ED 441 842 TM 030 877

AUTHOR	Yang, Wen-Ling
TITLE	The Effects of Content Homogeneity and Equating Method on the Accuracy of Common-Item Test Equating.
PUB DATE	2000-04-00
NOTE	30p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Equated Scores; Item Response Theory; Licensing Examinations (Professions); Physicians; *Scoring; *Test Content; Test Items
IDENTIFIERS	Accuracy; *Anchor Tests; *Common Item Effect

ABSTRACT

This study investigated whether equating accuracy improves with an anchor test that is more representative of its corresponding total test and whether such content effect depends on the particular equating method used. Scoring outcomes of a professional examination for a medical specialty were used. A total of 1,092 examinees took one form, and 1,149 took the other form. Item-sampling schemes were introduced to manipulate the content homogeneity and representation of anchor items so that their effects on equating accuracy could be studied with varied equating methods. Multiple criteria were proposed to evaluate equating accuracy. Overall, the study found that all the equating methods employed had accurate results to a moderate degree. They all produced more accurate results when the anchor items were more representative of their total test or when the content coverage of a test concentrated on fewer topics. In general, Item Response Theory-based equating was found to be more accurate than classical equating, but the small differences may not be substantial. To produce equating outcomes that are precise and less prone to bias, various equating methods should be considered for equating test forms for high-stakes examinations. (Contains 4 tables, 3 figures, and 22 references.) (SLD)

The Effects of Content Homogeneity and Equating Method On the Accuracy of Common-Item Test Equating

By

Wen-Ling Yang, Educational Testing Service

April, 2000

Paper prepared for the presentation at the Annual Meeting of the American Educational Research Association, New Orleans.

The Effects of Content Homogeneity and Equating Method On the Accuracy of Common-Item Test Equating

Introduction

Often in educational testing not all examinees take the same test on the same occasion. Alternate test forms with comparable scores therefore are needed to ensure test security. Equating scores on alternate test forms is crucial because of imperfect test construction, which fails to achieve comparable scores for alternate test forms. The equating practice is especially important for high-stake certification or licensure examinations. It is usually not practical to have or assume equivalent groups taking alternate test forms. As a result, one popular equating practice is to embed a set of common anchor items in alternate test forms. Various techniques were developed to address the need of the common-item equating design. However, characteristics of the set of common anchor items such as its length and content representation may influence equating outcomes (Budescu, 1985; Harris, 1991; Klein & Jarjoura, 1985; Petersen, Marco, and Stewart, 1982). Review of literature further suggested insufficient evidence of adequate equating methods and quality equating outcomes. It is mainly due to the lack of an absolute criterion in practice for evaluating equating accuracy (Harris & Crouse, 1993).

One purpose of this study is to investigate how much improvement in equating accuracy is obtained when an internal anchor set of items becomes more representative of the total test. Another purpose is to study to what degree such content effect depends on a particular equating method. The ultimate goal of this study is to improve test results and inform important educational decisions. Particularly, this study is to draw attention to equating tests with skewed score distributions. Tests without normal score distributions generally receive less attention than they should have. Given these purposes, the design of this study has these features:

- Sampling items from a large pool of items included in a real test, such that the item pool is conceptually an item universe for all the resulting sampled test forms.
- Manipulating the content representation of the anchor items of various item samples.
- Using multiple equating methods and multiple criteria for evaluating equating accuracy.

Four pairs of subtest forms were assembled with items sampled using four different sampling schemes. The item sampling resulted in sets of anchor items more or less representative of the total subtest forms. Each pair of the subtest forms were equated using both the classical and the IRT-based equating methods. The sampling of items from two original test forms and the set of common anchor items embedded in the two original forms enabled us to conceptualize an anchor universe consisting of all common anchor items in the original test. Scores computed for items in the anchor universe represent pseudo true scores, and they are used to evaluate the quality of equating outcomes produced by various common-item equating methods for various pairs of subtest forms. The design of this study offered an alternative way to evaluating equating outcomes by forming an absolute evaluation criterion (the pseudo true score) for assessing equating accuracy. We hope that not only the research findings will inform testing practice about the selection of anchor items and equating methods, but also future research can gain insights from the design of this study.

Note that this study includes an extensive review of literature, including the conditions of equivalency and equating guidelines, the assumptions and procedures of various equating methods, the features of the common anchor item equating, and the estimation of equating accuracy. Due to the space limit, however, the review is not presented in this paper.

Description of Data

Scoring outcomes of a professional in-training examination for a medical specialty were analyzed in this study. The observed scores on the two alternate forms of the test had negatively skewed score distributions because the test measured minimum competency of the medical specialty. The two test forms had 197 and 203 items respectively. A total of 1,092 examinees took the form with 197 items and other 1,149 examinees took the other form. The two groups taking different test forms were not randomly formed. These examinees were mostly students or graduates from medical schools in the US. All of the items in the two test forms were five-option multiple-choice items. Examinees' responses on individual items were scored as right or wrong and assigned with numerical values of 1 and 0 respectively. A set of 145 anchor items appeared in the same locations of the two test forms. The format and wording of these anchor items do not vary across alternate test forms.

A preliminary inspection on the test data showed that one examinee group scored slightly higher on the 145 anchor items (mean=107.721, s.d.=13.113) than another group (mean=105.457, s.d.=13.767). Therefore, it was likely that the two groups differed slightly in their abilities. Despite the limited demographic information for the examinee groups, sampling effects such as years of experience in medicine were examined to compare group differences in this study. The test takers took the medical test while participating in various in-training programs to prepare for the formal board certification examination. The passing standard for the test required a minimum of 75% of the test items be correctly answered. In general, these test takers were highly motivated to pass the in-training exam, because the in-training test provided them a valuable opportunity to familiarize themselves with the formal certification exams, it was assumed that the candidates had taken the test seriously. After receiving instruction and training from the in-training programs, most of the test takers were expected to master the knowledge or skills being tested.

The test data analyzed in this study generally met the requirements for equating. The test forms had sufficient number of items and the subtest forms created via item sampling were all reasonably long. The test items were administered and secured under standardized conditions. Some items had been administered in previous years under the same standardized testing situations and found to be satisfactory. The stems, alternatives, and stimulus materials for the common anchor items were identical for the two test forms. The scoring keys were clear and consistent for the two forms. The overall examinee group was reasonably large, exceeding 2,200 subjects.

In addition to its skewed score distributions, the test data is also distinctive for the hierarchical structure of its content. Although all test items in the medical test were from a single medicine-content domain, these items covered twenty-three core content areas that delineated the big medicine domain. In other words, the content of these test items represents categories of core knowledge or skills for the overall medical specialty. The hierarchy of the content specification of the test allowed us to study, via data manipulation, the effects of content homogeneity and content representativeness of anchor items on equating outcomes.

Design and Methodologies

In this section, we briefly summarize the basic design of this study in this section, including the item-sampling schemes, the equating methods, and the criteria used for evaluating the accuracy of equating outcomes. The tools used for various equating and analyses were also mentioned. In addition, major limitations for the present research are discussed.

Item Sampling

Four item-sampling schemes were used to create four pairs of alternate subtest forms. Each pair of the subtest forms shared a common set of anchor items. The item-sampling schemes were developed to manipulate the content homogeneity and the content representation of anchor items for the subtest forms. They include:

- Simple random sampling.
- Equal-weight domain random sampling, which randomly sampled equal number of items from each of the 23 core content areas.
- Proportional-weight domain random sampling, which randomly sampled a number of items proportional to the size of a core content area.
- Purposeful sampling, which simply included all of the items from the three largest core content areas.

As the original alternate test forms, these alternate subtest forms all had negatively skewed score distributions. To control for possible interaction effect between test length and content homogeneity, all of the subtest forms were created to have similar test length and anchor length.

Equating Method

To investigate to what extent the results of various equating approaches agree, this study estimated and compared the effects of linear, equipercentile, and IRT-based equating approaches on the accuracy of equating. The linear equating methods are straightforward and convenient in computation (Kolen & Brennan, 1987), but their results do not always meet all criteria for equivalent tests. Equipercentile equating is a frequently used non-linear equating approach, which is still based on observed score and has been known to have accurate results (Hills, Subhiyah, & Hirsch, 1988; Yen, 1985). Becoming popular in recent years, equating models based on the Item Response Theory have been found useful for equating using the common-item design (Cook & Eignor, 1991; Crocker & Algina, 1986).

Each pair of the resulting subtest forms from a particular item-sampling scheme were equated using four different methods for common-item equating. For the IRT-based equating methods, items and scores of the subtest forms were calibrated using a three parameter logistic (3-PL) IRT model with the assumption of unidimensionality. These four equating methods are:

- Tucker Linear Equating method (Braun & Holland, 1982; Kolen & Brennan, 1987), which represents classical linear equating in this study.
- Frequency-Estimation Equipercentile equating method with the Cubic-Spline Post Smoothing (Kolen & Brennan, 1995), which represents classical non-linear equating.
- The IRT-based Linear Transformation method (Hambleton & Swaminathan, 1990; Kolen & Brennan, 1995).
- The IRT Fixed-b method (Petersen, Cook, & Stocking, 1983; Hills, Subhiyah, & Hirsch, 1988).

Under the assumptions of the Tucker method, the linearly transformed scores on one test form will have the same mean and standard deviation as the scores on another form (Kolen & Brennan, 1987; Kolen & Brennan, 1995). Although the Tucker method requires equal reliability across test forms being equated, the impact of unequal reliability was not very critical in this study because each pair of the sampled test forms being equated were very similar in their content and had the same number of items. Based on observed score, the equipercentile equating aims at finding a score on one form of a test that has the same percentile rank as a score on the other form of the same test (Kolen & Brennan, 1995). The Frequency-Estimation method with Cubic-Spline smoothing technique were used because the two examinee groups were not too different.

The IRT-based equating methods have both theoretical advantages (Green, Yen, & Burket, 1989) and practical appeals (Cook & Eignor, 1991). The IRT-based equating methods were found to be more useful than linear equating methods when tests to be equated differed somewhat in content and length (Petersen, Cook, & Stocking, 1983). However, IRT-based equating results are often inconsistent, and the practical significance of their improvement in estimating equating accuracy remains unclear. Under the IRT assumptions of item and person invariance, linear transformation is reasonable for the non-equivalent-group anchor-item design of this study. This is because the difficulty and discrimination parameters for the common anchor items from alternate test forms are linearly related (Petersen, Cook, & Stocking, 1983; Hills, Subhiyah, & Hirsch, 1988). The fixed-b method is often used with LOGIST program (Hills, Subhiyah, & Hirsch, 1988). However, it is found that the fixed-b method also worked nicely with BILOG program (Yang, 1997). It was found that equating results yielded by the fixed-b method using BILOG were consistent with the results from IRT-based linear transformation method.

Criteria for Evaluating Equating Accuracy

To estimate the accuracy of equating outcomes from various equating methods across a variety of subtests, equating outcomes were compared to these four criterion measures:

- The equipercentile equating relationship resulted from equating the two original test forms.
- The equipercentile equating relationship resulted from equating pairs of subtest forms, created via different item-sampling schemes.
- The IRT-estimated pseudo true scores based on all of the 145 anchor items embedded in the original test forms.
- The estimated pseudo true scores based on the 145 anchor items on raw-score scale.

The first two of the above criteria (based on the equipercentile equating relationship) represent the traditionally used arbitrary criteria for evaluating equating accuracy. The arbitrary criterion derived from the original test forms was considered the most reliable measure among the four criteria. It is because this criterion was based on equating substantially longer test forms with more anchor items. The arbitrary criterion derived from the subtest forms was used to illustrate the potential bias due to the nature of an arbitrary evaluation criterion.

The last two criteria (based on the pseudo true scores) were innovative but specific to the design of present study. They were appropriate when examinee population, test items and testing occasion were considered fixed. These two evaluation criteria have the theoretical appeal because they were composed using all of the common-anchor items in the "anchor universe". Given the design of item sampling for creating subtest forms, the set of the 145 anchor items in the original test forms was conceptually the "anchor universe" for all the anchor item sets embedded in the resulting subtest forms. Since all of the examinees took all of the 145 common-anchor items, the estimated true scores on the 145 anchor items was appropriate in representing examinees' true scores. Although the criterion based on the pseudo-true-score on the raw-score scale does not possess any advantages from the Item Response Theory, it was chosen for this study because it might not overestimate the accuracy for IRT-based equating.

To measure the accuracy of equating, various equating outcomes were correlated to the respective criterion measures discussed above. Pearson Product Moment Correlation Coefficient (r) was computed to represent the degree of equating accuracy. The artifact of auto-correlation (an inflation in the correlation coefficient), due to the overlapping of items in the criterion and the equating outcome, was controlled. The second index used for measuring equating accuracy was the Root-Mean-Squared-Deviation (RMSD) (Klein & Jarjoura, 1985; Kolen & Harris, 1990; Livingston, Dorans, & Wright, 1990; Schmitt, Cook, Dorans, & Eignor, 1990). Equating outcomes from various equating methods were also correlated to each other for comparison purposes. Correlational plots were used to aid visual inspection.

Analysis Tools

The computer program used for IRT calibration is BILOG-MG, which yields marginal maximum likelihood (MML) estimates. The advantage of using this program is that the number of parameter estimates does not increase with the increasing number of examinees. Unix SAS and Excel were used to assist with data management, various computations required for equating and other statistical analyses. The equipercentile equating was facilitated by an extended version of the Common Item Program for Equating (CIPE) (Hanson, Zeng, & Kolen, 1995). The CIPE program is based on the frequency estimation methodology described by Kolen and Brennan (1995), and uses the Cubic Spline smoothing method (Kolen & Jarjoura, 1987; Kolen & Brennan, 1995) to post-smooth resulting equipercentile relationship. Standard errors of equivalent scores yielded by CIPE were graphed using Excel. Excel was also used for various sets of the smoothed equivalents yielded by a selection of eight degrees of smoothing parameters. The smoothness and conditions of "moment preservation" (Kolen & Brennan, 1995) were examined to determine an appropriate equating relationship between alternate test forms. These statistics were also computed by using Excel.

Research Limitations

The secondary nature of the data analyzed in this study limited the design of this study and the generalizability of study outcomes. Major limitations of this study regarding data manipulation, interpretation of equating accuracy and generalization of study outcomes are briefly summarized in this section.

Data Manipulation

The design of this study hinges on the test data available. As a result, the common-anchor-item design for equating was the only option for equating test forms in this study. Given the large number of anchor items in the original test forms, the subtest forms created in this study all had long anchors. It was thus difficult for this study to evaluate various effects on equating accuracy when only few anchor items were included in test forms.

Due to the uneven item distribution across the 23 core content categories of the original test forms, it was difficult for this study to create subtest forms with the same number of items and the same number of anchor items. Despite out effort of maintaining similar test lengths and sufficient anchor lengths across subtest forms, there was still a slight chance that differential anchor lengths influenced equating results. If the effect of anchor length were actually present, the effect of content representativeness of anchor items was likely to be confounded. Therefore, interpretations for such effects were made with cautions.

Interpretation of Equating Accuracy

The criteria used for evaluating equating accuracy and the indices used to represent equating accuracy had some inherent limitations. The consequence of using the arbitrary criteria for evaluating equating accuracy is self-evident. The major drawback is that such criteria do not address equating accuracy directly. Only the consistency between the criteria and the results of the other equating methods are measured. Therefore, the evaluation outcomes based on arbitrary criteria were interpreted with cautions.

As mentioned earlier, the pseudo-true-scores-based criteria were only appropriate when the examinee population and the testing occasion were considered fixed. That is, the criteria would only be effective for an equating situation where the examinees are from a population as same as or similar to the population for this study. The examinees also have to be tested under the same or a similar circumstance as the one for this study.

Generalization of Results

Sources of limitations for generalizing study outcomes include the characteristics of test items and test forms, the characteristics of examinees, and the equating models used. One important feature of the test analyzed in this study is its negatively skewed score distribution due to its minimum-competency test nature. Therefore, the study results based on such skewed score distribution should be carefully generalized to similar testing situations. As this study focused on a group of professionals in a medical field from the in-training programs, the results of this study should not be generalized to other populations that differ from the professional population in this study.

The IRT-based equating methods used in this study assume unidimensionality for the test forms being equating. Therefore, the equating results should not be generalized to testing context where multi-dimensionality prevails. In addition, generalization of the study findings should be limited to the context where the 3-PL IRT model applies. Similarly, generalization of the results of any other equating method should take into account the particular assumptions made by the method.

Results and Discussions

To ensure the quality implementation of the study design, characteristics of the subtest forms created from the two original test forms were inspected, examinee group differences were examined, and the adequacy of the 3PL IRT model for the IRT-based equating was scrutinized. These inspection outcomes are briefly summarized below, followed by a summary of various equating outcomes evaluated by respective criteria for evaluating equating accuracy.

Characteristics of Tests and Examinee Groups

Item and test analyses provided evidence of construct validity for all of the subtest forms created by item sampling. Specifically, these subtest forms had similar degree of reliability and moderate item difficulty. Table 1 presents the statistics for internal consistency as reliability measures for various subtest forms. Note that among the moderate reliability indices, the purposeful item-sampling scheme resulted in test forms with the highest internal consistency as expected. Also, each pair of the subtest forms had similar degree of internal consistency. The average item difficulty ranged from 0.688 to 0.759 across various subtest forms, indicating moderate difficulty level for an average item. The proportion of examinees answering an item correctly was used as the measure of item difficulty. In addition, the equal-variance two-tailed Student's *t*-test indicated that there were no statistically significant differences in average item difficulty between each pair of the subtest forms.

The desired item-sampling effect on content homogeneity and the content representation of anchor items were also confirmed. Items from each of the subtest forms were generally correlated to their total test positively and moderately, providing evidence of homogeneity. The equal-variance two-tailed Student's *t*-test further indicated that there were no statistically significant differences in item-total correlation between each pair of the subtest forms. In consistency with the item-sampling schemes, the purposeful sampling resulted in test forms with the strongest item-total correlation. Likely, the variation in the item-total correlation across various subtest forms was in part due to the item sampling effect, as planned by the study design.

Correlation between the score on the anchor items and the score on the unique items of a test ($r_{\text{anchor,unique}}$) not only provides empirical evidence for item homogeneity, the statistic can also be used as an index of efficiency for test equating (Budescu, 1985). The bigger the value of the $r_{\text{anchor,unique}}$, the more precise the parameters will be estimated for the combined group in equating. All of the indices of equating efficiency for the variety of subtest forms in this study were statistically significant, suggesting efficient common-item equating. The values of the $r_{\text{anchor,unique}}$ ranged from .44 to .54, indicating similarities between anchor items and the non-anchor items to a moderate degree.

Across various subtest forms, the score on the anchor test also correlated significantly with the total test score to a considerable degree. The observed values of the $r_{\text{anchor,total}}$ ranged from .861 to .968. Despite the inflation in the $r_{\text{anchor,total}}$ due to auto-correlation, the large value of the $r_{\text{anchor,total}}$ was considered robust. It provided evidence of content representativeness of anchor test. Despite the large values for all of the $r_{\text{anchor,total}}$, various anchors in this study were more or less representative of their corresponding total tests. This proves the success of the planned manipulation for the content representation of anchors. Specifically, the $r_{\text{anchor,total}}$ decreased as the content specificity of the subtest changed. For one subtest form, the $r_{\text{anchor,total}}$ decreased from .968 (Purpose Sampling) to .939 (Equal-Weight Domain Random Sampling) to .924 (Proportional-Weight Domain Random Sampling) to .861 (Simple Random Sampling). And for the alternate subtest form, it decreased from .968 (Purposeful Sampling) to .939 (Equal-Weight Domain Random Sampling) to .925 (Proportional-Weight Domain Random Sampling) to .863 (Simple Random Sampling).

Although studies for the demographic attributes of the two examinee groups indicated slight differences between the two examinee groups, the differences were not serious. Despite the limited availability of the data, these differences between the two examinee groups were examined: ability, years of experience, and program participation of the examinees. This study found that the examinee groups performed slightly differently on the same set of anchor items. For the 145 anchor items in the original test forms, one group (mean=107.721, s.d.=13.113) scored slightly higher than the other group (mean=105.457, s.d.=13.767). The difference was statistically significant at α=.05 significance level (t=3.987, df=2,239, p=.0001). For each of the four pairs of subtest forms, there was also significant ability difference between the two examinee groups. To further examine the group differences, average

Table 1 - Reliability (Internal Consistency) of Subtest Forms Created via Various Item-Sampling Schemes

Item-Sampling Scheme	Subtest Form	Cronbach's α
Simple Random Sampling	A	0.658
	B	0.713
Equal-Weight Domain Random Sampling	A	0.684
	B	0.690
Proportional-Weight Domain Random Sampling	A	0.662
	B	0.691
Purposeful Sampling	A	0.774
	B	0.768

item difficulties were computed for the anchor items and unique items respectively and the results were presented in Table 2. Across various subtest forms, one examinee group (the group taking Form B in Table 2) was associated with slightly larger values of item difficulty based on the anchor items. It confirmed that this group had slightly higher ability than the other group.

Despite the statistical significance found for the group differences, the differences between the two groups might not have practical significance. It is because all of the effect sizes for the group differences were relatively small, compared to the group means and standard deviations. The group differences could be attributed to the non-random selection or assignment of the examinees in the examination. It is noted that examinee-group disparity can be a threat to the accuracy of Tucker linear equating, and Levine equally reliable method is recommended for such case (Kolen & Brennan, 1987). The Tucker equating method was used in this study because it is found previously that the Tucker method and the Levine method yielded almost identical results for the original test forms.

The means and the standard deviations for the years of experience were very similar for the two examinee groups. Assuming equal variances (F=0.093, p=0.76), the two-tailed *t*-test for the group mean difference suggested no statistically significant difference (t=0.659, df=2,226, p=0.510). Therefore, the two examinee groups in this study had the same number of years of experience. The examinees taking one test form were from 64 different in-training programs, and those who took the other test form were in 62 different programs. In total, there were 109 programs. Most of the time, examinees from the same program took the same test form. Information was lacking for the studies of inter-program differences, such as differential curricular design or various instructional methods. As a result, group differences due to program variations could not be examined. Given more demographic information about the examinee groups, group differences and their potential influences on the results of equating could be more thoroughly examined.

Applicability of the 3-PL IRT Calibration

The adequacy of the 3PL IRT model for the IRT-based methods for equating test forms with negatively skewed score distribution was evaluated. The three-parameter IRT model incorporates a guessing parameter for item and score calibration. Based on a 3-PL IRT model, the IRT-based equating in theory has the advantage of accounting for the guessing factor, which is likely to be present for the minimum competency examination with multiple-choice items. In addition, the satisfactory equating outcomes of the two IRT-based methods (presented later in this paper) provide some empirical evidence of the applicability of the 3-PL IRT model.

Equating Outcomes

Intermediate results and final equating outcomes of various equating methods for each of the four sets of subtest forms were compared and summarized in tables and graphs. However, due to limited space, this paper will highlight only the equating outcomes of most interest.

IRT-based Equating

For each pair of the subtest forms, the equivalent ability estimates resulted from the IRT-based linear transformation method and the fixed-b method correlated strongly. It indicated similarities between the equating outcomes of these two methods. Across various subtests, the Pearson *r*s were all statistically significant and had values close to 1. It is therefore concluded that the two IRT-based equating methods

Table 2 - Examinee Group Differences in Average Item Difficulty Based on Anchor and Unique Items

Item-Sampling Scheme	Subtest Form	$\overline{p}_{\text{AnchorItem}}$	$\overline{p}_{\text{UniqueItem}}$
Simple Random Sampling	A	0.722	0.654
	B	0.736	0.670
Equal-Weight Domain Random Sampling	A	0.767	0.650
	B	0.781	0.705
Proportional-Weight Domain Random Sampling	A	0.720	0.683
	B	0.734	0.690
Purposeful Sampling	A	0.703	0.664
	B	0.728	0.691

Note. 1. The item difficulty (p_i) is defined to be the % of examinees getting item "i" correct, and the average item difficulty is:

$$\overline{p} = \frac{\sum_{i=1}^{n} p_i}{n},$$ where n is the total number of items.

2. In this study, 1,092 examinees took Form A, and 1,149 took Form B.

were very similar in determining individual examinee's standing in the entire examinee group. Nevertheless, the dependent-samples *t*-test showed that the mean difference between the ability estimates resulted from the two IRT-based methods was significant at a conservative significance level of $\alpha = 0.01$ ($p<0.001$). The conservative significance level was used to control for the total error rate due to the multiple hypothesis tests for various subtests. The significant test results suggested that the outcomes of the two IRT-based methods were not as close as indicated by the Pearson *r*s.

It is noted that the large values for the *t*-statistics for the significance tests were partly due to the small standard errors of the mean differences and the large sample sizes (degrees of freedom). Also, the effect sizes of the mean differences for all the four subtests were all very small. They suggested that the differences between the outcomes resulted from different IRT-based equating methods might not be practically significant. Plotting the resulting equivalent ability estimates of one IRT-based method against the estimates of the other method, the correlational graphs in Figure 1 illustrated the positive and strong relationship between the outcomes of these two equating methods. While the fairly solid straight lines in the plots suggested strong linear relationship, the thick and coarse lines suggested less than perfect relationship. Overall, at the two ends of the ability scale, the outcomes of the two IRT-based equating methods were more similar than the outcomes at the middle range of the ability scale.

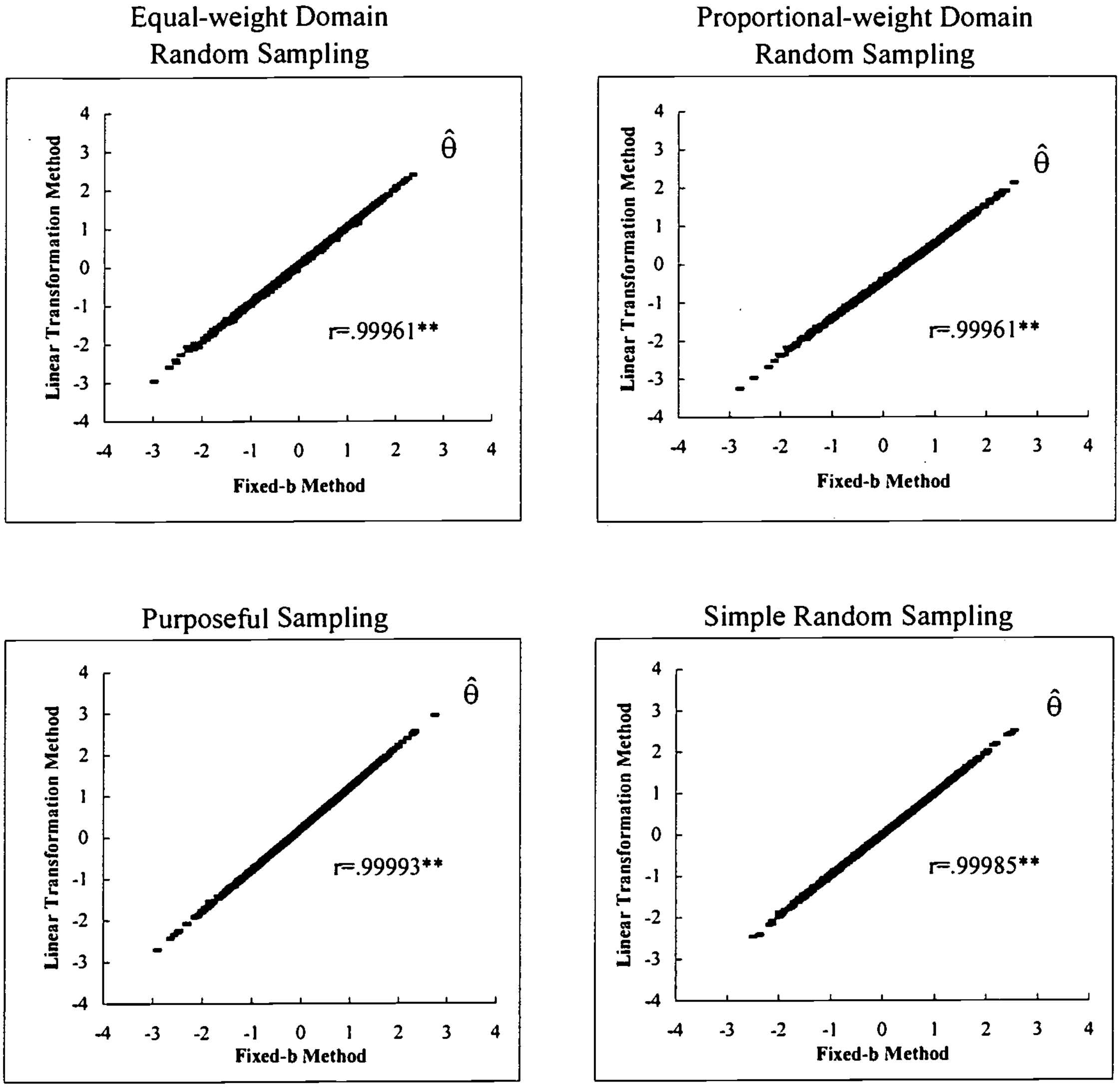
The true score estimates were obtained by applying the IRT true score formula (Lord, 1980) to the outcomes of the two IRT-based equating methods:

$$\hat{T} = \sum_{i=1}^{n} p_i(\theta) = \sum_{i=1}^{n} \{c_i + (1-c_i)/[1+Exp^{-1.7a_i(\theta-b_i)}]\} ,$$

where $\hat{T}$ is the estimated true score, $p_i(\theta)$ is the probability of getting item i correct given examinee ability θ, n is the number of items, a_i is the item discrimination for item i, b_i is the item difficulty for items i, and c_i is the pseudo-chance level (guessing) for item i (Hambleton & Swaminathan, 1990). As expected, for each subtest, the correlation between the resulting true-score estimates from the two IRT-based methods was fairly strong and statistically significant.

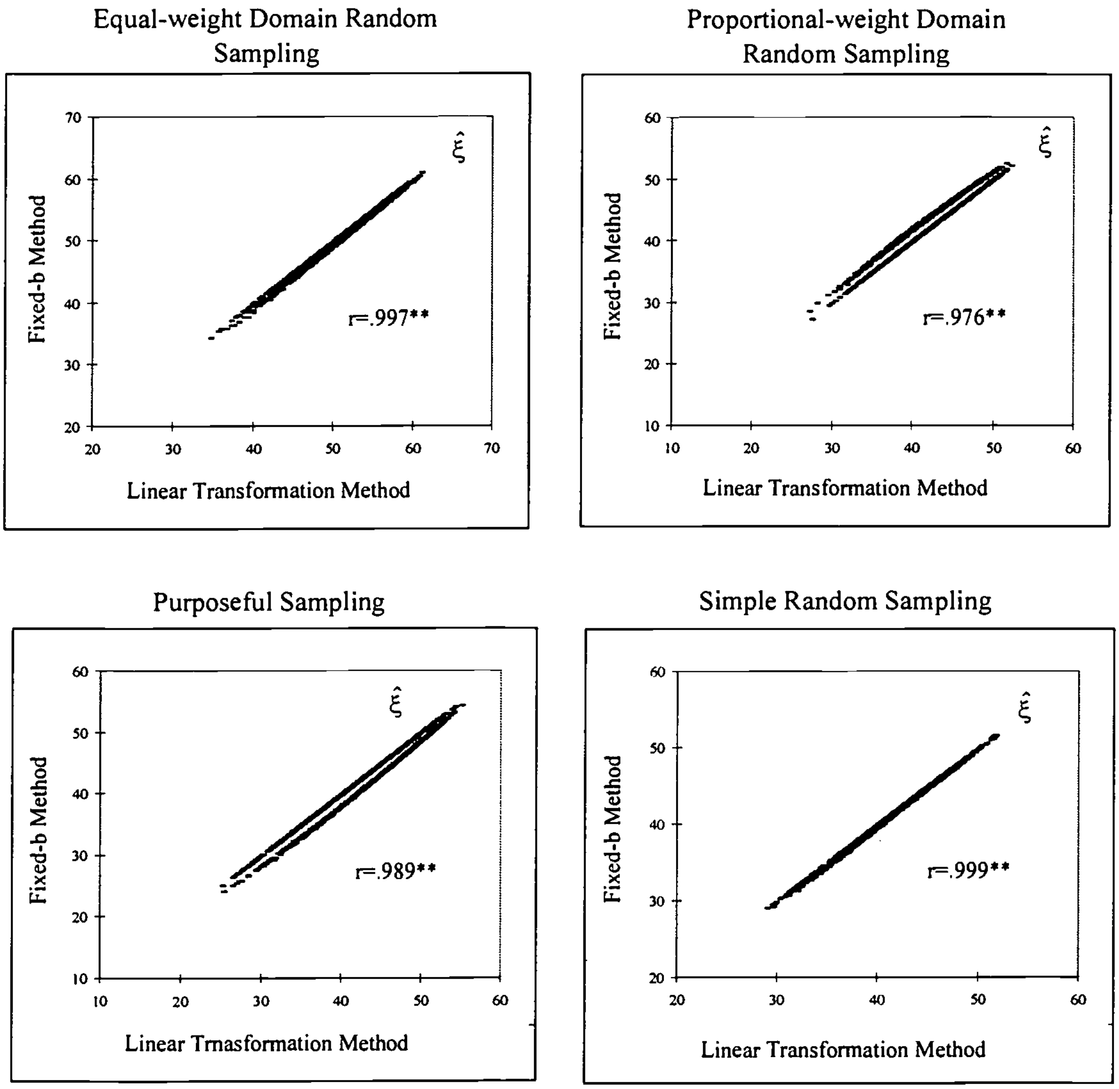
The correlational graphs in Figure 2 illustrated the relationship between the outcomes for the two IRT-based equating methods, based on the equivalent true-score estimates. As the plots in Figure 1, the graphs in Figure 2 showed positive and strong relationship between the outcomes of the two IRT-based methods. However, the plots in Figure 2 were more revealing in showing the differences between the two equating methods. While the scattered data points formed pretty solid straight lines for subtests created by Equal-Weight Domain Random Sampling and Simple Random Sampling, the data plots for subtests resulted from Proportional-Weight Domain Random Sampling and Purposeful Sampling clearly suggested more than one relationship lines. It suggested that the resulting true-score estimates of one method did not correspond to the true-score estimates of the other method on a one-to-one basis. In other words, a subgroup of examinees receiving the same score when one IRT-based equating method was used might receive different scores when the other IRT-based method was in use. An inspection on the resulting estimates of equivalent true scores yielded by the two IRT-based methods confirmed this phenomenon.

In addition, the formation of the phenomenal plots in Figure 2 looked linear but somewhat curvilinear. Like the plots in Figure 1, it suggested that the outcomes of the two IRT-based methods were



Note. (1) θ is the examinee ability; (2) **- Significance level less than 0.01.

Figure 1 - Relationship Between the Ability Estimates of the Two IRT-Based Equating Methods



Note. (1) ξ is the equivalent true score; (2) **- Significance level less than .01.

Figure 2 - Relationship Between the True Score Estimates of the Two IRT-Based Equating Methods

more similar for the cases receiving extreme scores than for the cases in the middle range of the true-score scale. When examinees scored either extremely high or extremely low, the two equating methods often ranked these examinees in the same order. The slightly non-linear relationship indicated that using the Pearson *r*, which is only appropriate for linear relationship, for summarizing the relationship between the outcomes of the two IRT-based equating methods could be misleading.

Smoothing the Equipercentile Equating Outcomes

This study used the frequency estimation method (Kolen & Brennan, 1995) for equipercentile equating. To increase equating precision, after obtaining the frequency-estimation equipercentile equivalent scores, this study applied the Cubic Spline Post-smoothing method (Kolen & Jarjoura, 1987; Kolen & Brennan, 1995) to smooth the equivalent scores. A total of eight smoothing parameters were specified (s=0.01, 0.05, 0.10, 0.20, 0.30, 0.50, 0.75, and 1) for post-smoothing. These parameters yielded smoothed equivalent scores differing in their degree of smoothing (Hanson, Zeng, & Kolen, 1995). They controlled the amount of the average squared standardized difference between the smoothed and the unsmoothed equating outcomes.

The resulting eight sets of smoothed equivalent scores were inspected graphically and statistically to determine which of the eight smoothing parameters resulted in the least amount of smoothing required for a smooth equipercentile equating function. For graphical inspection, each of the eight sets of smoothed equivalent scores was graphed with the set of unsmoothed equivalent scores, and a standard error band was constructed around the unsmoothed equating outcomes to facilitate visual inspection. The adequacy of various smoothed equating outcomes were in part judged by their smoothness and deviations from the unsmoothed equating outcomes, shown in the graphs. Figure 3 presents a set of eight graphs as an example of the graphical inspection.

In addition to the graphical inspection, the four moments--mean, standard deviation, skewness, and kurtosis--of the resulting smoothed equivalent scores were estimated to evaluate the smoothing requirement of “moment preservation” (Kolen & Brennan, 1995). The estimation outcomes for the moments of the eight sets of equivalent scores were summarized. The moments of the smoothed equivalent scores were compared to the moments of the unsmoothed equivalent scores such that the most appropriate smoothing parameter could be identified. An appropriate smoothing parameter will result in a smooth equipercentile equating function that does not depart too much from the unsmoothed equating outcomes.

Using the graphical inspection techniques and following the “moment preservation“ requirements, smoothing parameters were selected for various equipercentile equating outcomes. The final equipercentile equivalent scores yielded by these smoothing parameters appeared to be smooth and were not too far apart from the unsmoothed results. Their four moments were also close to those of the unsmoothed equivalent scores. Without introducing substantial bias into the smoothing process, the use of these smoothing parameters improved the precision of the equipercentile equating in estimating the equivalent scores (Kolen, 1991).

The smoothing requirement of "moment preservation" also requires that the moments of the equated scores on one form of a test to be close to those on the other form of the same test (Kolen & Brennan, 1995). This property is desired for both random group equating design and common-item non-equivalent group design. However, for the non-equivalent group design used in this study, it is a lot more difficult to examine this property and the interpretation will not be as clear as for the random

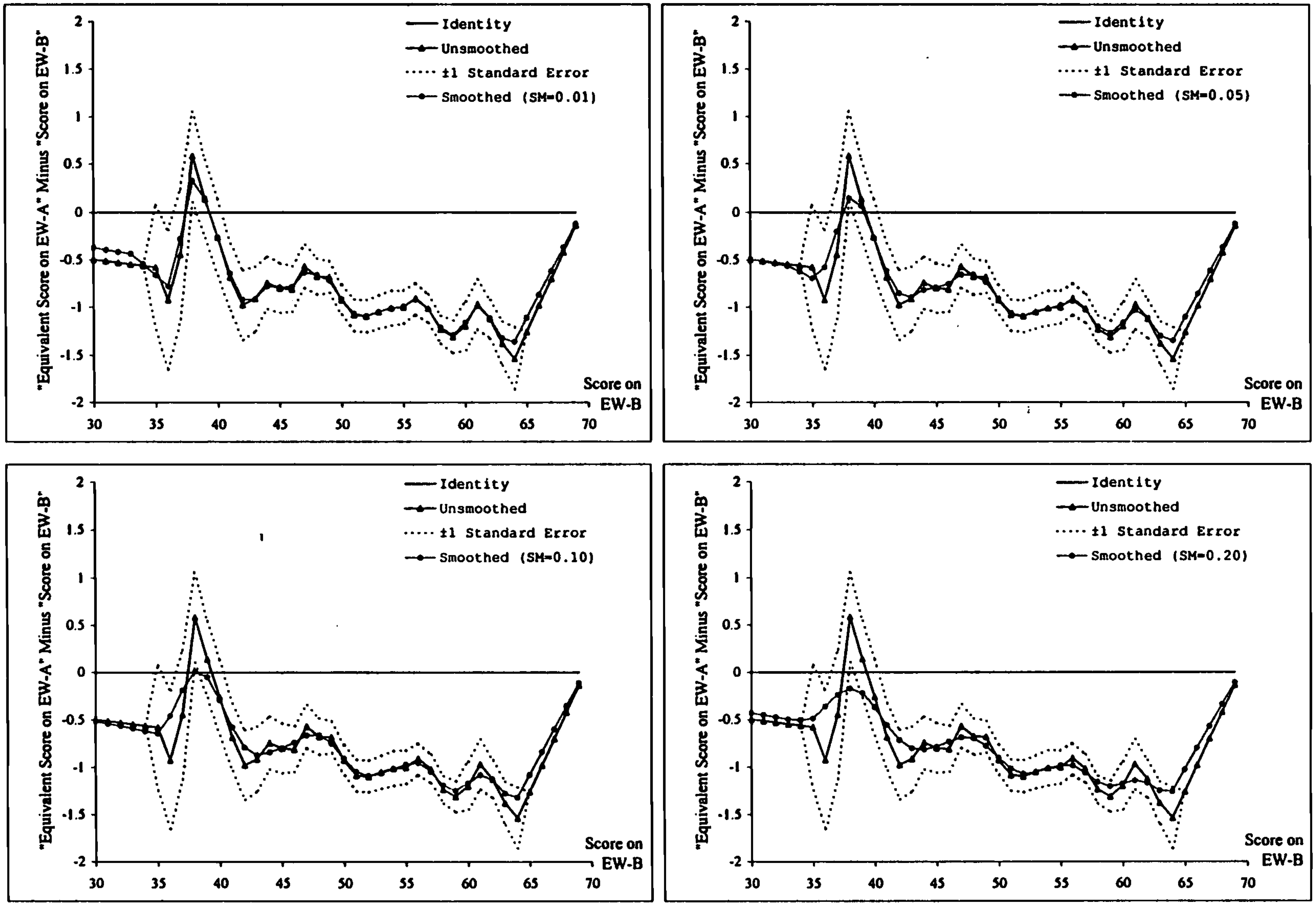


Figure 3 - An Example of the Score to Score Equivalents by Various Degrees of Smoothing for a Subtest

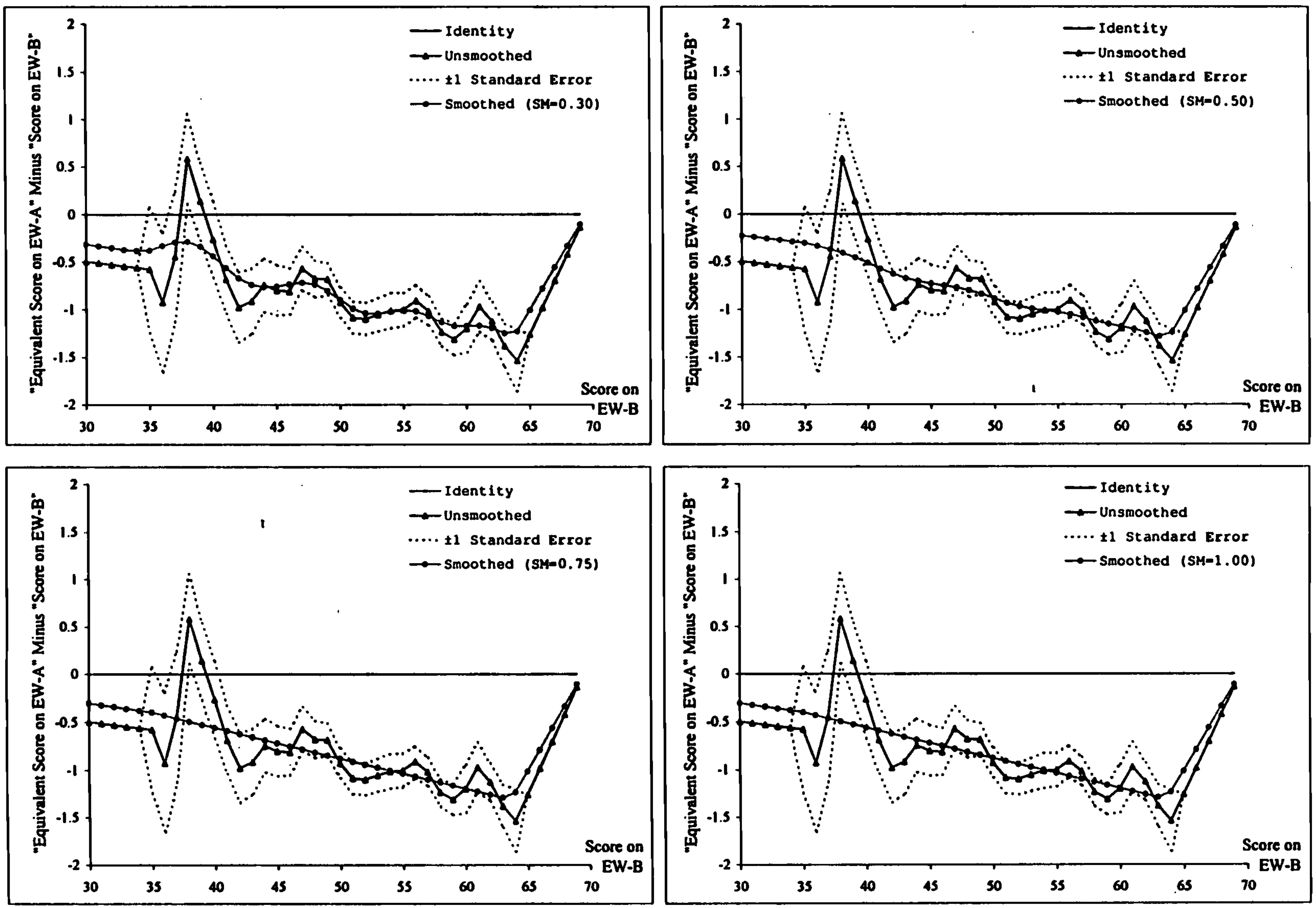


Figure 3 (cont'd)

group design (M. J. Kolen, personal communication, May 6,1997). This study therefore could not directly assess the "moment preservation" on one form for the particular population taking the other form because of missing data. In addition, the moments in this study depended on the particular assumption made by the frequency estimation method used in this study. The frequency estimation method assumes that, for both forms of a test, the conditional distribution of total score given each common-item score is the same in both populations.

Tucker Linear Equating

For each of the four subtests, Tucker linear method found an equating equation that transformed scores on one form to a set of new scores comparable to the scores on the other form. The Tucker equating equation was derived by defining a synthetic population, assuming equal conditional variances and same linear regression functions for the two populations, and estimating the means and variances for the synthetic population (Kolen & Brennan, 1987; Kolen & Brennan, 1995). Using the resulting Tucker equations, equivalent scores were established for the two forms of each sampled test.

Similarities among Outcomes of Various Equating Methods

Overall, this study found positive and significantly strong relationships among equating outcomes of various methods. Nonetheless, the weaker relationship between the IRT-based equating outcomes and the non-IRT equating outcomes reflected fundamental differences between the two methodologies. The dependent-samples *t*-test found small but statistically significant differences between the outcomes of the two IRT-based methods. However, the small differences might not have practical significance due to the small effect size and the large sample size. In general, individual examinees were ordered in a similar way regardless of the equating method used. Table 3 summarizes the relationship among various equating outcomes for various subtests.

The nearly perfect positive relationship between the outcomes produced by the Tucker method and the frequency-estimation equipercentile method, shown in Table 3, suggested that the two methods almost rank ordered individual examinees in the same way. For each of the four subtests, the Pearson *r* was close to 1 ($r > 0.999$). Correlational graphs further confirmed the similarities between the outcomes of the two non-IRT equating methods. The relationship between the IRT-based equating outcomes and the non-IRT equating outcomes were not as strong as the relationship between the outcomes of the two IRT-based methods. They were also not as strong as the relationship between the outcomes of the two non-IRT methods. The Pearson *r* between the IRT-based and the non-IRT equating outcomes ranged from 0.944 to 0.973 (see the underlined correlation coefficients presented in Table 3). The above findings reflect the logical and fundamental differences between the IRT-based equating approach and the classical equating approach.

Evaluation of Equating Accuracy

The accuracy of equating outcomes was evaluated using the four criteria described in the section for design and methodologies. An index of equating accuracy was obtained by computing the correlation coefficient for the relationship between the equivalent scores of an equating method and the criterion scores based on each of the four criteria for evaluating equating accuracy. Table 4 summarizes the evaluation results of equating accuracy for various methods and various subtests. Note that Table 4 only included the estimation outcomes of equating accuracy resulted from three evaluation criteria. For the evaluation criterion based on the equipercentile equating for individual subtests, its evaluation outcomes were presented in Table 3 (see the bordered numerical entries of the Pearson

Table 3 - Relationship among Various Equating Outcomes for Various subtests

Pearson Correlation Coefficient (*r*)		Equating Method															
		Tucker Linear method				Equipercentile Method				IRT-Based Linear Transformation Method				IRT-Based Fixed-b Method			
		SR	EW	PW	PS	SR	EW	PW	PS	SR	EW	PW	PS	SR	EW	PW	PS
Tucker Linear Method	SR	1.000															
	EW	0.782	1.000														
	PW	0.795	0.782	1.000													
	PS	0.810	0.755	0.795	1.000												
Equipercentile Method	SR	1.000	0.783	0.795	0.810	1.000											
	EW	0.782	1.000	0.782	0.754	0.783	1.000										
	PW	0.795	0.781	0.999	0.794	0.795	0.781	1.000									
	PS	0.810	0.755	0.795	1.000	0.811	0.754	0.794	1.000								
IRT-Based Linear Transformation Method	SR	0.965	0.780	0.791	0.834	0.964	0.780	0.790	0.834	1.000							
	EW	0.786	0.964	0.787	0.776	0.786	0.964	0.786	0.776	0.811	1.000						
	PW	0.768	0.750	0.944	0.773	0.767	0.750	0.944	0.774	0.785	0.776	1.000					
	PS	0.795	0.739	0.771	0.973	0.794	0.739	0.770	0.973	0.842	0.771	0.748	1.000				
IRT-Based Fixed-b Method	SR	0.963	0.779	0.790	0.831	0.962	0.779	0.789	0.831	0.999	0.810	0.791	0.837	1.000			
	EW	0.783	0.961	0.785	0.770	0.784	0.960	0.784	0.770	0.808	0.997	0.789	0.759	0.810	1.000		
	PW	0.785	0.771	0.963	0.801	0.784	0.771	0.963	0.801	0.804	0.798	0.976	0.793	0.802	0.795	1.000	
	PS	0.799	0.740	0.774	0.972	0.799	0.739	0.773	0.972	0.846	0.772	0.782	0.989	0.846	0.771	0.795	1.000

Note. All of the Pearson correlation coefficients were significant at α=.01.

SR=Simple Random Sampling; EW=Equal-Weight Domain Random Sampling; PW=Proportional-Weight Domain Random Sampling; PS=Purposeful Sampling.

Table 4 - Accuracy of Equating Outcomes of Various Equating Methods based on Different Evaluation Criteria

		Criterion for Evaluating Equating Accuracy		
		Based on Pseudo True Score		Arbitrary
Equating Method	Subtest	Total Raw Score on the 145 Anchor Items	IRT-estimated True Score based on the 145 Anchor Items	Frequency-Estimation Equipercentile Equating for the Original Test Forms
Tucker Linear Equating	SR	0.832	0.819	0.884
	EW	0.859	0.829	0.880
	PW	0.860	0.839	0.883
	PS	0.892	0.898	0.903
Frequency-Estimation Equipercentile Equating	SR	0.832	0.819	0.884
	EW	0.858	0.829	0.880
	PW	0.859	0.838	0.882
	PS	0.892	0.898	0.903
IRT-Based Linear Transformation Method	SR	0.856	0.860	0.897
	EW	0.877	0.870	0.893
	PW	0.845	0.839	0.864
	PS	0.894	0.917	0.897
IRT-Based Fixed-b Method	SR	0.854	0.858	0.896
	EW	0.873	0.865	0.889
	PW	0.870	0.867	0.888
	PS	0.895	0.916	0.898

SR=Simple Random Sampling; EW=Equal-Weight Domain Random Sampling; PW=Proportional-Weight Domain Random Sampling; PS=Purposeful Sampling.

Note. All of the indices of equating accuracy, represented by the Pearson *r* between the criterion scores and the resulting equivalent scores of an equating method, were statistically significant at α=.01.

correlation coefficients in Table 3).

Evaluation Using the Pseudo-True-Score-Based Criteria

Using the two evaluation criteria based on the pseudo true score, this study found that the IRT-based equating outcomes might be more accurate than the outcomes of the non-IRT methods. The difference in equating accuracy between the IRT-based method and the non-IRT-based method were small. However, statistical significance tests for these differences suggested that these differences were statistically significant at α=0.05 level. Nonetheless, the differences of small size might not have practical significance. The advantage of the IRT-based equating method in enhancing equating accuracy therefore remained unclear. While such small improvement may not have practical significance for equating in some occasions, it can be valuable for some other occasions that demand for a higher equating precision. The information about the degree of equating accuracy will also help to decide which equating method to use for a particular equating context.

For each of the four subtests, the accuracy of equating outcomes varied slightly from method to method. However, all of the four equating methods produced the least accurate outcomes for the subtest created by the simple random item-sampling scheme. Regardless of the equating method used, equating outcomes often were most accurate for the subtest created by the purposeful sampling scheme. In addition, equating outcomes for the other two subtests were estimated to be equally accurate. The important implications of such findings are:

- For common-item equating practice, it is important to include anchor items that are representative of the total test in content to improve equating accuracy.
- To improve equating accuracy, it is useful to construct test forms with more content homogeneous items, or to limit the content coverage of test forms to a small number of topics.

Using the two evaluation criteria based on the pseudo true score, the Tucker method and the frequency-estimation equipercentile method almost produced the same degree of equating accuracy. Overall, no method-test interaction effect was found for equating accuracy. That is, the relative equating accuracy of different equating methods were not dependent on the particular test forms being equated, and the equating accuracy for different sampled tests were independent of the particular equating method used. In addition, the two evaluation criteria based on the pseudo true scores led to vary similar outcomes about the equating accuracy of various methods and subtests.

Applicability of the Arbitrary Criteria

Evaluation results based on the two arbitrary criteria of evaluating equating accuracy confirmed that the use of an arbitrary criterion would lead to erroneous assessment outcomes of equating accuracy, as suggested in the literature.

Compared to the two pseudo-true-score-based criteria, the arbitrary criterion based on equating the original test forms using the equipercentile equating method often resulted in inflated indices for equating accuracy. It was in part due to the artifact of auto-correlation, which was more serious than ever because each of the subtests was a subset of the criterion test. In addition, the outcomes of the IRT-based equating methods were not always estimated to be more accurate than those outcomes from the non-IRT methods. Even when the IRT-based equating outcomes appeared to be more accurate, the improvement over the other methods often were not as good or significant as before (when the other evaluation criteria were used). Ideally, the criterion based on the equating outcomes of a longer test should be more reliable and be more effective for studying equating accuracy. However, the advantage was smeared by the criterion's inherent limitation that resulted in inflated estimate of equating accuracy, as well as its arbitrary nature. It reiterated the importance of the accuracy of an arbitrary criterion itself, when the criterion is used to evaluate the effectiveness of the other equating outcomes.

It is a common practice to use some arbitrary criterion for evaluating equating accuracy. However, the estimation of equating accuracy based on an arbitrary criterion is often biased because of the subjectivity of the particular criterion used. It is one of the particular interests of this study to investigate the potential bias due to the arbitrary nature of a criterion for evaluating equating accuracy. The arbitrary criterion used to study the estimation bias was the outcome of the equipercentile equating for each of the various subtests. In addition to the index of equating accuracy measured by the Pearson *r*, the root-mean-squared deviation (RMSD) statistic was also used as a second measure for equating accuracy. The RMSD is appropriate for the estimation of equating accuracy because the resulting equivalent scores being evaluated and the criterion scores were on the same scale.

The estimation outcomes for equating accuracy based on the above criterion differed from the outcomes resulted from the other criteria for evaluating equating accuracy. While the outcomes from other criteria suggested moderate equating accuracy for various methods on equating pairs of subtest forms, this criterion produced much higher degree of equating accuracy for various methods. It suggested potential bias due to the use of an arbitrary criterion. A dramatically different finding from this criterion is that the outcome of the Tucker linear method was found to be significantly more accurate than the outcomes of the IRT-based methods. Specifically, the indices of equating accuracy for the Tucker method over various subtests were all close to 1, suggesting nearly perfect accuracy. This finding clearly indicated that the arbitrary criterion was biased in evaluating equating accuracy. It overestimated the accuracy of the Tucker equating. Relatively speaking, this non-IRT-based criterion might underestimate the outcomes of the IRT-based equating.

Advantage of Multiple Criteria and Indices for Equating Accuracy

The use of multiple criteria for evaluating equating accuracy in this study proves to be very informative. Comparisons between the resulting evaluation outcomes of various criteria rendered an opportunity for thoroughly studying the effectiveness of various equating methods and the effect of content homogeneity on equating accuracy. The use of multiple criteria guarded the estimation of equating accuracy from being biased by a single arbitrary criterion.

Overall, the RMSD statistics agreed with the Pearson *r* statistics in estimating equating accuracy. However, there were still small discrepancies between the two measurement indices. Comparing the estimation outcomes resulted from these two statistics, it is clear that different statistics used to represent equating accuracy may lead to somewhat different estimations for equating accuracy. Therefore, it is important to select an index of equating accuracy that has well known features and advantages. The nature of the statistics used to represent the degree of equating accuracy should also be taken into account when interpreting the evaluation results for equating accuracy.

Summary for Evaluating Equating Accuracy

In estimating equating accuracy, it was found that all of the four equating methods yielded similarly accurate results to a moderate degree. The similarities could be attributed to the fact that there were more anchor items than non-anchor items in the original test forms, and hence all the sampled test forms had long anchor sets. Another plausible explanation was that the pairs of test forms being equated were already very similar before being equated. However, using the two pseudo-true-score-based evaluation criteria, IRT-based equating outcomes were found to be slightly more accurate than the outcomes of Tucker linear equating and frequency-estimation equipercentile equating. The small differences were statistically significant at $\alpha=.05$, but the small improvement might not be of practical significance. Overall, the use of the 3-PL model for IRT-based equating in equating minimum competency test forms was found to be satisfactory.

All of the methods yielded more accurate equating outcomes when anchor items were more content representative, or when the test forms were content homogeneous. Therefore, to improve equating outcomes, it is recommended to include more content representative anchor items or to have the content coverage of a test concentrate on fewer topics. If the degree of equating accuracy becomes critical when equating test forms with negatively skewed score distributions, IRT-based equating may be a better choice. Otherwise, either IRT-based equating or classical equating will yield adequate results, given the anchor items are content representative. The choice of a particular equating method should take into account factors such as cost and effectiveness, computation resources, and policy requirement of an equating program regarding an acceptable accuracy level. Whenever feasible, various equating methods should be applied to equate tests for high-stake examinations.

It was also found that the evaluation criterion based on the IRT-estimated pseudo-true-score was not biased in overestimating the accuracy of IRT-based equating outcomes. Generally, the two pseudo-true-score-based criteria yielded consistent evaluation outcomes. Estimation biases were found in using the arbitrary evaluation criterion. It overestimated the accuracy of the Tucker equating outcomes. As suggested in previous literature, it was concluded that the use of an arbitrary criterion would lead to erroneous outcomes for assessing equating accuracy. Overall, the use of multiple criteria for evaluating equating accuracy was informative. It compensated for the drawback of lacking an absolute good criterion.

Suggestions

Based on the results and discussions of this study, the following suggestions are recommended for improving the common-item equating practice and research in the future:

- When test forms being equated have negatively skewed score distributions and a higher degree of equating accuracy is needed, the IRT-based equating is preferred over the classical equating. However, both approaches should be considered in the context of the particular testing program.
- For constructing alternate test forms, it is desirable to include anchor items that are more representative of the total test.
- Given more flexibility in data manipulation, research in the future should seek for better control over test length and anchor length to gain a clear perspective for the effect of content homogeneity and representation on equating accuracy.
- In future research, the design of this study can also replicated with fewer common anchor items to investigate whether various equating methods still result in similar equating outcomes to a profound degree.
- Future studies can treat some common items in this study as unique items and equate the pairs of subtests using various equating methods to learn whether a small set of anchor items is as efficient and effective as a large anchor set in pursuing equating accuracy. The minimum number of anchor items required for equating test forms with negatively skewed score distributions can therefore be estimated, with consideration of efficiency and effectiveness.

- To help reduce the risk of introducing bias due to the use of one single criterion for evaluating equating accuracy, the applicability of using multiple criteria should be considered for future equating study.
- The efficiency and effectiveness of various indices used to represent equating accuracy need to be explored.
- Given more information on the characteristics of examinee groups, issues regarding the construct validity of a test should be investigated.
- Issues of test dimensionality remain largely unexplored because of the complexity of the content specificity for the test analyzed in this study. In the future, it may be useful to have a narrow focus on the test content such that dimensionality issues can be better investigated .
- It may be of interest to study the possibility of using the strong true score model that treats test items as a random sample to equate alternate test forms.
- The applicability of the study design to the context of performance assessment can be investigated.
- If a cutoff score in a particular score range is desired, future studies can be designed to establish a reasonable criterion for evaluating equating accuracy that will result in more accurate (that is, correctly pass or fail the test takers) and reliable estimation around the cut-off point.

Conclusions

This study investigated whether equating accuracy improves with an anchor test that is more representative of its corresponding total test, and whether such content effect depends on the particular equating method used. The major goal is to improve test results, on which critical educational decisions are based. The analysis results are expected to inform the practice of educational testing in many ways, leading to better selection of anchor items and equating methods, improved evaluation of equating accuracy, and a more precise, efficient, and fairer testing practice.

In this study, item-sampling schemes were introduced to manipulate the content homogeneity and representation of anchor items such that their effects on equating accuracy could be studied with various equating methods. Multiple criteria were proposed to evaluate equating accuracy, which should help overcome the common drawbacks of inadequate evaluation resulting from the use of one single arbitrary criterion. Overall, this study found all the equating methods employed had accurate results to a moderate degree. They all produced more accurate results when the anchor items were more representative of their total test, or when the content coverage of a test concentrated on fewer topics. Therefore, it is recommended to include anchor items that fully reflect the overall content coverage of a test to improve equating accuracy. In general, IRT-based equating was found to be more accurate than the classical equating. However, the small differences may not be substantial. As a rule of thumb, IRT-based equating is generally preferred if there is a need of a higher degree of equating accuracy. Otherwise, either IRT-based equating or classical equating can be applied in practice, depending on an equating program's policy regarding the acceptable level of accuracy, budget plan, and resources availability. The equating approach that better fits these requirements or considerations will result in a more efficient yet effective outcome. Nevertheless, for equating test forms for high-stake examinations, various equating methods and multiple criteria for evaluating equating accuracy should be considered if the time and cost were not an issue. The purpose is to produce equating outcomes that are precise and less prone to bias.

References

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 (pp. 9-49). New York: Academic.

Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. Journal of Educational Measurement, 22, 13-20.

Cook, L. L., & Eignor, D. R. (1991). An NCME instructional module on IRT equating methods. Educational Measurement: Issues and Practice, 10, 37- 45.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Chicago: Holt, Rinehart and Winston, Inc.

Green, D. R., Yen, W. M., & Burket, G. R. (1989). Experiences in the application of item response theory in test construction. Applied Measurement in Education, 2, 297-312.

Hambleton, R. K., & Swaminathan, H. (1990). Item response theory: Principles and applications. Boston: Kluwer-Nijhoff Publishing.

Hanson, B. A., Zeng, L., & Kolen, M. J. (1995, October). Equating Computer Programs. (Available from Michael Kolen, ACT, 2255 N. Dubuque Street, Iowa City, IA 52243).

Harris, D. J. (1991). Equating with non-representative common item sets and non-equivalent groups. Paper presented at the American Educational Research Association, Chicago.

Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. Applied Measurement in Education, 6, 195-240.

Hills, J. R., Subhiyah, R. G., & Hirsch, T. M. (1988). Equating minimum-competency tests: Comparisons of methods. Journal of Educational Measurement, 25, 221-231.

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. Journal of Educational Measurement, 22, 197-206.

Kolen, M. J. (1991). Smoothing methods for estimating test score distributions. Journal of Educational Measurement, 28, 257-282.

Kolen, M. J., & Brennan, R. L. (1987). Linear equating models for the common-item nonequivalent-populations design. Applied Psychological Measurement, 11, 263-277.

Kolen, M. J., & Brennan, R. L. (1995). Test equating: Methods and practices. New York: Springer-Verlag.

Kolen, M. J., & Harris, D. J. (1990). Comparison of item preequating and random groups equating using IRT and equipercentile methods. Journal of Educational Measurement, 27, 27-39.

Kolen, M. J., & Jarjoura, D. (1987). Analytical smoothing for equipercentile equating under the common item nonequivalent populations design. Psychometrika, 52, 43-59.

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating Methods works best? Applied Measurement in Education, 3, 73-95.

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.

Petersen, N. C., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating models. In P. W. Holland and D. B. Rubin (Eds.), Test Equating. New York: Academic Press, 71-135.

Schmitt, A. P., Cook, L. L., Dorans, N. J., & Eignor, D. R. (1990). Sensitivity of equating results to different sampling strategies. Applied Measurement in Education, 3, 53-71.

Yang, W. L. (1997, March). The effects of content mix and equating method on the accuracy of test equating using anchor-item design: comparisons of linear and IRT equating. Paper presented at the annual meeting of the American Educational Research Association, Chicago. (ERIC Document Reproduction Service No. ED 409 334)

Yen, W. M. (1985). Tau equivalence of vertical equating using three-parameter item response theory and Thurstonian procedures. Paper presented at the meeting of the American Educational Research Association, Chicago.

AERA

ERIC®

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

TM030877

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: The effects of content homogeneity and equating method on the accuracy of common-item test equating.

Author(s): Wen-Ling Yang

Corporate Source: Educational Testing Service

Publication Date: 4/27/2000

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ↑ ☒	Level 2A ↑ □	Level 2B ↑ □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature: Wen-Ling Yang	Printed Name/Position/Title: Wen-Ling Yang Measurement Statistician	
Organization/Address: Educational Testing Service Rosedale Road, MS 11-L Princeton, NJ 08541	Telephone: (609) 683-2088	FAX: (609) 683-2130
	E-Mail Address: wyang@ets.org	Date: 4/25/2000

(over)

ERIC® Clearinghouse on Assessment and Evaluation

University of Maryland
1129 Shriver Laboratory
College Park, MD 20742-5701

Tel: (800) 464-3742
(301) 405-7449
FAX: (301) 405-8134
ericae@ericae.net
http://ericae.net

March 2000

Dear AERA Presenter,

Congratulations on being a presenter at AERA. The ERIC Clearinghouse on Assessment and Evaluation would like you to contribute to ERIC by providing us with a written copy of your presentation. Submitting your paper to ERIC ensures a wider audience by making it available to members of the education community who could not attend your session or this year's conference.

Abstracts of papers accepted by ERIC appear in *Resources in Education (RIE)* and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of *RIE*. Abstracts of your contribution will be accessible through the printed, electronic, and internet versions of *RIE*. The paper will be available **full-text, on demand through the ERIC Document Reproduction Service** and through the microfiche collections housed at libraries around the world.

We are gathering all the papers from the AERA Conference. We will route your paper to the appropriate clearinghouse and you will be notified if your paper meets ERIC's criteria. Documents are reviewed for contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality. You can track our processing of your paper at **http://ericae.net.**

To disseminate your work through ERIC, you need to sign the reproduction release form on the back of this letter and include it with **two** copies of your paper. You can drop of the copies of your paper and reproduction release form at the ERIC booth (223) or mail to our attention at the address below. **If you have not submitted your 1999 Conference paper please send today or drop it off at the booth with a Reproduction Release Form.** Please feel free to copy the form for future or additional submissions.

Mail to: AERA 2000/ERIC Acquisitions
The University of Maryland
1129 Shriver Lab
College Park, MD 20742

Sincerely,

Lawrence M. Rudner

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE

ERIC/AE is a project of the Department of Measurement, Statistics and Evaluation at the College of Education, University of Maryland.