree (1980). The second method involved taking a sample of 2,000 observations at random from the expected observed score distribution (EOSD), found using a method of Lord and Wingersky (1983), for a test. Score distributions were tabulated for each simulated administration. For each test length, 100 equipercentile equatings and smoothings were then performed using the methods described below. The smoothings and equatings were the same for the operational and simulated data, and are described following the description of the operational data.

Criterion Equatings

The preparation of simulated tests allows total control of the simulated test situation. It is, therefore, possible to know in advance the criterion or "true" equating of the tests used. IRT makes possible several approaches to the determination of the criterion equating. It is possible, for example, to determine the true scores associated with various abilities (or θ values) and equate true scores through common θ. Analogously, a variant of true-score equating can be performed, and for each integer true score on the experimental test, the corresponding θ can be computed (usually by means of inverse interpolation); then the score on the reference test which corresponds to that value of θ can be found. This method has the advantage of giving equated scores for each number-correct true score, and interpolation of tabled values is not required.

True scores are never known in actuality, however; the above method thus is not entirely appropriate. The method used to establish the criterion equatings for the simulations used in the present study is based on the EOSD for each test. The algorithm developed by Lord and Wingersky (1983) was used to prepare distributions of expected observed scores for each of the six simulated tests. In an EOSD, each score has associated with it a proportion of examinees, not a frequency. The distributions model the result of administering the test to an infinitely large number of examinees and observing the relative frequency of each score.

The EOSD method of establishing a criterion equating is appropriate because the aim of the present research was to determine methods of smoothing which compensate for the relatively small sample sizes that must be used operationally. By comparing the small-sample equatings ($N = 2{,}000$) with those that result from an "infinite" sample (i.e., those based on the EOSD), the extent of improvement resulting from smoothing is directly observable. The criterion equatings, then, were the unsmoothed equipercentile equatings which result

from using the EOSDs in the unsmoothed equipercentile method.

Operational Data

The operational data were taken from a set of ASVAB scores with very large sample sizes (approximately 100,000 examinees) for three roughly parallel forms of each of several subtests. Among those subtests were two forms of Mathematics Knowledge (25 items) and two forms of Electronics Information (20 items). In addition to the frequency distributions of test scores for all examinees, there were available 100 samples of 2,000 scores for each of the four subtests (two forms each of Mathematical Knowledge and Electronics Information). The samples were drawn at random without replacement from the larger samples of 100,000 examinees. Two test lengths were thus available in the operational data: 20 and 25 items.

The test lengths used were constrained in part by the availability of data and in part by the aim of increasing the generalizability of the study by employing a number of different test lengths for operational and simulated tests. For the operational data, criterion equatings were established by using the full sample of 100,000 examinees. Although that sample equating is not totally error-free, it is based on a sample 50 times as large as the samples of size 2,000 and thus was expected to have sample-dependent errors only approximately one-seventh as large as those found in the small equatings. As with the simulated data, the criterion equatings were unsmoothed equipercentile equatings, as described below. As with the simulated data, 100 reduced-sample equatings were made for each of the test pairs, both without smoothing and with each of the 14 smoothing methods.

Equatings

All test equatings were performed using the equipercentile method described by Lindsay and Prichard (1971). For the unsmoothed equatings and the equatings to which only postsmoothing was to be applied, the raw frequency files were equated. When the equatings involved presmoothing, the smoothed frequency estimates were equated. Following the equatings and smoothings (which are described below), each test or simulated test had associated with it a criterion equating, an unsmoothed equating, and 14 smoothed equatings, one for each of the smoothing methods used.

Smoothing Methods

Most of the smoothing methods require no description beyond that given above. Two of the postsmoothing methods, however, are more complex and require further description. The fifth postsmoothing method was the fitting of a logistic ogive to the data. The ogive was fit by the method of the simplex, which is an iterative, rather than an optimal, method. The method requires an initial estimate of the four parameters (upper and lower asymptotes, slope, and location) which define the ogive; it then successively finds better and better sets of parameters.

The procedure used here for fitting the cubic spline departed in three ways from that used by Kolen (1984). First, Kolen fit two spline functions, one using the equated experimental test scores as the dependent scores, and the other using the reference test scores as the dependent variables. The final equated values were obtained by averaging the equatings resulting from the use of those two spline functions. In order to retain comparability with other smoothing methods used in this research, the experimental test was used as the dependent variable in fitting the spline.

The second departure involved difficulties which were encountered with cubic spline smoothing at lower ends of the score distribution. Kolen (1984), finding similar difficulties at both ends, addressed it by applying the splines only in the interval of test scores ranging from the .5th to the 99.5th percentile. The shortest of his tests, however, was 40 items, and few examinees scored at either of the extremes. Smoothing by means of cubic splines as described by Reinsch (1967) requires an estimate of the standard errors of the y variables at each duck, but at the lower end point, where frequencies are at or near zero, the standard errors are not defined or do not exist. For the purposes of this

investigation, the end standard errors were assigned the value of the closest defined standard error, where "closest" means the numerically closest integer score.

Initially, smoothing methods relied heavily on human judgment and experience in passing a line among the points. The hope of users of the more analytic smoothing methods has been that an optimum or nearly optimum method might be found so that judgmental methods would not be necessary. Smoothing could then be automated and thus replicable and objective. The work of Kolen (1984), whose cubic splines have been among the most effective postsmoothing methods described in the literature, has not avoided the necessity of intervening judgment in the application of the smoothing process. For the current study, however, when over 500 applications of the smoothing technique were required, automated smoothing was a necessity. Thus, the third departure was the use of standard errors in the cubic spline fitting procedure. Kolen's (1984) procedure for achieving "moderate" smoothing was used by allowing the smoothing parameter to take the value of $K/2$, where K equals the test length plus 1.

Analysis of Equating Results

Each of the five tests, three simulated and two operational, had associated with it one criterion equating, 100 unsmoothed equatings based on sample sizes of 2,000 (called the "small sample"), and 100 sets of 14 smoothed equatings based on the same samples. The question of interest was the effect of the smoothings on the accuracy of the equatings.

The measures used to define the accuracy of the equatings are based on the concept of deviations. A deviation is a difference between an equated score obtained with a small sample and an equated score based on a criterion equating. At each observed score on the experimental test, the corresponding score on the reference test was found using the criterion equating. The equated scores were found as decimal fractions not rounded to the nearest integer. The score corresponding to the same experimental test score was then found for the unsmoothed small-sample equating and for each of the 14 smoothed equatings. The differences between the equated score based on the criterion equating and the equated score based on the small-sample equatings were found for each possible score on the experimental test, for the unsmoothed and for the smoothed equatings, for all 100 replications.

These differences, or deviations, were the raw data used for evaluating the smoothings. A deviation, D, associated with a given score on an experimental test, unsmoothed or smoothed by a particular method, is thus defined by the formula $D = x - x'$, where x is the equated score based on criterion equating and x' is the equated score based on small-sample equating. Each test thus has as many deviation scores, D, as there are items on a test, plus 1 (for a score of 0). For each of the 100 small-sample equatings, the deviations at each score were combined across equatings to give a general measure of deviation at each score.

Three such deviation measures were computed. The first measure is the root mean square deviation (RMSD), found by taking the square root of the sum of the squares of the deviations across all 100 samples. The second measure is the average absolute deviation (AAD), which is simply the mean of the absolute value of the deviations computed across all samples. The third measure is the average of the signed values of the deviations (ASD), found by taking the mean of the deviations across all 100 replications. ASD differs from AAD in that the absolute values are not found before the mean is computed. Positive values of ASD indicate that the small-sample equating resulted in a value which was generally lower than the criterion equating values, whereas negative values indicate the opposite. These three measures, RMSD, AAD, and ASD, were found for each score point on each test for the unsmoothed equatings and for each of the 14 smoothed equatings, across all 100 sample equatings. The three measures of deviation taken together allow an evaluation of the effects of the smoothing methods.

AAD and RMSD both give numbers which represent the unsigned magnitude of an average deviation. AAD is the arithmetic mean of absolute values, while RMSD has the effect of weighting (or

emphasizing) the deviations which are far from the criterion equating. ASD averages the deviations as does AAD, but it includes their sign. The resulting ASD shows how far the mean of the equated values for all 100 samples is above or below the value given by the criterion equating. This is a significant value for two reasons. First, equipercentile test equating has not been shown to be statistically unbiased; ASD estimates how large the ASD actually is. Second, methods which reduce RMSD or AAD may increase ASD. Thus, RMSD, AAD, and ASD must be considered together in evaluating a smoothing technique.

Standard Errors

Two cross-checks were made to ensure the accuracy of the methods used to determine RMSD. First, the standard errors of equipercentile equating were determined using a formula derived by Lord (1982). The resulting standard errors, one at each score level which was at or above chance, or expected guessing score level, were compared to (1) RMSDs obtained from the simulated test administrations, (2) those obtained from the operational data, and (3) those obtained from the results of the jackknifing. The observed RMSD values should be empirical estimates of the same standard errors which the Lord (1982) standard errors represent. In each case, the data from the criterion equating were used to develop the standard errors.

As described above, the criterion equatings for the simulations were established on the basis of expected observed scores, which correspond to "infinite" sample sizes, while the criterion equatings for the operational data were based on 100,000 cases. In order to make the standard errors of the criterion equatings comparable to the RMSD figures calculated from the samples of size 2,000, a figure of 2,000 was used to represent the sample size in calculating the standard errors for the criterion equatings, although the proportions called for by the formula were those obtained from the full criterion equating samples.

Lord's (1982) method allows calculation of standard errors for the unsmoothed case of equipercentile equating. The simulations discussed above allow empirical estimations of the standard error for all of the smoothing conditions as well as for the unsmoothed condition. The agreement or nonagreement of the Lord formula values with the values generated through simulation indicate the extent to which the simulations and evaluations, in the unsmoothed case, are behaving as intended. There is no corresponding analytical cross-check on the values of the standard errors in the smoothed cases because formulas for standard errors in those cases do not exist.

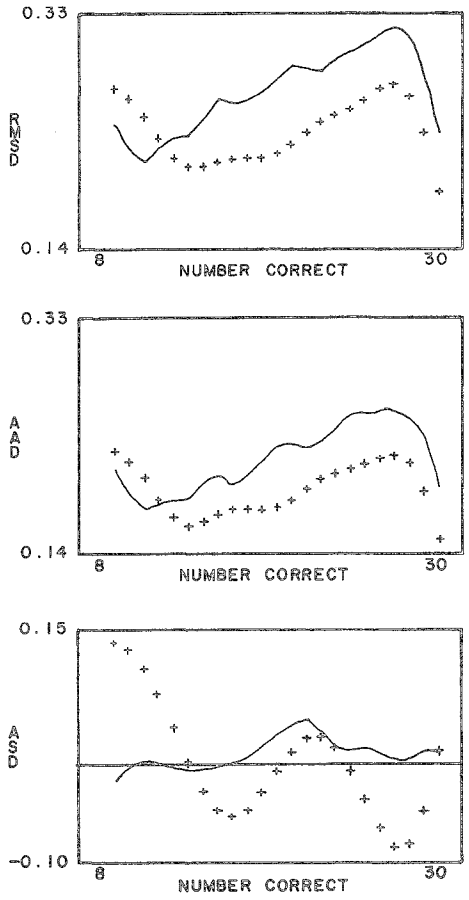
In order to provide corroborating or noncorroborating estimates of the standard errors for smoothed equatings, the equatings were conducted with the use of Tukey's statistical jackknife. The jackknife (Mosteller & Tukey, 1977) provides an estimate of the standard error of a procedure regardless of whether analytical formulas for such errors are available. Estimated standard errors were computed for each of the nonchance score levels on the tests; these standard errors were averaged over all such test scores. Thus, each test combined with each smoothing method resulted in a mean standard error of equating as estimated by the jackknife and as estimated by the RMSD of the small-sample equatings.

Results

Fairbank (1985) provided extensive graphical presentation of the results of each of the five tests smoothed by each of the smoothing methods. Summaries of the data are included here, but the figures are too extensive to reproduce in full. Two of these figures are presented here, illustrating (1) the effects of a particularly effective case of presmoothing with the negative hypergeometric, and (2) a modestly effective postsmoothing with cubic splines.

Figures 1 and 2 each represent one test length and one method of smoothing. Each figure is divided into three panels. Each panel shows measures of deviation as a function of the raw score on the experimental test, both with and without smoothing. In each figure, the top panel shows the effect of smoothing on RMSD, the middle panel shows its effect on AAD, and the bottom panel shows the effect on ASD. Two functions are shown on each

Figure 1
Effect of Presmoothing With the Negative Hypergeometric on Measures of Deviation for a Simulated 30-Item Test
(Solid Lines Represent Unsmoothed Equatings; + Signs Represent Smoothed Equatings)



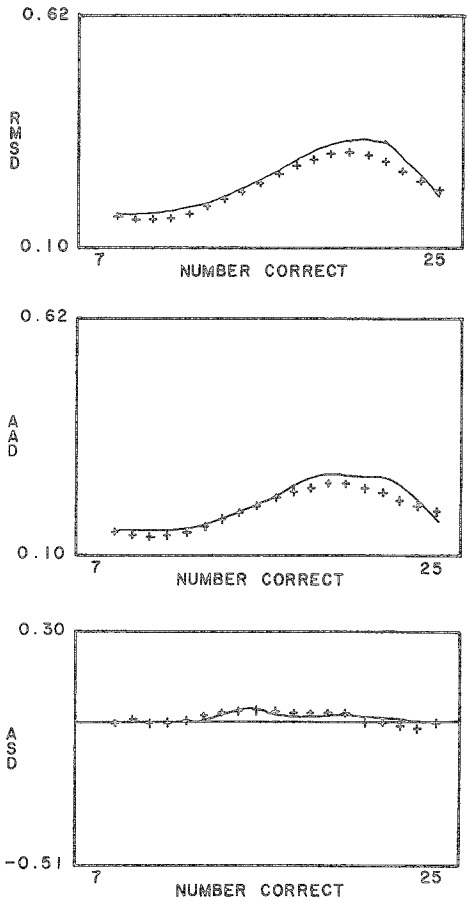
panel of each figure. The continuous line shows the RMSD, AAD, or ASD which results from equating samples of size 2,000 without smoothing, while the + signs indicate the RMSD, AAD, or ASD when the same samples are equated with smoothing. Each point on the graph is an average of the deviations over the 100 samples; deviations are differences between the criterion equating and the small-sample equating. When there is a horizontal line plotted in a graph, that line represents zero deviation. For the figure panels which depict AAD and RMSD, + signs which lie below the continuous line indicate that an improvement, or a reduction of deviations, resulted from smoothing. The situation with ASD is slightly more complex, because ASD may be either positive or negative. Improvement, or reduction of ASD, is indicated when the + signs lie either between the continuous line and the x-axis of the graph, or closer to the x-axis than the continuous line. In order to show the more relevant deviations effectively, the figures do not present information on the levels of RMSD, AAD, or ASD at test scores below the guessing level for each test.

Figure 1 shows the effects of presmoothing with the negative hypergeometric, as applied to a simulated 30-item test. The top two panels of the figure show that both RMSD and AAD are reduced by the use of the smoother, while the bottom panel shows that ASD is sharply increased at many score points for the test. Thus there is a trade-off between reduced RMSD and increased ASD. Figure 2, in contrast, shows the results of postsmoothing the 25-item operational test with cubic smoothing splines. It is seen that there is slight but consistent improvement in RMSD and AAD with the use of the spline smoother, but very little effect on ASD.

Tables 1 and 2 present a summary of the deviations associated with the unsmoothed equatings and with four of the most effective smoothed equatings. A briefer summary is presented in Table 4 for all 14 equatings. Table 1 shows the RMSD, AAD, and ASD as averaged across all test scores above the guessing level with no smoothing. The averages of the ASD were taken over the absolute values of ASD so that positive and negative values would not cancel out.

Table 2 shows results for the smoothed equatings. The averages of RMSD, AAD, and ASD are presented as proportions of the deviations in Table 1. Thus, values less than 1 indicate that smoothing reduced the deviation, while values greater than 1 indicate an increase in deviations. For example, a value of .9 indicates that a particular smoothing method reduced the mean measure of a deviation to 90% of its unsmoothed value, when that mean was taken over all scores on a test which were above chance level. The table indicates the effects of the smoothers in a global sense. The effects are

Figure 2
Effect of Postsmoothing With Cubic Smoothing Splines on Measures of Deviation for an Operational 25-Item Test
(Solid Lines Represent Unsmoothed Equatings; + Signs Represent Smoothed Equatings)



averaged over all scores above chance and thus may obliterate the locally high deviations. The standard errors of equating and the related measures of RMSD, AAD, and ASD are summary measures of the extent to which a test equating is subject to sample-dependent error.

Jackknifing

Table 3 presents the standard errors of unsmoothed equating as estimated by Lord's (1982) analytic formula with those estimated by means of repeated reduced-sample equating (i.e., RMSD from simulated or operational tests) and those estimated by means of the jackknife. The standard errors are presented in the metric of test items. They are averaged over all test scores which are higher than chance level. The standard errors thus indicate that standard errors of equating vary with the length of the tests, and vary with the method used to estimate them. More important, however, is that the sizes of the errors differ only slightly with the different methods of estimating them.

Table 4 presents the standard errors of smoothed equatings, as estimated by the RMSD of the reduced-sample equatings (100 samples of 2,000) and as estimated by Tukey's jackknife (Mosteller & Tukey, 1977). Again, in order to facilitate evaluating the effects of smoothing, the RMSD and jackknifed estimates of the error associated with the smoothed equatings are presented as proportions of the RMSD and error associated with the unsmoothed equatings. Thus, values in Table 4 greater than 1.0 indicate that smoothing increased the RMSD, whereas values less than 1.0 indicate a reduction in RMSD.

Discussion

To evaluate the effects of smoothing, particularly its effects on deviations, it is helpful to consider such deviations within the context of the accuracy of ability or achievement tests more generally. The standard errors of equating discussed here are not the only measurement errors which arise in the testing process. There are also standard errors of measurement that are intrinsic to any test which is not perfectly reliable. Equation 2 relates reliability (r_{xx}), standard error of measurement (SEM), and test score standard deviation (SD):

$$\text{SEM} = \text{SD}(1 - r_{xx})^{1/2} \quad . \qquad (2)$$

Thus, the SEM for the experimental 15-item test, based on a reliability (KR-20) estimate of .80 and a standard deviation of 3.28, is 1.47. Similarly, the SEM for the experimental 30-item test is 2.20, and that for the experimental 50-item test is 2.74 (values of the SD are typical of those found for tests used in the present study). The corresponding av-

Table 1
RMSD, AAD, and ASD Averaged Over All Samples At All Scores Above Chance Level for Unsmoothed Equatings, for Simulated and Operational Tests, As a Function of Test Length

Test Length	RMSD	AAD	ASD
Simulated Tests			
15	.134	.106	.009
30	.269	.214	.016
50	.439	.348	.028
Operational Tests			
20	.184	.145	.029
25	.242	.192	.015

erage standard errors of equating as estimated by Lord's formula, given in Table 4, are .15, .30, and .51. Thus the standard error of equating ranges from approximately only 10% to 20% of the standard error of measurement. A smoothing method which reduces RMSD by 20% will thus reduce total test error by 2% to 4%.

Presmoothers

Table 4 shows that smoothing by the method of three-point moving medians had no overall beneficial effect. Frequently it resulted in less accurate equatings than unsmoothed equatings. Similar results were obtained from the use of five-point moving medians. There is no consistent beneficial effect, and frequent deleterious effects, on all three measures of deviation. Whatever local gains are achieved are offset by losses elsewhere.

The method of three-point moving weighted averages, the results of which are given in detail in Table 2, showed generally encouraging results, although the gains are modest. The gains are particularly evident on the 15-item simulated test and the 20-item operational test. There was a modest increase in the ASD at the high score levels in both tests. The method of five-point moving weighted averages has generally negligible effects on all three measures of deviation. The result of applying the method of five-point moving weighted averages with root transformation was virtually identical to the result of applying the method of five-point moving weighted averages without root transformation, as described above. There was no significant benefit achieved. Smoothing by the method of 4253H Twice was generally ineffective and resulted in local increases and local decreases in the measures of deviation.

The results of smoothing by means of the negative hypergeometric, as shown in detail in Table 2, show consistent improvement in RMSD and AAD as a consequence of smoothing. The effects are particularly impressive with the simulated tests, presumably in part because the criterion equatings for those tests are nearly perfect, not estimated from very large samples. The gains are not uniform across the tests. On the shorter tests at lower scores, the measures of RMSD and AAD actually increased as a consequence of using the negative hypergeometric. The beneficial effects of the negative hypergeometric do not extend to the measures of ASD. The ASD increases both globally and locally, sometimes quite dramatically. These increases were expected at the lower end of the test, where guessing is a factor, but increases at the upper end were not expected. It must be noted, however, that the ASD figures were initially low (see Table 1), so that a tripling of ASD may still denote an acceptably low level.

The question of the amount of ASD that can be considered acceptable is complex. Until there are equating methods which can be shown to be consistent, sufficient, efficient, and unbiased, it will

be necessary to balance such properties against each other to determine the mix which is optimal for a given purpose. The largest increase in ASD occurred for the 50-item test. The increase, by a factor of approximately 7.5, resulted in an increase in the mean ASD (Table 1) from .015 score points to .11 score points. The mean RMSD for the same test was .24 without smoothing, and .23 with smoothing. Thus, for the 50-item test used in this study the increase in ASD was greater than the reduction in RMSD, although the resulting ASD was only half the magnitude of the RMSD.

Table 2
Proportion of Mean Deviation for RMSD, AAD, and ASD (Based on Unsmoothed Data in Table 1) for Simulated and Operational Tests, Using the Method of 3-Point Moving Weighted Averages, the Method of Negative Hypergeometric, the Method of Cubic Splines, and the Method of 5-Point Moving Weighted Averages

Smoothing Method, Test Type, and Test Length	RMSD	AAD	ASD
Method of 3-Point Moving Weighted Averages			
Simulated Tests			
15	.962	.963	1.183
30	.974	.975	.846
50	.979	.981	1.303
Operational Tests			
20	.953	.948	1.100
25	.969	.970	1.606
Mean	.967	.968	1.208
Method of Negative Hypergeometric			
Simulated Tests			
15	.891	.903	2.919
30	.865	.867	3.596
50	.852	.861	3.453
Operational Tests			
20	.905	.908	2.008
25	.966	.989	7.479
Mean	.896	.906	3.891
Method of Cubic Splines			
Simulated Tests			
15	.914	.917	1.548
30	.935	.927	2.086
50	.984	.984	1.773
Operational Tests			
20	.935	.932	.956
25	.928	.927	1.364
Mean	.939	.937	1.545
Method of 5-Point Moving Weighted Averages			
Simulated Tests			
15	.984	.985	1.115
30	.990	.989	.980
50	.994	.993	1.013
Operational Tests			
20	.985	.985	.995
25	.990	.990	1.069
Mean	.989	.989	1.035

Table 3
Standard Errors (Averaged Over All Scores Above Chance Level) of Unsmoothed Equating Estimated by Three Methods As a Function of Test Length for Simulated and Operational Tests

	Simulated			Operational	
Method of Estimation	15 Items	30 Items	50 Items	20 Items	25 Items
Lord's Formula	.15	.30	.51	.18	.25
Average of 100 Samples	.13	.27	.44	.18	.24
Jackknifing	.15	.25	.47	.17	.23

An increase in ASD may be more acceptable when two tests are equated so that they may be used interchangeably than the same increase would be when the objective of the equating is to replace one operational test with another. If two tests are used interchangeably, then a systematic tendency to deviations in one direction on one test will be offset by scores on the other test. Thus, if the forms of the test are administered at random to examinees, there will be no expected advantage to any examinee. If, in contrast, a test is equated to another so that the older test may be replaced, then ASD will result in equated scores which give results that differ systematically from the scores expected on the test which was replaced.

Why does the negative hypergeometric smoothing method outperform the other presmoothers? One likely reason is that it takes into account all of the information in a distribution's mean and standard deviation in arriving at the smoothed frequency for each point. The other presmoothers respond only to local conditions and so may incorporate, rather than eliminate, some sample-dependent local fluctuations. Furthermore, among the seven presmoothers investigated, only the negative hypergeometric is based on a mathematical model of testing. The other smoothers work by applying general algorithms which have been shown to be useful in a wide variety of circumstances. It appears that those smoothers do not bring the sample score distributions closer to the shape of the distribution of the parent population, whereas the negative hypergeometric does. However, the negative hypergeometric does so at the cost of increased ASD at some specific test scores.

Postsmoothers

The use of both linear and quadratic regression postsmoothing resulted in modest reductions of RMSD and AAD at the middle score ranges, but increases at the upper ranges. The increases in the deviations of the upper score ranges were especially prominent in the 50-item simulated test. Improvements in RMSD and AAD were partially offset by increases in ASD. Deviation measures tended to be high at the upper end, especially with the 50-item test. Use of cubic polynomial regression smoothing had less benefit than did quadratic regression in most cases, but it also caused less increase in RMSD at high scores, and less of an effect on ASD.

Because a cubic function can follow a given curve more accurately than can a quadratic function, it would be expected that the cubic regression smoothing would lead to more accurate equating than linear or quadratic regression smoothing. Findings to the contrary suggest that the cubic functions may have been following and fitting sample-dependent fluctuations in the individual equatings. Smoothing by means of orthogonal regression had effects which were very similar to those which resulted from the use of linear regression. The deleterious effects at the high end of the test, however, were less pronounced. Postsmoothing by means of the logistic ogive resulted in modest reductions in RMSD and AAD, at the usual cost of increases in ASD, and with the previously noted problems at the highest scores.

Smoothing by cubic smoothing splines (Table 2) provides the most promising results among the postsmoothing methods. There were modest re-

Table 4
Proportional Change in Standard Errors (Magnitude of Standard Error Estimates for Smoothed Equatings When Expressed As Proportions of the Corresponding Unsmoothed Equatings) for 14 Smoothing Methods, As a Function of Test Length, for Simulated and Operational Tests

Smoothing Method and Method of Estimation	Simulated 15 Items	Simulated 30 Items	Simulated 50 Items	Operational 20 Items	Operational 25 Items
Presmoothing					
3-Point Moving Median					
RMSD 100 Samples	1.00	1.00	1.06	1.00	.99
Jackknifing	1.07	1.00	1.02	1.05	1.02
5-Point Moving Median					
RMSD 100 Samples	1.02	.99	1.03	1.02	1.01
Jackknifing	1.16	1.01	1.00	1.00	1.02
3-Point Moving Weighted Averages					
RMSD 100 Samples	.96	.97	.98	.95	.97
Jackknifing	.98	.98	.98	.94	.98
5-Point Moving Weighted Averages					
RMSD 100 Samples	.99	.99	.99	.99	.99
Jackknifing	1.01	1.00	.99	.98	1.00
5-Point Moving Weighted Averages With Root Transformation					
RMSD 100 Samples	1.00	.99	1.00	.99	.98
Jackknifing	1.01	1.00	1.00	.98	1.01
4253H Twice					
RMSD 100 Samples	1.02	1.01	1.03	.98	.99
Jackknifing	1.09	.99	.97	.98	.98
Negative Hypergeometric					
RMSD 100 Samples	.89	.87	.85	.91	.97
Jackknifing	.89	.80	.88	.81	.87
Postsmoothing					
Linear Regression					
RMSD 100 Samples	.97	1.13	1.13	.92	1.24
Jackknifing	.76	.82	.93	1.08	.88
Quadratic Regression					
RMSD 100 Samples	.99	1.03	1.75	.97	.96
Jackknifing	.97	.92	.99	.99	.86
Cubic Regression					
RMSD 100 Samples	1.08	1.05	1.32	1.12	.99
Jackknifing	.99	1.03	1.45	1.19	.93
Orthogonal Regression					
RMSD 100 Samples	.87	1.01	.96	.88	1.18
Jackknifing	.70	.85	2.15	1.09	.89
Logistic Ogive					
RMSD 100 Samples	.87	.97	.94	.88	1.17
Jackknifing	.70	.85	2.03	1.09	.88
Cubic Splines					
RMSD 100 Samples	.91	.94	.98	.94	.93
Jackknifing	1.00	1.01	1.01	.99	.93
5-Point Moving Weighted Averages					
RMSD 100 Samples	.98	.99	.99	.99	.99
Jackknifing	.95	.99	.99	.98	.99

ductions throughout in the amounts of RMSD and AAD; no end-point problems occurred at either end, and the problem of increases in ASD was not particularly severe. Table 4 shows that the gains, though modest, are consistent for the cubic smoothing spline method. Finally, as Table 2 shows, the effect of postsmoothing by five-point moving weighted averages was very minor, but consistently beneficial at all scores and with all three measures.

Jackknifed Estimates

The close agreement of the standard errors as estimated by the three methods, shown in Table 3, supports the contention that each of the three estimation methods is both appropriate and correctly executed. Although there are slight differences in the estimates, they are not large enough to call into question the appropriateness of the methods. As Table 4 shows, the mean RMSD and jackknifing methods do give somewhat divergent results in some cases; therefore, a conservative criterion for the recommendation of adopting a smoothing method is that the method should appear advantageous with both estimation techniques. The method best meeting that criterion at all test lengths is the method of presmoothing by the negative hypergeometric.

Conclusions

One presmoother and one postsmoother stand out as deserving further study and consideration for future operational use. The presmoother is the negative hypergeometric; the postsmoother is the cubic smoothing spline. When its effect is estimated by jackknifing, the cubic smoothing spline was not effective in reducing RMSD with the 20-item operational test, nor with any of the simulated tests. There was, however, consistent improvement resulting from the use of the smoothing splines as measured by RMSD. This divergence of measures of effectiveness suggests the need for further study before unequivocal recommendations can be made.

Presmoothers other than the negative hypergeometric were either ineffective, inconsistent in their effects, or have associated with them disadvantages such as greatly increased ASD. Divgi (1983) likewise found merit in the use of the negative hypergeometric, although he also found that the three- and four-parameter beta binomial distributions were more effective than the negative hypergeometric. The lack of effectiveness of the other presmoothers may say less about the presmoothers than it does about the robustness of equipercentile equating. The various cumulative frequency counts used in equating may be degraded by all of the smoothers except the negative hypergeometric.

The cubic smoothing spline has a number of intuitively appealing characteristics: It can follow a curve of any shape, it can pass as close to the fit points as appropriate, and it is theoretically neutral in the sense that its use does not depend on the applicability or appropriateness of any statistical theory of testing. Its effectiveness, which is also reported by Kolen (1984), is thus not surprising. Although the improvements due to the splines were modest, the fact that there is no concomitant increase in ASD makes their use particularly attractive. The cubic smoothing splines perform, in effect, exactly what hand smoothing attempts: It passes a theoretically neutral curve among the points.

Limitations

The present study is limited in several respects, all of which may tend to reduce its generalizability to other applications. First, only five tests were used: two operational and three simulated. Generalizations to other tests may be inadvisable if the tests do not statistically resemble those used for this study. Second, the tests used, especially the simulated tests, may be more similar to each other than are most operationally equated tests. Generalization to less similar tests is of questionable appropriateness. Third, all equated pairs were pairs of tests of the same length, a condition not always found operationally.

Another issue of potential importance could not be investigated using the current methodology. One of the particularly significant advantages of equipercentile test equating is that when tests equated by the equipercentile method are used interchangeably to select only persons who score at or above a certain percentile, then there is no expected ad-

vantage to any examinee in taking any particular form of the test in place of any other form. It is not clear that presmoothed equipercentile equatings retain that property. In applications where such percentile invariance is an essential consideration, the use of presmoothing should await further research.

Finally, it may be asked whether any benefit would be realized from smoothing both before and after equating, thus combining the two methods. Four combinations of presmoothers and postsmoothers were investigated, but not reported here because of limitations of space. Such combinations did not give better results than the more effective method in each pair when used separately.

Recommendations

Among the presmoothing methods, the negative hypergeometric deserves consideration for operational use. If any of the presmoothers studied here is to be adopted, then the negative hypergeometric would be the most appropriate. It has the effect of reducing RMSD by about 10%, a benefit which could also be achieved by increasing sample size by about 20%. Among the postsmoothers, gains were not as evident with linear, quadratic, and cubic regression smoothing, as had been anticipated. In those cases where an a priori decision has been made that the smoothing shall be linear, the use of orthogonal regression should be favored over the use of standard regression. Where the shape of the regression fitting is not determined in advance, then the use of cubic splines appears appropriate. These two postsmoothing methods, orthogonal regression and cubic splines, are appropriate for operational use with tests similar to those studied here, and may be useful with other tests if further research confirms their usefulness.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 508–600). Washington DC: American Council on Education. (Reprinted by Educational Testing Service, Princeton NJ, 1984.)

Cureton, E. E., & Tukey, J. W. (1951). Smoothing frequency distributions, equating tests, and preparing norms. (Abstract of presented paper.) *American Psychologist, 6,* 404.

Divgi, D. R. (1983). *Comparison of some methods for smoothing score distributions*. Paper presented at the meeting of the American Educational Research Association, New Orleans.

Fairbank, B. A., Jr. (1985). *Equipercentile equating: The effects of presmoothing and postsmoothing on the magnitude of sample-dependent errors* (AFHRL-TR-84-64). Brooks Air Force Base TX: U.S. Air Force Human Resources Laboratory.

Holland, P. W., & Rubin, D. B. (1982). *Test equating*. New York: Academic Press.

Keats, J. A., & Lord, F. M. (1962). A theoretical distribution for mental test scores. *Psychometrika, 27,* 59–72.

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

Lindsay, C. A., & Prichard, M. A. (1971). An analytical procedure for the equipercentile method of equating tests. *Journal of Educational Measurement, 8,* 203–207.

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

Lord, F. M. (1982). The standard error of equipercentile test equating. *Journal of Educational Statistics, 7,* 165–174.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Wesley.

Lord, F. M., & Wingersky, M. S. (1983). *Comparison of IRT observed score and true score equatings* (RR-83-26-ONR). Princeton NJ: Educational Testing Service.

Madansky, A. (1959). The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association, 54,* 173–205.

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression*. Reading MA: Addison-Wesley.

Ree, M. J. (1980). AVRAM: Adaptive vector and response automation method [Computer program abstract]. *Applied Psychological Measurement, 4,* 277–278.

Ree, M. J., Mullins, C. J., Mathews, J. J., & Massey, R. H. (1982). *Armed Services Vocational Aptitude Battery: Item and factor analysis of forms 8, 9, and 10* (AFHRL-TR-81-55, AO-A113 465). Brooks Air Force Base TX: U.S. Air Force Human Resources Laboratory.

Reinsch, C. H. (1967). Smoothing by spline functions. *Numerische Mathematik, 10,* 177–183.

Tukey, J. W. (1977). *Exploratory data analysis*. Read-

ing MA: Addison-Wesley.

United States Military Entrance Processing Command (1984). *Test manual for the Armed Services Vocational Aptitude Battery* (Document No. DoD 1304.12AA). North Chicago IL: Author.

Velleman, P. F. (1980). Definition and comparison of robust nonlinear data smoothing algorithms. *Journal of the American Statistical Association, 75*, 609–615.

Velleman, P. F., & Hoaglin, D. C. (1981). *Applications, basics, and computing of exploratory data analysis*. Belmont CA: Duxbury Press division of Wadsworth.

Acknowledgments

This research was conducted under contract with the U.S. Air Force Human Resources Laboratory, Project 7719, Force Acquisition and Distribution System, Task 771918, Personnel Qualification Tests. The opinions expressed in this paper are those of the author and not necessarily those of the United States Air Force. The author thanks Malcolm Ree and Capt. Toni Wegner of the Manpower and Personnel Division of the Air Force Human Resources Laboratory. Their discussions, suggestions, and support were essential to this project. In addition, suggestions and insights provided by Michael Levine and Mark Reckase in the early stages of the research were most valuable and are appreciated.

Author's Address

Send requests for reprints or further information to B. A. Fairbank, Jr., Performance Metrics, Inc., 5825 Callaghan Road, Suite 225, San Antonio TX 78228, U.S.A.