DOCUMENT RESUME

ED 271 478 TM 860 265

AUTHOR Cook, Linda L.; Petersen, Nancy S.
TITLE Problems Related to the Use of Conventional and Item Response Theory Equating Methods in Less than Optimal Circumstances.
PUB DATE Apr 86
NOTE 39p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Descriptive (141)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Educational Research; *Equated Scores; Error of Measurement; Evaluation Methods; *Latent Trait Theory; Measurement Techniques; *Psychometrics; Research Needs; *Research Problems; Sample Size; *Sampling; Scaling; Statistical Distributions; *Statistical Studies; Testing Programs; Test Items
IDENTIFIERS *Anchor Tests; Smoothing Methods

ABSTRACT

This paper examines how various equating methods are affected by: (1) sampling error; (2) sample characteristics; and (3) characteristics of anchor test items. It reviews empirical studies that investigated the invariance of equating transformations, and it discusses empirical and simulation studies that focus on how the properties of anchor tests affect conventional and item response theory equating results. Rather than offering a cookbook procedure for obtaining accurate equating results, the paper provides some practical suggestions for practitioners to follow when equating and describes some needed research. More research is needed toward improving the use of analytic techniques for smoothing or modeling marginal and bivariate frequency distributions. More testing programs need to evaluate the extent to which their operational equating results are population invariant and to examine the similarity in equating results for major population subgroups taking the test forms at different administration dates. An investigation is needed of how lack of content representativeness of the linking items, or the differences in ability levels of the new and old form groups affect: (1) item parameter scaling and their interaction with the number of linking items, (2) the position of the linking items, and (3) the scaling procedure used. (LMO)

# Problems Related to the Use of Conventional and Item Response Theory Equating Methods in Less than Optimal Circumstances

Linda L. Cook and Nancy S. Petersen[1,2]

Educational Testing Service

Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, April, 1986



[1]The authors' names are in alphabetical order.

[2]The technical assistance of Daniel Eignor and Marilyn Wingersky, in the preparation of some sections of this paper, is gratefully acknowledged.

# Problems Related to the Use of Conventional and Item Response Theory Equating Methods in Less than Optimal Circumstances

Linda L. Cook and Nancy S. Petersen[1,2]

Educational Testing Service

## INTRODUCTION

Many psychometricians view score equating as a subjective art with theoretical foundations since, in practice, we never know the true relationship between scores on different forms of the same test. Furthermore, in practice, the data never satisfy the assumptions of the various equating models. Two forms of the same test are never perfectly parallel, equally reliable, or unidimensional. Seldom do we have the ideal data collection design in which all test-takers take both test forms. And, sample sizes are seldom as large as desired.

Consequently, when equating scores on two forms of a test we need to determine how similar the test forms are in content, difficulty, and reliability. And, if an anchor test is used, we need to determine the extent to which the anchor test mirrors the properties of the total tests. It is also important that we gather as much information as possible about the extent to which the samples to be used in equating are similar in composition and ability and representative of the population for whom the test is intended. Further, we need to evaluate the extent to which the data fit the

---

[1]The authors' names are in alphabetical order.

[2]The technical assistance of Daniel Eignor and Marilyn Wingersky, in the preparation of some sections of this paper, is gratefully acknowledged.

various equating models and how robust the models are to sampling error and lack of model fit in order to make a sound subjective judgment as to which model is most likely to produce an equating that will be sufficiently accurate for the intended score use.

The purpose of this paper is to discuss some practical problems encountered using conventional and/or IRT methods for score equating. In particular, we will discuss how equating results may be affected by (1) the use of analytic techniques for smoothing empirical distributions, (2) sample characteristics, and (3) properties of the items included in an anchor test. It is likely that these problems are too complex to lend themselves to straightforward statistical solutions. However, more research in these areas could lead to practical guidelines for dealing with these issues.

## EQUIPERCENTILE EQUATING AND SMOOTHING

In equipercentile equating, a transformation is chosen such that raw scores on two tests are considered to be equated if they correspond to the same percentile rank in some group of examinees (Angoff, 1984). Equipercentile equating is based on the definition that the score scales for two tests are comparable with respect to a certain population if the score distributions for the two tests are identical for that population (Braun and Holland, 1982; Lord, 1950). That is, scores on form X and form Y are equated on population P if the distribution of transformed Y scores is the same as the distribution of untransformed X scores. Equipercentile equating of scores on the two forms of a test can be thought of as a two-stage process (Kolen, 1984). First, tabulate the relative cumulative frequency distributions for the two forms to be equated. Second, obtain equated scores (i.e., scores with the same percentile

ranks) on the two forms from these relative cumulative frequency distributions.

In practice, equipercentile equating is not as straightforward as the above description sounds. For example, there will seldom be a raw score on form X with the same exact percentile rank as a given raw score on form Y. Thus, to find pairs of scores with the same percentile rank, it is necessary to make the distributions continuous. Consequently, some subjective elements, such as choice of interpolation method, must be introduced into the equating process. Most practitioners use linear interpolation although curvilinear procedures could be used. However, while choice of interpolation procedure will have some affect on the conversion line, a greater problem with equipercentile equating is probably its susceptibility to sampling error.

Equipercentile equating, unlike other equating methods, makes no assumptions about the tests to be equated. All the procedure does is compress and stretch the score units on one test so that its raw score distribution coincides with that on the other test. Equipercentile equating is simply an empirical procedure and, as such, is very data dependent. This is a problem particularly when sample sizes are small.

When sample sizes are small, it is not uncommon to obtain score distributions with zero frequencies for scores within the observed score range and only a few observations, if any, for extreme scores. Conversions based on such data are usually somewhat irregular and step-like. These irregularities are unlikely to be closely reproduced if the equating were redone using different samples of the same size. And, if the equating were redone using very large samples, it is likely that most of the irregularities would disappear and that the conversion would be a rather nice monotonic-increasing function. To mitigate the effects of sampling error on equipercentile

equating, it may be that practitioners should routinely use analytic techniques for smoothing or modeling the frequency distributions and/or the resulting conversion line.

A wide variety of analytic techniques for smoothing empirical distributions have been described in the statistical literature. The rolling weighted average of frequencies method described by Angoff (1984) and attributed to Cureton and Tukey (1951), the nonparametric probability density estimation procedures, such as kernal estimators, described by Tapia and Thompson (1978), and the techniques, such as moving medians, for smoothing empirical distributions described by Tukey (1977) are methods applicable to a wide range of distributional forms requiring minimal statistical assumptions. Regression-based polynomial functions, smoothing cubic spline functions (Reinsch, 1967), and theoretical distributions can be fitted to empirical distributions. Keats and Lord (1962) have suggested use of the negative hypergeometric distribution and Brandenburg and Forsyth (1974) have suggested use of the Pearson Type I (general beta) function. A procedure, based on item response theory, for estimating the population observed-score distribution has been described by Lord (1980).

A number of the analytic smoothing procedures described above have been used by practitioners. Few studies, however, have been conducted that were explicitly designed to evaluate the extent to which equipercentile equating can be made more robust to sampling error by use of analytical smoothing techniques. A notable exception is a recent study by Fairbank.

Fairbank (1985) investigated a variety of analytical techniques for smoothing empirical distributions (presmoothers) and conversion lines (postsmoothers) to determine whether statistical smoothing could increase the accuracy of equipercentile equating. Presmoothing techniques used were moving

medians, rolling weighted averages, and the negative hypergeometric. Postsmoothing techniques used were the logistic ogive, cubic splines, rolling weighted averages, and linear, quadratic, cubic, and orthogonal regression. For the tests used in the study, the most effective technique was the negative hypergeometric and the most effective postsmoother was cubic smoothing splines. It was also found that combining presmoothers and postsmoothers did not result in an improvement beyond that obtained with the more effective of the combined pair used alone.

The results of the Fairbank study suggest that it may be more beneficial in equipercentile equating to smooth the empirical distributions than to smooth the resulting conversion line. Possibly, these results are due in part to the fact that the procedures used to smooth the conversion lines were regression based. Most definitions of equated scores require that the conversion be symmetric (Angoff, 1984; Lord, 1980). For example, if a score, of 45 on test X is equated to a score of 50 on test Y, then a score of 50 on test Y must equate to a score of 45 on test X. Since the regression of X on y is generally not the same as the regression of Y on x, the use of a regression based technique for smoothing the conversion line will destroy the symmetry of that conversion.

The smoothing problem becomes more complex when an anchor test data collection design is used. In an anchor test design, one form of the test is administered to one group of examinees, a second form to a second group of examinees, and a common test to both groups. The groups may be random groups from the same population or they may be non-equivalent or naturally-occurring groups that, consequently, vary in systematic ways. In either case, scores on the anchor test can be used to estimate performance of the combined group of examinees on both the new and old forms of the test, thus simulating by

statistical methods the situation in which the same group of examinees take both forms of the test. Ideally, the anchor test would be composed of questions like those in the two forms to be equated. And, the higher the correlation between scores on the anchor test and scores on the new or old form, the more useful the data.

It may be inappropriate to apply analytic techniques for smoothing empirical distributions to data collected via an anchor test design. Independent application of such techniques to the four marginal distributions prior to equating may destroy the bivariate relationship between each test form and the anchor test, upon which the success of the equating depends. For data collected via an anchor test design, analytic techniques for smoothing the bivariate distributions prior to equating should be more appropriate; however, little work has been done in this area as it relates to equating. At Educational Testing Service, we are in the process of investigating the use of analytic methods for smoothing two-way contingency tables. The methods that we are considering are based on generalized log-linear models that choose the smoothest distribution on the two-way table such that certain key features of the observed data are preserved. For example, one might preserve the correlation, the quadrant totals, and the marginal means, variances, and skewness; or, one might preserve only the correlation and the marginal means and variances. Models that preserve fewer features of the data do more smoothing. And, residuals and chi-square tests can be used to help evaluate the quality of the fit between various models.

More research needs to be done on the use of analytic smoothing techniques in the equating process. It is possible that smoothing may significantly increase the robustness of equating results, particularly when sample sizes are small. Currently, we know too little about the effects of smoothing on

equating results. Too few studies such as that by Fairbank have been undertaken. And, it is possible that those results may not be generalizable to other testing situations. Different-shape distributions may call for different techniques.

## POPULATION INVARIANCE AND SAMPLE SELECTION

There is no universally accepted definition of equated scores. Instead, a variety of definitions have been discussed in the literature (Angoff, 1984; Holland and Rubin, 1982; Lord, 1977, 1980). However, underlying most of these definitions is the requirement that the equating transformation be population independent. That is, the equating transformation should be the same regardless of the group from which it is derived.

Whether or not an equating function based on one population also works for another population, is an empirical question that can be tested with data (Braun and Holland, 1982). For example, one can investigate whether or not the equating transformation between scores on two forms of a mathematics test derived using random samples from the population taking the test forms is the same as that derived using samples of all males. The question of population invariance, however, can be compounded by the manner in which the data are collected for the equating experiment. The most commonly used methods of data collection are the anchor test design (described in the preceding section) and the random groups design in which the two test forms are given to random samples from the same population.

Many testing programs offer multiple administrations of their tests, and, examinees who choose a particular administration at which to take the test may vary in nonrandom, systematic ways. For example, those taking the test at one

administration may be more able than those taking it at another administration. Or, those taking the test at one administration may have more relevant coursework than those taking it at another administration. Using a random-groups design, one might find that the equating function was essentially invariant for subgroups taking the test on one administration date, but if the two forms were administered and equated using data from another administration, the two equating functions would differ. In an anchor test design, if the two groups used for the equating take the test on different administration dates, the possiblility exists that they may not be subgroups from the same population. If this situation exists, the equating function obtained from such a design may be very problematic. In this section we will review a number of empirical studies designed to evaluate the population invariance properties of equating transformations.

Kingston, Leary, and Wightman (1985) examined the feasibility of using IRT true-score equating (employing the three-parameter logistic model) to equate scores on the Verbal and Quantitative measures of the Graduate Management Admission Test (GMAT). Population invariance of the equating transformation was of particular interest to these researchers. To investigate this question, Kingston, et al. collected data on two forms of the GMAT that were offered at the same administration. Six samples of examinees were selected to provide data for the experimental equatings. Two samples were random, one sample consisted of all males, another of all females, one of "younger" students (ages 21-23), and one of "older" students (29 years of age or older). The results of the study indicated that the equatings were very consistent across the random and subgroup populations for both the Verbal and Quantitative measures. The researchers concluded that differences among converted scores for equatings performed on the six subgroups were negligible.

Angoff and Cowell (1985) examined the population independence of equating transformations derived using conventional linear and equipercentile procedures applied to forms of the homogeneous Graduate Records Examination (GRE) quantitative test and a specially constituted heterogeneous GRE verbal-plus-quantitative test. Both tests were equated using random samples from the entire population as well as random samples from subgroups defined by sex, race, field of study, and level of performance. The researchers evaluated the invariance of the equating transformations by examining departures of the transformations based on the subgroups from the population transformation (based on the random samples selected from the entire population). Discrepancies between the population and subgroup transformations were evaluated in terms of empirically determined standard errors of equating.

Some discrepancies between subpopulation transformations and the population transformation were noted for both the homogeneous and heterogeneous test. The subpopulation producing the largest discrepancies was a very able Physical Science group. Because the majority of the discrepancies were not significant for the homogeneous test, the researchers concluded ". . . that, at least for this homogeneously constructed test--and presumably for other homogeneous tests--the assumption of population independency is unchallenged" (p. 71). The Physical Science subgroup presented a more serious problem for the heterogeneous test. Angoff and Cowell concluded that discrepancies between the Physical Science subgroup equating transformation and the transformation obtained from the equating based on the total population could possibly be attributed to the fact that the two forms gave unequal weight to questions from the physical sciences and, thus, were not strictly parallel. This lack of parallelism appeared to affect the equating results obtained for the Physical Science group.

The Kingston, Leary, and Wightman (1985) and Angoff and Cowell (1985) studies employed equating samples that were similar in level and dispersion of ability, i.e., males and females (or other subgroups used for equating) may have had different ability levels, but the two samples of males that were used for the form-to-form equatings were similar in ability level. In contrast to the studies reviewed above, the studies we will review next employ new and old-form groups who took the test on different administration dates and consequently may differ somewhat in ability level and other characteristics that may affect their test scores.

Cook, Eignor, and Taft (1985) examined the results of equating two forms of a secondary school biology achievement test, which had been constructed to be reasonably parallel to each other. Their study employed one old-form sample and two different new-form samples. The old-form sample was randomly selected from a fall administration of the test. One new form sample was randomly selected from a spring administration of the test and the second sample was randomly selected from a fall administration. It should be noted that students taking the biology test in the spring are typically able students who have recently completed a course in biology. Students taking the test in the fall are less able students, the majority of whom have not formally studied biology for six to eighteen months. It seems logical to suspect that some degree of forgetting (if, for example, immediate recall is more important for some content classifications than for others) could affect what the test is actually measuring for the spring and fall populations and hence affect total test scores as well as performance on the anchor test items.

Table 1 contains summary statistics which describe the performance of the three samples on the two forms of the total test and the common anchor test

(58 common items included as part of the total score on both the new and old forms of the test). It can be seen, from the data presented in Table 1, that the two new and old form fall samples are very similar in their performance on the 58 item anchor test. On the other hand, the new-form spring sample performs very differently on the anchor test items.

---

Insert Table 1 about here

---

Differences between performance on the common items of the spring and fall samples is best illustrated by an examination of the plots shown in Figure 1. These plots show the relationship between equated delta values (transformed item difficulty values, Henrysson, 1971) for the fall old form and spring and fall new form groups. It is obvious, from examination of the plots shown in Figure 1, that the item difficulty indices (deltas) demonstrate more scatter for the spring-new-form/fall-old-form combination than they do for the fall-new-form/fall-old-form pairing. Further evidence of the disagreement between these indices for the spring/fall combination and their close agreement for the fall/fall combination, is shown by the correlation coefficients given below the plots. These data strongly indicate that the 58 common items contained in the new and old biology test forms measure the same underlying constructs for the two fall samples, but are differentially difficult for the spring group. Thus, it is quite likely that the two biology test forms, even though constructed to be very parallel, measure different skills or constructs depending upon whether they are administered to a spring or fall group.

---

Insert Figure 1 about here

---

The two biology test forms were equated to each other using linear and equipercentile observed score equating methods (see Angoff, 1984, for a description of these methods). IRT true-score equating based on the three-parameter logistic model (Lord, 1980, p. 193) was also performed. All equatings were carried out first using the spring-new-form/fall-old-form combination and second using the fall-new-form/fall-old-form pairing. The results of these equatings are presented in Table 2. Purusal of the data shown in Table 2 indicates that all of the equatings using the spring-new form/fall-old-form combination resulted in scaled-score means at least 15 points higher than those based on the fall-new-form/fall-old-form combinations.

---

Insert Table 2 about here

---

Several questions may be asked about the results presented in Table 2. For one, are the equatings discrepant due to the differences in ability level of the new and old-form samples or are they discrepant because the test is measuring different (non-parallel) constructs for the spring and fall groups?

Another, closely related question is: Will an equacing transformation determined by using samples from one of these groups remain invariant for the other group? This latter question was investigated by Cook (1984). In her study, two different forms of the biology test were equated using new and old-form samples from fall administrations and then the equating was repeated using new and old form samples from spring administrations. Although the

spring and fall groups differ in level and dispersion of ability, the two spring samples used for the equatings were similar to each other as were the two fall samples. The situation is similar to that investigated by Angoff and Cowell (1985) and Kingston, et al. (1985). The results of these equatings are compared in Table 3. One can see, from examination of these data, that the equating transformation determined using the spring/spring combination results in reported scores, that are 10 points higher, throughout most of the score range, than those obtained by a transformation determined using fall/fall samples. These results suggest that the spring and fall groups taking the biology test may not be subgroups from the same population and that the biology test may not be measuring the same thing for these two populations.

---

Insert Table 3 about here

---

A final question that comes to mind, when reviewing the data presented in Table 2 is: Why are the IRT true-score results as affected by differences in group ability as are the results based on the observed score equating methods? Lord (1984) makes the point that, if the IRT model holds, true scores will be equated for all subpopulations of examinees. The fact that the results based on the spring-new-form/fall-old-form and fall-new-form/fall-old form groups differ so much seems to indicate that the assumptions of the IRT model are not met or, as previously hypothesized, the spring and fall groups are not subgroups of the same population.

To summarize, Kingston, Leary, and Wightman (1985) and Angoff and Cowell (1985) found little differences among equating transformations obtained from subgroups of a specific population when the tests to be equated were constructed to be parallel and homogeneous in content and the equating samples

were similar in ability. Angoff and Cowell found that when heterogeneous tests were equated under the same conditions, the results were not stable across all subgroups.

The results of the studies conducted by Cook, Eignor, and Taft (1985) and Cook (1984) may be contrasted to those obtained by Kingston et al. and Angoff and Cowell. Cook et al. found that when relatively parallel forms of an achievement test were equated using groups of students who took the tests on different administration dates, both conventional and IRT equating results were seriously affected. They concluded that the disparate equating results were obtained because students taking the test at the different administrations differed in relative recency of their course work. This difference in recency of training interacted with test content. Thus the test measured different constructs depending upon the sample of examinees to whom the test was administered.

## PROPERTIES OF ANCHOR TEST ITEMS

In the previous section of this paper, empirical studies that investigated the invariance of equating transformations were reviewed. The studies focused on how equating results may be affected by examinee characteristics. Some of the studies involved the use of an anchor test data collection design. When using an anchor test design, one must also be concerned with the properties and characteristics of the anchor test items in relation to the total test.

As mentioned earlier in this paper, an anchor test is used to reduce equating error resulting from differences in ability between new and old form groups. The anchor test may consist of common items that are scattered throughout the new and old form (internal anchor test); or the anchor test

may appear in a separately-timed section of the test (external anchor test).

In the context of IRT equating, anchor tests are usually referred to as linking tests or linking items. The linking items are used to "scale" item parameter estimates. If, prior to equating, new and old test forms are given to groups that differ in level of ability, the IRT parameter estimates for the two forms will be on different scales. It is well known that IRT equating requires that the item parameter estimates for two test forms be on the same scale prior to equating, and that the quality of the equating depends upon how well the item scaling is accomplished.

In this section of the paper, we will review empirical and simulation studies that focus on how the properties of anchor tests affect conventional and IRT equating results. With regard to conventional equating methods, the properties of the anchor test that will be discussed are length, parallelism with the tests to be equated, and consistency of item difficulty for new and old form groups. With regard to IRT equating, we will discuss anchor test properties such as length, difficulty, and size of the standard errors of estimation of the items; and we will also discuss studies that compare methods for placing item parameter estimates on the same scale (concurrent calibration versus an item transformation method).

Klein and Kolen (1985) investigated the relationship between anchor test length and accuracy of results obtained using conventional equating methods. The test of interest was a certification test which contained 250 multiple-choice items. These researchers, using data from a fall administration of the test, separated examinees into similar and dissimilar-ability-level groups.

Within each group, they equated the test to itself several times using the Tucker observed score method and anchor tests of 20, 40, 60, 80 and 100 items. The results of their study indicated that, when groups are similar in ability, anchor test length has little effect on the quality of equating. However, when the groups used for equating differ in level of ability, length of the anchor test becomes very important. The authors concluded that, "When the tests being equated were very similar, or in this particular case, identical, and the groups of examinees very similar, substantially more-accurate equating was not obtained by lengthening the anchor test. However, longer anchor tests did result in more-accurate equating when the groups of examinees were dissimilar" (p. 10). They emphasize that the results of this study are based on anchor tests that correspond very closely to the total test with respect to content representation, difficulty, and discrimination.

Results of the Cook, Eignor and Taft (1985) study, discussed earlier in this paper are also pertinent to a discussion of anchor test length. Figure 2 contains plots of the 58 common items used in the equatings previously discussed, 36 common items chosen by content experts to represent concepts in biology most likely to remain stable across the spring and fall groups, 29 common items for which delta values (item difficulty indices) changed the least for the spring and fall groups and, 29 common items for which delta values changed the most for the two groups. The plots shown in Figure 2 compare delta values obtained for the different item sets given to the very able spring group and the less able fall group. Figure 3 repeats the plots shown in Figure 2; however, the comparison is between delta values obtained for the two fall groups that are similar in level of ability. It is fairly clear, from the data presented in Figures 2 and 3, that, for all subsets of items (different anchor tests) the items are differentially

difficult for the spring and fall groups and similar in level of difficulty for the two fall groups. The question is: What affect does this have on the equating results?

---

Insert Figures 2 and 3 about here

---

Table 4 presents results for the equatings, based on the various sets of anchor test items, for the spring/fall and fall/fall sampling combinations. Notice, when the groups differ in level of ability (spring/fall samples), the different anchor tests yield very disparate equating results. However, when the groups are similar in level of ability (fall/fall samples) the various anchor tests yield equating results that are in close agreement. These findings, in conjunction with those obtained by Klein and Kolen, strongly indicate that when groups differ in level of ability (as they typically do for anchor test designs), special care must be taken when selecting the set of common items constituting the anchor test.

---

Insert Table 4 about here

---

Klein and Jarjoura (1986) investigated, for conventional equating methods, the importance of the content representation of the anchor test. For their study, they equated a 250 item multiple choice test to itself through three intervening links or anchor tests. The success of the equating was judged by how closely the identity relationship of equating a test to itself was recovered. For the representative chain of equatings, they used three 60-item anchor tests, all representative of the content of the total tests. For the nonrepresentative chain, the first anchor consisted of 101 items, the second

of 105 items and the third of 60 items. Only the 60-item anchor was representative of the total test content. Both Tucker observed-score and Levine true-score equating methods were used. Based on the results of their study, the authors concluded that it was quite important to use content-representative anchors. They explained the importance in the following way. "Consider an extreme example in which two test data groups differed on only some of the content areas. If a nonrepresentative anchor consisted of items from only the content area for which there were no differences, it would fail to reflect the true differences between the groups on the full test form" (p. 203).

The results of the three studies reviewed indicate that the properties of an anchor test can seriously affect conventional equating results. The number of items included in an anchor test, as well as the content representativeness of the items, appear to be important variables. However, these variables seem to decrease in importance as the equating samples become more similar in level and dispersion of ability. Since anchor test designs are usually used in situations where groups differ in ability level, the results of these studies have serious implications for this type of design.

Item response theory equating applications use a variety of procedures for placing parameter estimates from separate item calibrations on the same scale. Cook and Eignor (1981) and Stocking and Lord (1983) have provided detailed descriptions of many of the commonly-used transformation methods. For score equating applications, users of the three-parameter logistic model and the computer program LOGIST (Wingersky, 1983) make use of a set of linking (common) items on each test to place parameter estimates for items appearing on two or more test forms on the same scale. This "scaling" can be essentially accomplished in two ways. One way, referred to as concurrent

calibration (Petersen, Cook, and Stocking, 1983), involves the estimation of item parameters for the test forms and the linking items in a single calibration run. The second procedure involves estimation of item parameters for one or more test forms along with a set of linking items in one calibration run and the estimation of item parameters for one or more different forms and the linking items in another calibration run. The item parameter estimates for the test editions are then placed on the same scale via an item scaling procedure. A current procedure used by a number of researchers is referred to as the "characteristic curve method" (Stocking and Lord, 1983). Once item parameter estimates for the various forms have been placed on the same scale, it is possible to equate scores on the test forms using IRT true-score equating procedures (Lord, 1980).

Most of the research that has been conducted, to date, has essentially addressed the question of how many common items are necessary to place item parameter estimates on the same scale prior to IRT true-score equating. Vale, Maurelli, Gialluca, Weiss, and Ree (1981) investigated the problem using simulated data with 5, 15, and 25 common items and three different shapes of the linking item section test information curve: peaked, normal, and rectangular. Vale et al. assumed that good estimates for the linking items were already known, and they required that there be enough linking and unique items to get good ability estimates. For each examinee, the researchers obtained two estimates of ability, one from the linking items and the other from the unique items. The estimates were used to determine the transformation required to put the unique items onto the common scale. For this method of placing parameter estimates on a common scale, Vale et al. found that 15 to 25 items were necessary. They also found that the linking item sections with a rectangular or normal information function gave better

results than those with a peaked information function. McKinley and Reckase (1981) studied the number of common items problem in the context of the construction of large item pools. They worked with real data from a multidimensional achievement test covering a number of different areas of achievement. McKinley and Reckase concluded that 5 items were not adequate, 25 items were better than 15, but 15 items were adequate for linking with a concurrent calibration design.

Raju, Edwards, and Osberg (1983) studied linking test size in the context of vertical equating; they also used real data. Unlike the previously described studies, Raju et al. made use of the Rasch model along with the three-parameter logistic model. Parameter estimates were derived in separate calibration runs; linking constants were determined by setting equal the standard deviates of item difficulty estimates obtained for the common items in the separate calibrations. Raju et al. found, for both models, that short linking tests, with as few as six or eight items, performed almost as well as longer linking tests containing twice or three times as many items. They also found the three parameter model to provide more acceptable equating results.

Wingersky and Lord (1984) studied the number of linking items problem in the context of concurrent calibration. In the most extreme case studied, Wingersky and Lord found that two good linking items (items with small standard errors) worked almost as well as a set of 25 common items.

Wingersky, Cook, and Eignor (1986) investigated the affects on IRT true-score equating results of the characteristics of the linking items. The study was carried out using the three-parameter logistic item response theory model and Monte Carlo procedures. So that the simulated data reflected actual test data, the true item parameters were taken from the estimated parameters obtained from LOGIST calibrations of item responses obtained from selected

administrations of the verbal sections of the College Board Scholastic Aptitude Test (SAT-V). The characteristics of the items were investigated for two of the common linking or scaling designs: concurrent calibration and the characteristic curve transformation method. The authors investigated the affects on these two scaling designs of using linking tests consisting of 10, 20, and 40 items. In addition, the affects on equating of two different characteristics of the parameter estimates of the common or linking items were investigated: (1) items with parameter estimates having standard errors of estimation (SEE) similar to those found in typical SAT-V common item sections, and (2) items with parameter estimates which have small standard errors of estimation. Finally, the affect on true-score equating results of using peaked and uniform distributions of abilities to estimate item parameters was investigated.

The results of the Wingersky, Cook, and Eignor study showed very little difference in equating results based on placing item parameter estimates on the same scale using a concurrent calibration procedure or a characteristic curve method of scaling. As expected, the authors found, for both scaling methods, that the accuracy of the equating results improved as the number of linking items was increased. The characteristic curve transformation method seemed to require slightly more items than the concurrent calibration procedure. A surprising finding was that, for both scaling procedures, linking items chosen to have SEEs similar to those typically found for SAT-V equating items, provided slightly better equating results than those deliberately chosen to have small SEEs. They also found that the equating results were slightly better when a uniform distribution of abilities rather than a peaked distribution of abilities was used to estimate parameters for the linking items.

Kingston and Dorans (1984) examined what they referred to as "context effects" on IRT true-score equating. They defined context effects as occurring when examinees respond differently to an item depending upon its location within a test. The researchers investigated the susceptibility to item location effects of 10 item types from the Graduate Record Examination (GRE) General Test. To study location effects, they administered two versions of Form B of the GRE to random samples from the same population. One version contained items in the typical operational location; the second version contained the same items in nonoperational locations. In general, Kingston and Dorans found some practice and fatigue effects for most of the item types they studied. To evaluate the affect of item location on IRT true-score equating, they equated the two versions of Form B to Form A (a form of the GRE General Test that had been previously placed on scale) and compared the results. They found that the two equatings of the Verbal measure of the test agreed fairly closely. In contrast, the equating of the Quantitative measure that resulted from the use of items in the nonoperational position showed a small but consistent bias which the researchers attributed mostly to practice effects on the data interpretation items. The equating results that were the most profoundly affected were those for the Analytical measure. The analytical item types showed extreme sensitivity to item location which was reflected in a difference in means of almost 30 converted score points between the equatings of Form B with items in the operational and nonoperational positions.

Kingston and Dorans concluded that the results of their study demonstrated susceptibility to location effects depends upon the item type. Review of the analytical item types showed the items to be quite complex with extensive and complicated directions. They hypothesized that once the directions were

understood, the items were fairly easy to handle. Hence, the difficulty of an individual item depends upon how many items of the same type precede it.

The results of the Kingston and Dorans study have important implications for equating applications such as IRT based pre-equating. IRT pre-equating is a variant of IRT true-score equating that uses a data collection design that involves the calibration of items (typically using pretest data) and subsequent equating of test forms prior to a test's administration. Because items will usually not appear in the same position in a final form of the test as they do when they are precalibrated, the appropriateness of an equating transformation derived from pre-equating requires that the parameter estimates of the items not be influenced by location effects.

Eignor (1985) examined the possibility of pre-equating the Verbal and Mathematical sections of the Scholastic Aptitude Test (SAT) using pretest items administered in a variable section which is given along with the operational sections of the SAT. For his study, Eignor calibrated the Verbal and Mathematical sections of two SAT forms using pretest data and pre-equated these forms to existing SAT forms. The results of the pre-equating were compared to those obtained when the test forms were operationally equated using IRT true-score equating. Eignor concluded that the results of the pre-equating of the Verbal section for one form were quite successful, while those obtained for the second form exhibited discrepancies between the pre-equating results and the operational results of upwards of 20 scaled-score points.

The pre-equating of the mathematical sections of both forms studied provided results at least as discrepant as those for the most problematic verbal section and, thus, were a source of concern. Eignor hypothesized that discrepancies between difficulty parameter estimates obtained when items were

calibrated in multiple pretest sections and then as part of an intact final-form (resulting in differences between pre-equating and operational equating transformations) could possibly be attributed to the type of location effects described by Kingston and Dorans (1984). However, Eignor's design did not allow him to test this hypothesis. Eignor also suspected that discrepancies in the pretest and intact final form item difficulty parameter estimates, and the resultant IRT equatings, may have been the result of differences in the ability levels of the multiple groups used for calibration. Recall, Kingston and Dorans carried out their study using random groups of examinees from the same administration; thus, differences in group ability would not be a factor in the results they observed. It seems logical to conclude that, if it is possible to obtain results such as those described by Kingston and Dorans for the Analytical measure of the GRE, when groups are similar in level of ability, then a design such as Eignor's, where item parameter estimates (and ultimately the equating based on these estimates) are subject to both location effects and differences in group ability, is tenuous to say the least.

The IRT studies reviewed reached varying conclusions as to the effect of anchor test length on equating results. The Vale et al. (1981) and the Raju et al. (1983) studies looked at linking test size in the context of separate calibration runs. Vale, et al. suggest that at least 15 items may be necessary while Raju et al. found adequate linking with as few as 6 common items. The McKinley and Reckase (1981) and Wingersky and Lord (1984) studies looked at linking test size in the context of concurrent calibration. McKinley and Reckase suggest that 15 items are needed for adequate linking, while Wingersky and Lord suggest that as few as five items may be needed. The Wingersky, Cook, and Eignor (1986) study suggests that one obtains improved results for both a concurrent calibration procedure and a characteristic curve

transformation method for longer linking tests. Furthermore, they found little differences in equatings based on item parameter scalings performed using the concurrent calibration procedure and the characteristic curve method. In addition, their results indicate that slightly more accurate equatings are obtained if a uniform, rather than a peaked distribution of ability is used to estimate parameters for the linking items.

The conventional equating studies reviewed clearly indicate that the properties of an anchor test are of greater concern as the samples used to equate tests diverge more in level of ability. The results of the study by Cook, Eignor, and Taft (1985) dramatically illustrate how equating results can be very disparate for different sets of common items if the groups used for equating differ in level of ability. On the other hand, these same common item sets yield results that exhibit a high level of agreement when the groups are similar in ability level. The results of the study by Klein and Kolen (1985) are similar to those obtained by Cook, Eignor, and Taft; i.e., anchor test length only became a serious factor when the groups differed in level of ability. The results of the study by Klein and Jarjoura (1986) indicate that content parallelism of the anchor test to the total test is an important factor when the two groups used for the equating differ in level of ability. Although the question was not specifically addressed in any of the IRT studies reviewed, it is likely that, similar to the results obtained for the conventional studies, the properties of linking items used for IRT item parameter scaling interact with differences in group ability.

It is clear, from the reviews of the Kingston and Dorans (1984) and Eignor (1985) studies that evaluated the affects of the properties of linking items on IRT true-score equating, that item position is a key factor. However, since it is usually possible for most equating designs to maintain the position of

linking items at least reasonably well, this factor only becomes a major concern for IRT pre-equating designs. As an aside, it is interesting to note that, although the authors of this paper cannot cite a formal study, it is an established practice when using conventional equating, to maintain the position of anchor test items across new and old test forms that are to be equated. One suspects that if IRT true-score equating is so affected by item location, conventional equating results probably are similarly affected.

## CONCLUSIONS

In this paper, we have discussed how various equating methods are affected by (1) sampling error, (2) sample characteristics, and (3) characteristics of anchor test items. Unfortunately, we are unable to offer, as an outcome of this discussion, a cookbook procedure for obtaining accurate equating results. Instead, we will try to make some practical suggestions for practitioners to follow when equating and attempt to describe some needed research.

Sparcity of data for equating is a common problem for testing programs that offer tests for certification or licensure purposes or that offer tests that measure competence in a particular subject area. And, the small sample problem is exacerbated when the tests are long. Some research has been done that suggests that the accuracy of equipercentile equating may be improved through the use of analytic techniques for smoothing or modeling marginal and bivariate frequency distributions. More formal research in this area is needed. There is a need for simulation and empirical studies designed to evaluate the effectiveness of analytic smoothing techniques for recovering the underlying distribution when sample size, test length, and distributional shape, are varied. Such research could lead to guidelines for the

practitioner and, thus, help to eliminate some of the "subjectiveness" in operational equipercentile equating decisions.

Definitions of equated scores require that the equating transformation be the same regardless of the group from which it is derived. Some of the studies reviewed suggest that this may not be too serious a problem for forms of a homogeneous test that are constructed to be as similar as possible in all respects. However, most testing programs offer their tests on multiple dates throughout the year; and, most of the studies reviewed have not examined whether their results would hold up if the same test forms were re-equated using samples from a different administration date. Examinees who take a test on different administration dates are self-selected and, thus, may vary in systematic ways (such as in recency of relevant coursework) which may be related to test performance. This is probably a greater problem for achievement-type tests than for aptitude-type tests. As the Cook, Eignor, and Taft (1985) study indicated, an achievement test may measure different skills and abilities for groups taking the test on different administration dates. More testing programs need to evaluate the extent to which their operational equating results are population invariant. They need to examine the similarity in equating results for major population subgroups taking test forms at different administration dates. If the resulting equating transformations are not the same, it may be necessary to take this into account when selecting samples and administration dates for conducting equating. It may also be necessary to explicitly describe, to test score users and recipients, the characteristics of the group for whom scores can be considered to have the same meaning.

The effectiveness of an anchor test depends on how closely the test mirrors, in content and statistical properties, the characteristics of the

total tests that are to be equated. The studies we reviewed indicate that the extent to which conventional equating results are altered by departures from this ideal situation depends, to a large extent, on differences between the level and dispersion of ability of the equating samples. It is apparent that when groups vary in ability (the typical anchor test situation), special care must be taken to ensure that the anchor test is a miniature of the total test. It is also apparent from results of studies, such as that done by Cook, Eignor, and Taft (1985), that anchor test items should be examined to determine if they are differentially difficult for the new and old form groups.

The studies that examined the affect of the characteristics of the linking items on IRT true-score equating focused on several different properties of the items. The most compelling results were those found in the Kingston and Dorans (1984) and Eignor (1985) studies. The results of these studies indicate that the scaling of the parameter estimates depends, to some extent, on the relative position of the linking items in the new and old forms of the test.

Results of the IRT equating studies that examined the affect of the number of linking items, the size of the standard errors of estimation of the item parameters, and the type of scaling procedure, did not provide clear guidelines. It is important to note that none of these studies systematically investigated (as was done in the conventional anchor test studies) the affect on item parameter scaling of (1) lack of content representativeness of the linking items or (2) the differences in ability levels of the new and old form groups. An investigation of how both of these factors affect item parameter scaling and their interaction with the number of linking items, the position of the linking items, and the scaling procedure used, would certainly be worthwhile. In the absence of such a

study or studies, it would appear that, when equating tests are administered to groups that differ in level of ability, regardless of whether one is using IRT true-score or conventional equating procedures, one should choose common items that are a miniature of the total tests to be equated, and make sure that these items remain in the same relative position when administered to the new and old form groups. It would also seem prudent to evaluate the differential difficulty of the common items administered to the equating samples, particularly when equating samples come from different administration dates.

The practical issues that we have chosen to address in this paper are among those that we frequently encounter in our ongoing work at Educational Testing Service. We have tried to address these issues using available research and our own personal experience. In so doing, a number of additional IRT and conventional equating issues in need of clarification have become apparent. We are optimistic that as more equating research is done and as more psychometricians gain practical equating experience, many of the issues will be resolved.

REFERENCES

Angoff, W. H. (1984). Scales, norms, and equivalent scores. Princeton, NJ: Educational Testing Service. (Reprint of chapter in R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, DC: American Council on Education, 1971.)

Angoff, W. H., & Cowell, W. R. (1985). An examination of the assumption that the equating of parallel forms is population independent (RR-85-22). Princeton, NJ: Educational Testing Service.

Brandenburg, D. C., & Forsyth, R. A. (1974). Approximating standardized achievement test norms with a theoretical model. Educational and Psychological Measurement, 34, 3-9.

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

Cook, L. L. (1984). Equating refurbished achievement tests. Unpublished Statistical Report. Princeton, NJ: Educational Testing Service.

Cook, L. L., Eignor, D. R., & Taft, H. (1985). A comparative study of curriculum effects on the stability of IRT and conventional item parameter estimates (RR-85-38). Princeton, NJ: Educational Testing Service.

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

Eignor, D. R. (1985). An investigation of the feasibility and practical outcomes of pre-equating the SAT verbal and mathematical sections (RR-85-10). Princeton, NJ: Educational Testing Service.

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

Henrysson, S. (1971). Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, DC: American Council on Education.

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

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

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. Applied Psychological Measurement, 8, 147-154.

Kingston, N., Leary, L., & Wightman, L. (1985). An exploratory study of the applicability of item response theory methods to the Graduate Management Admissions Test (RR-85-34). Princeton, NJ: Educational Testing Service.

Klein, L. W., & Jarjoura, D. (1986). The importance of content representation for common-item equating with nonrandom groups. Journal of Educational Measurement, (in press).

Klein, L. W., & Kolen, M. J. (1985). Effect of number of common items in common-item equating with nonrandom groups. Paper presented at the annual meeting of AERA, Chicago.

Lord, F. M. (1950). Notes on comparable scales for test scores (RE-50-48). Princeton, NJ: Educational Testing Service.

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

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

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". Applied Psychological Measurement, 8, 453-461.

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.

Raju, N. S., Edwards, J. E., & Osberg, D. W. (1983). The effect of anchor test size in vertical equating with the Rasch and three-parameter models. Paper presented at the annual meeting of NCME, Montreal.

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

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

Tapia, R. A., & Thompson, J. R. (1978). Non-parametric probability density estimation. Baltimore, MD: The Johns Hopkins University Press.

Vale, C. D., Maurelli, V. A., Gialluca, K. A., Weiss, D. J., & Ree, M. J. (1981). Methods for linking item parameters (AFHRL-TR-81-10). Brooks AFB, TX: Air Force Human Resource Laboratory.

Wingersky, M. S., Cook, L. L., & Eignor, D. R. (1986). Specifying the characteristics of linking items used for the item response theory item calibration. Paper presented at the annual meeting of AERA, San Francisco.

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. Applied Psychological Measurement, 8, 347-364.

B:LCNPNCME.REF

TABLE 1

Raw Score Summary Statistics for Biology Tests and Common Item Sets*

OLD FORM (FALL SAMPLE)

| Test | n | N | Mean | Standard Deviation | Equating Test-Total Test Corr. |
|---|---|---|---|---|---|
| Old Form | 99 | 2408 | 46.33 | 18.26 | |
| Anchor Test | 58 | | 25.62 | 11.42 | .96 |

NEW FORM (SPRING SAMPLE)

| Test | n | N | Mean | Standard Deviation | Equating Test-Total Test Corr. |
|---|---|---|---|---|---|
| New Form | 95 | 3892 | 53.71 | 17.61 | |
| Anchor Test | 58 | | 32.89 | 11.42 | .97 |

NEW FORM (FALL SAMPLE)

| Test | n | N | Mean | Standard Deviation | Equating Test-Total Test Corr. |
|---|---|---|---|---|---|
| New Form | 95 | 3653 | 44.74 | 17.56 | |
| Anchor Test | 58 | | 25.65 | 11.27 | .96 |

*Raw score summary statistics for the two new form samples may be directly compared. However, due to differences in test difficulty and test length, comparisons of these statistics should not be made with those obtained for the old form of the test.

TABLE 2

Biology Test Scaled Score Summary Statistics Resulting from Equating Method/Equating Sample Combinations

| | Equating Method | | | | | |
|---|---|---|---|---|---|---|
| | Linear | | Equipercentile | | IRT | |
| Sample Combination | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| Spring-New-Form/ Fall-Old-Form | 585 | 104 | 582 | 102 | 586 | 102 |
| Fall-New-Form/ Fall-Old-Form | 569 | 103 | 567 | 103 | 568 | 103 |

TABLE 3

Biology Test Raw to Scale Linear Conversions Resulting from Fall and Spring New-Form/Old-Form Combination

| Raw Score | Fall New Form/ Fall Old Form | Spring New Form/ Spring Old Form |
|---|---|---|
| 100 | 790 | 800 |
| 90 | 740 | 750 |
| 80 | 680 | 690 |
| 70 | 630 | 640 |
| 60 | 570 | 580 |
| 50 | 520 | 530 |
| 40 | 470 | 470 |
| 30 | 410 | 420 |
| 20 | 360 | 370 |
| 10 | 310 | 310 |
| 0 | 250 | 260 |

TABLE 4

Scaled Score Summary Statistics Resulting From Equating Method/ Common Item Set/Equating Sample Combinations*

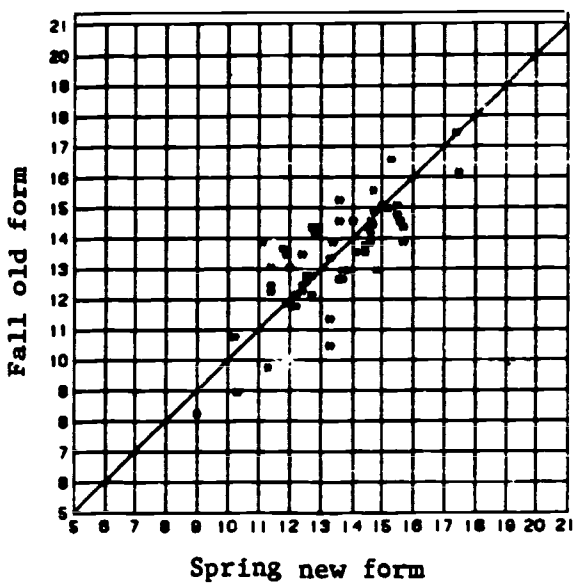
Spring New Form Sample/Fall Old Form Sample

| Common Item Equating Section | Equating Method: Linear Mean | Linear S.D. | Equipercentile Mean | Equipercentile S.D. | IRT Mean | IRT S.D. |
|---|---|---|---|---|---|---|
| 58 items | 585 | 104 | 582 | 102 | 586 | 102 |
| 36 items | 579 | 102 | 574 | 102 | 581 | 103 |
| 29 items demonstrating smallest difference in delta values | 539 | 103 | 541 | 103 | 545 | 102 |
| 29 items demonstrating largest difference in delta values | 624 | 105 | 608 | 99 | 619 | 97 |

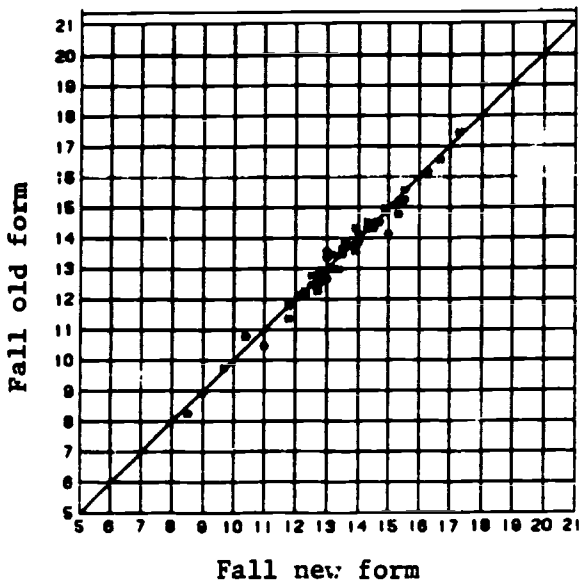
Fall-New-Form-Sample/Fall-Old-Form-Sample

| Common Item Equating Section | Equating Method: Linear Mean | Linear S.D. | Equipercentile Mean | Equipercentile S.D. | IRT Mean | IRT S.D. |
|---|---|---|---|---|---|---|
| 58 items | 569 | 103 | 567 | 103 | 568 | 103 |
| 36 items | 570 | 102 | 570 | 102 | 570 | 102 |
| 29 items demonstrating smallest difference in delta values | 567 | 101 | 567 | 102 | 567 | 102 |
| 29 items demonstrating largest difference in delta values | 570 | 102 | 570 | 103 | 569 | 103 |

*Raw score frequency distributions used to compute scaled-score summary statistics were obtained from spring total group (N=23,405).

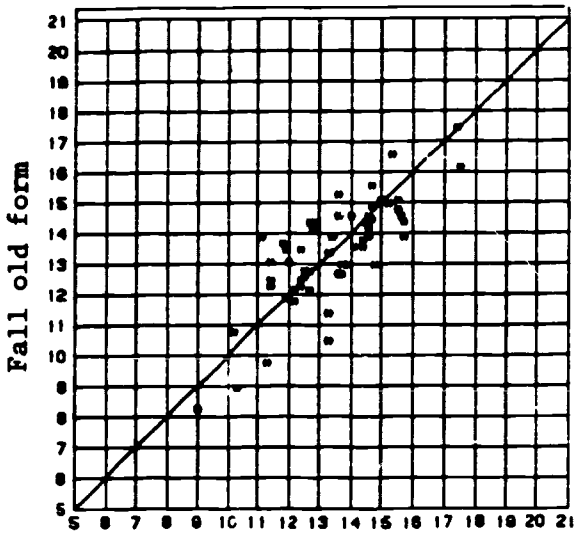



r=.79



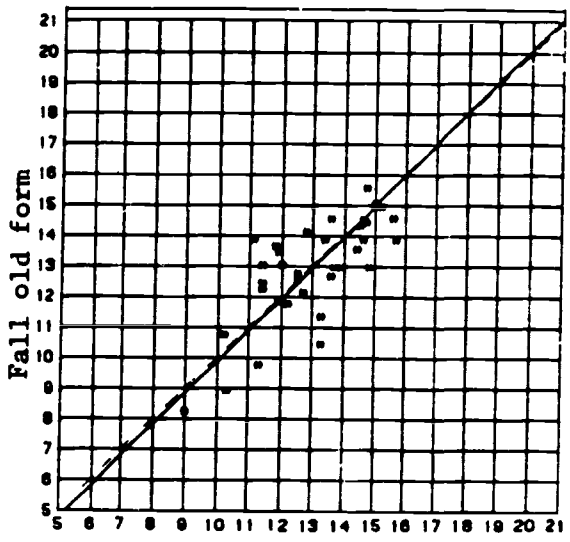

r=.99

Figure 1: Biology test plots of equated delta values for spring and fall new form and fall old form samples for 58 common items.



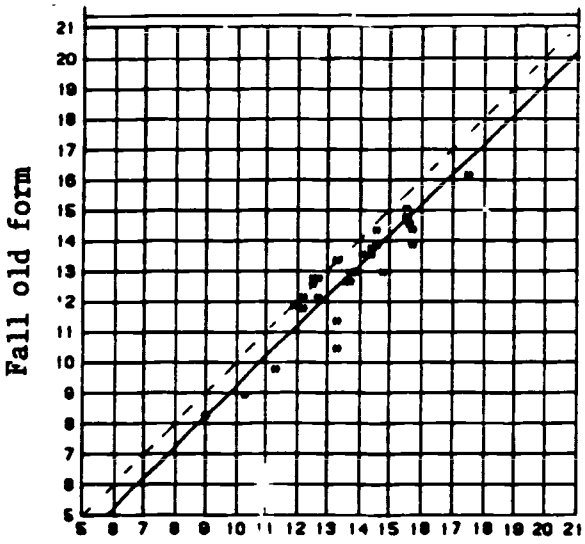

58 common items

r=.79



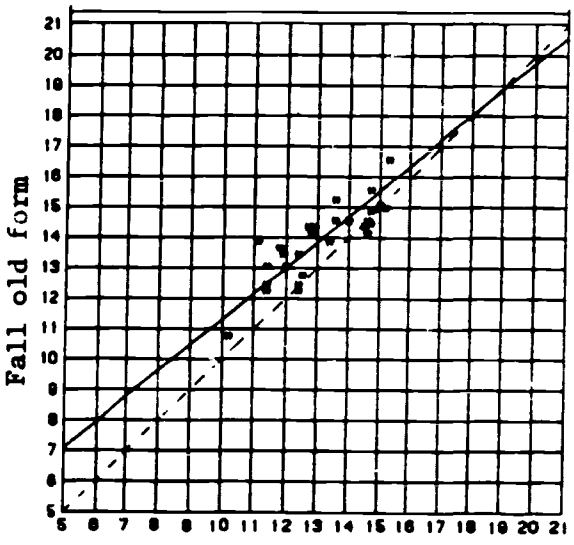

36 common items chosen by content experts

r=.74




29 common items demonstrating smallest differences in delta values

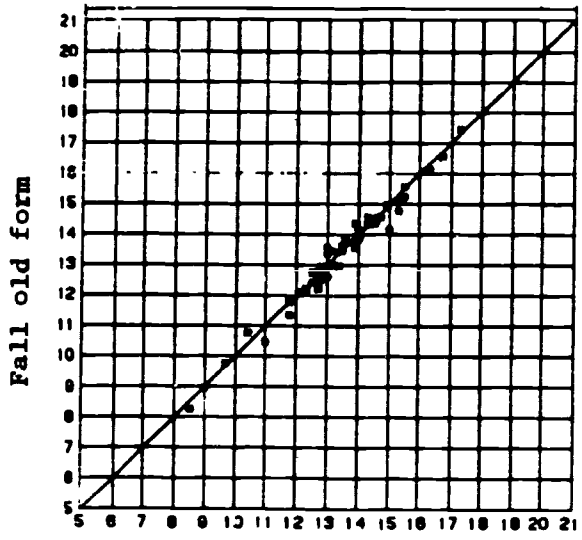
r=.92




29 common items demonstrating largest differences in delta values

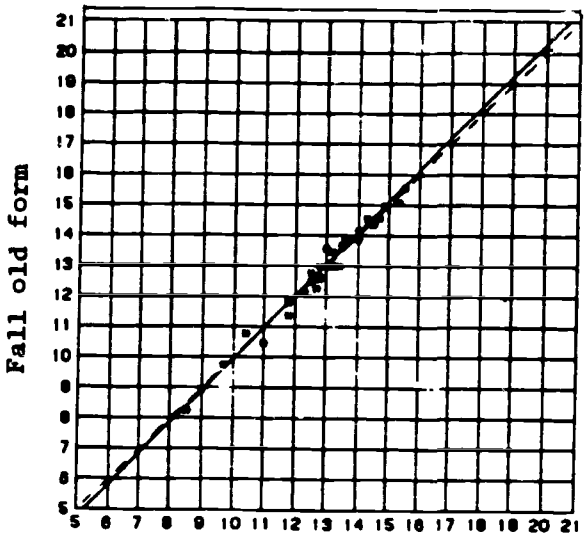
r=.87

Figure 2: Biology test plots of spring new form equated deltas versus fall old form equated deltas for the 58 common items and the three subsets.



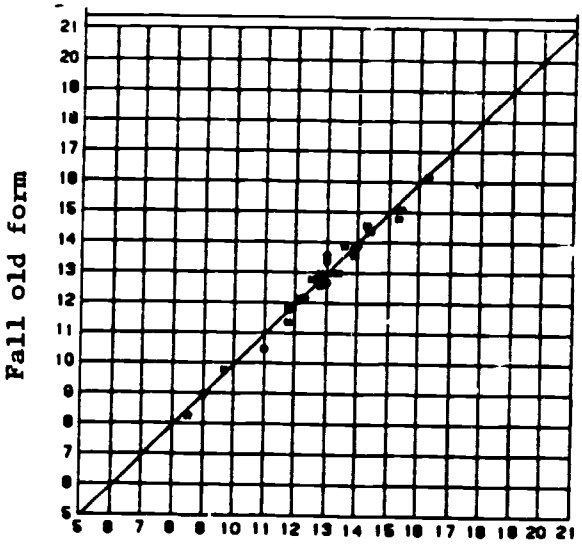

58 common items

r=.99



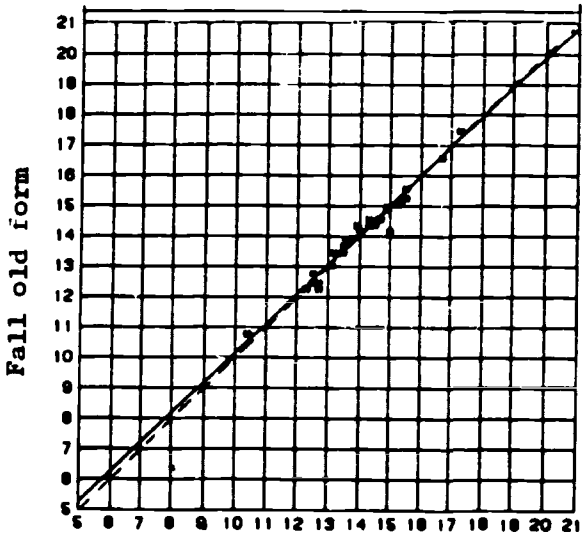

36 common items chosen by content experts

r=.99




29 common items demonstrating smallest differences in delta values

r=.99




29 common items demonstrating largest differences in delta values

r=.98

Figure 3: Biology test plots of fall new form equated deltas versus fall old form equated deltas for the 58 common items and the three subsets.