Effects of Equating Strains on Equating Error in the Common-Item Nonequivalent Groups Design

Jae-Chun Ban · Bradley A. Hanson · Deborah J. Harris
(KICE · CTB/McGraw-Hill · ACT, Inc.)

I. Introduction

Many large scale educational testing programs make use of alternate forms of the same test, which are then equated to establish interchangeability among forms. Alternate forms are constructed with different items to be as similar as possible in their content and statistical specifications. These testing programs usually have several distinct annual test administration dates (e.g., in March, June, and October) with different test forms given each date.

Examinees who test at different times of the year may differ in their overall level of achievement due to such things as recency or amount of relevant coursework. If such group differences exist, it is recommended that a new form be equated to a form from the same time of year rather than to a form from a different time of year to obtain greater equating stability and to keep the scores more interchangeable by maximizing the similarity of groups used in the equatings (Kolen & Brennan, 1995). As an illustrative equating plan, if a test is administered in March, June, and October, and the new March forms are always equated to previous March forms, June forms are equated to previous June forms, and October forms are equated to previous October forms.

The similarity of groups is an important factor in obtaining good equating, particularly when the common-item nonequivalent groups design is employed to equate forms from different test dates (Harris & Kolen, 1986; Harris, 1987; Cook & Petersen, 1987). In the common-item nonequivalent groups design two groups of examinees from different populations are each administered different test forms that have a subset of items in common.

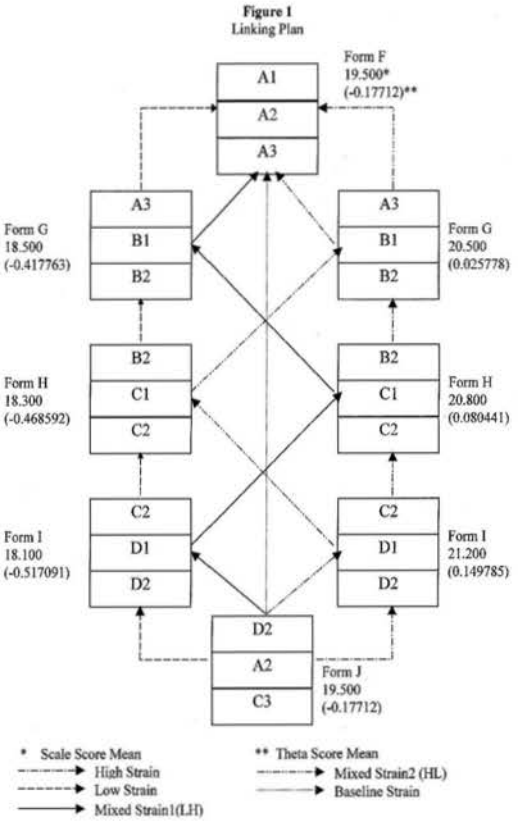
Although the equating a new form to an old form from the same time of year is ideal from the perspective of using similar groups, it has a significant problem: development of equating strain. Suppose a test is administered on different test dates and statistical characteristics (mean and standard deviation) of the sample administered each form are different on each test date because of different overall levels of achievement. Different linking plans may be made for each time of year. If this linkage pattern were extended over several years, it would be called an equating strain, and the scores from the forms in different strains would be no longer comparable.

Equating strains can lead to systematic error in which examinees earn higher scale scores on one form than on another form (Kolen & Brennan, 1995). Equating strains can also lead to random error. Random equating error is present whenever samples from populations of examinees are used to estimate equating relationships. Systematic equating error results from violations of the assumptions and conditions of the particular equating methodology used. Smoothing techniques, violations of statistical assumptions, improperly implemented data collection designs, and the substantial differences between the equating group and the operational group are some examples of systematic error sources (Kolen & Brennan, 1995). When the common item nonequivalent groups design is used for equating, the group differences across administrations, particularly if substantial, also affect the amount of systematic equating error. Both random and systematic errors tend to accumulate over time, especially if a large number of test forms are involved in the equating process.

Several ways to avoid strains for the common-item nonequivalent groups design were suggested by Kolen and Brennan (1995): minimizing the number of links that affect the comparison of scores on adjacent forms, minimizing the number of links back to the initial form, and so on.

When a score scale has been in place over a number of years, however, strains are inevitable unless all new forms link back to a single old form. The effects of equating strains on equating errors has not been extensively examined.

The purpose of this study was, first, to investigate the extent to which random and systematic errors of equated scores in the common-item nonequivalent groups design arise when several equating linkage patterns are constructed with only higher groups mean forms, with only lower groups mean forms, or with a mixture of lower and higher groups mean forms (see Figure 1 where different equating linkage paths are illustrated). Second, this study investigated which equating methods frequently used for the common-item nonequivalent groups design are the least affected by the equating linkage paths that are used.



Ⅱ. Methodology

1. Constructing Forms

This study used four 60 item ACT Mathematics forms (ACT, Inc., 1997) denoted here as Forms A, B, C, and D. Randomly equivalent groups took Forms A (2,696 examinees), B (2,670 examinees), C (2,617 examinees), and D (2,651 examinees). Forms A, B, C, and D do not have any items in common.

Five 60 item Mathematics forms that contained common items were constructed from items on the original four forms. Items on each of the original four forms were divided into three sets of 20 items, such that the content and statistical characteristics of the three sets were as similar as possible. For Form A the three sets of 20 items are denoted A1, A2, and A3. Item sets on the other forms are similarly named. The item sets were combined to construct five forms denoted F, G, H, I, and J. Form F contained item sets A1, A2, and A3, and so was identical to Form A. Form G contained item sets A3, B1, and B2, so that Forms F and G had 20 items in common. The forms were constructed so that pairs of consecutive forms had 20 items in common (e.g., Forms F and G had 20 items in common, Forms G and H had 20 items in common, etc.). This formed a common item equating chain of 5 forms (J > I > H > G > F), where J is the newest form to be equated to the base Form F through Forms I, H, and G. In addition, Form F and J had 20 items in common to allow Form J to be directly equated to Form F. The item sets used in Forms F, G, H, I, and J, and the sequence of forms in the equating chain are given in Figure 1.

2. Population Differences and Constructing Equating Strains

Item responses for the forms in the equating chain were generated from high, middle, and low population distributions of ability (IRT theta distributions). The population distributions were normal, with standard deviation one and with means varying to represent low, middle, and high ability populations. To choose the means, ACT Assessment scale score means observed on national test dates for the past several years were reviewed. The three lowest scale score means were considered as means for the low populations, while the three highest scale score means were considered as means for the high populations, and a middle scale score mean was used as the mean for the middle population. Scale score means were converted to means on the IRT thetas scale. Data for the new form (Form J) and the base form (Form F) were generated to have a theta mean of 0.177. Two sets of data were generated for the three middle forms in the chain of equatings: one from low populations (with means of 0.42, 0.47, and 0.52 for Forms G, H, and I), and one from high populations (with means of 0.03, 0.08, and 0.13 for Forms G, H, and I). In Figure 1, the ACT Assessment scale score on Form A and the corresponding theta mean (in parenthesis) for each form are shown.

Forms J and F were administered to the simulees of the middle ability population. Forms G, H, and I were administered either to a low ability population simulees or a high

ability population simulees (high or low strains). This study considered five paths linking the forms using the low, high, mixed1(LII), mixed2(IIL) and baseline strains as represented in Figure 1. Equating using the low strain is represented by the long dashed lines (through Forms G, H, and I on the left side of Figure 1). Equating using the high strain is represented by the mixed long and one short dashed lines (through Forms G, H, and I on the right side of Figure 1). Equating using the mixed1(LII) strain is represented by the solid line alternating between the two strains. The line containing a long and two short dashes that alternates between the two strains represents the mixed2(IIL) strain. The dotted line shows a direct equating, called baseline, from Form J to Form F, that will be compared to the equatings based on the longer chains of forms.

3. Data Generation Procedures

For this simulation study, dichotomous item response data were generated for simulees. Item parameter estimates for each item on each form were obtained from BILOG (Mislevy & Bock, 1990) assuming a three parameter logistic IRT model using equating data from actual administration of the ACT. The estimates were treated as if they were parameter values in the generation of simulee responses. The specific procedures for data generation were:

1. Generate random ability parameters for 3,000 simulees from the normal population distribution with a standard deviation of one and means varying to represent low, middle, and high ability populations (e.g., mean of 0.42 for Form G for low population),
2. Compute the probability of a simulee correctly answering an item,
3. Generate a random number from the uniform distribution for each item,
4. Compare the generated number from step 3 to the probability from step 2. Assign a one (correct) if the random number is less than or equal to the probability; otherwise assign a
   zero (incorrect),
5. Repeat steps 2 through 4 for each item,
6. Repeat steps 2 through 5 for each simulee,
7. Repeat steps 1 to 6 for each form.

4. Equating Methods

The equating methods under consideration in this study are linear and curvilinear methods appropriate for the common item nonequivalent groups design. Curvilinear methods encompass both the frequency estimation equipercentile equating (the frequency estimation, Angoff, 1984) and IRT methods.

The common item nonequivalent groups design involves two populations. Two populations must be combined to obtain a single population, synthetic population, for defining an equating relationship. The synthetic group is a weighted combination of two nonequivalent groups (Braun & Holland, 1982). In this study, the weight of the synthetic group was one for new group and zero for old group for linear methods, the frequency estimation method, and the IRT observed score method .

In linear equating, the means and standard deviations on the two forms for a particular

synthetic population are set equal. The linear conversion equating, $l_{Y_S}(x)$, for observed scores in synthetic population S is defined in the following way:

$$l_{Y_S}(x) = \frac{\sigma_S(Y)}{\sigma_S(X)}[x - \mu_S(X)] + \mu_S(Y), \quad (1)$$

where x and y refer to the test scores to be equated, S(X) and S(Y) refer to the standard deviations of Form X and Form Y scores in synthetic population, and S(X) and S(Y) refer to the means of the two forms in synthetic population.

Three linear equating methods were considered in this study: the Tucker method, the Levine observed score method, and the Levine true score method (Kolen & Brennan, 1995; Petersen, Kolen, & Hoover, 1989). In the common-item nonequivalent groups design, the common items are used to adjust for population differences in performance on the non-common items. This requires strong statistical assumptions because each examinee comes from only one population and takes only one form. These linear equating methods differ in terms of their statistical assumptions (Kolen & Brennan, 1995). The Tucker and Levine observed score methods transform observed scores on Form X to observed scores on Form Y, while the Levine true score method is based on the estimated true scores of Form X and Form Y.

The frequency estimation method estimates the cumulative distributions of scores on Form X and Form Y for a synthetic population. In the frequency estimation method with the nonequivalent groups design, the Form X distribution is set equal to the Form Y distribution for a synthetic group of examinees. Form X scores converted using equipercentile equating methodology have approximately the same mean, standard deviation, and distributional shape (Kolen, 1988). The equipercentile equating function is found by locating scores on Form X that have the same percentile rank as scores on Form Y.

IRT equating methods involve estimating item parameters on the two forms, placing parameter estimates on the two forms on a common scale, and equating test scores. In the IRT true score method, the true scores on one form associated with a given is considered to be equivalent to the true score on another form associated with that. The IRT observed score method uses the IRT model to produce an estimated distribution of the observed number-correct score on each form, which then are equated using the equipercentile method (Kolen & Brennan, 1995, p. 181).

5. Equating Procedures

For 500 sets of simulated samples, six equatings were computed for each of the five strains (high, low, two mixed, and baseline strains) in Figure 1. The equating methods used were: (1) Tucker method, (2) Levine observed score method, (3) Levine true score method, (4) frequency estimation method, (5) IRT observed score method, and (6) IRT true score method. Form F was treated as the oldest form and Form J as the newest form. After equating a new form to an old form in the sequence of Figure 1 (J -> I -> H -> G -> F), the conversion table of raw-to-scale scores for each method was saved for other subsequent equatings.

For the IRT equatings, the computer

program BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) was used to estimate item parameters for each form. Since the item parameters estimated from the separate runs of the BILOG-MG were not on a common scale, the computer program ST (Hanson & Zeng, 1995a) was used to put the item parameter estimates on a common scale via the Stocking-Lord scale transformation procedure (Stocking & Lord, 1983). Then, the computer program PIE (Hanson & Zeng, 1995b) was employed to conduct IRT true and observed score equatings.

6. Criteria

The fitted observed score distributions based on the three parameter IRT model used to generate the data were calculated for Forms F and J. The population equating function was taken to be the equipercentile equating function computed using the fitted observed score distributions. The following specific procedures were used to obtain the population equating function:

1. Use BILOG to obtain item parameter estimates for Forms F and J using the three-parameter IRT model and then treat them as true values,
2. Obtain a discrete approximation of the normal distribution with mean .17712 and standard deviation 1.0 and use this as the population distribution for Forms F and J. This was done by using 100 equally spaced points for ranging from -4.0 to 4.0, computing the density at each point, and standardizing the distribution so that the sum of the probabilities across the 100 points was 1,
3. Compute the IRT observed equating function using the PIE program, and treat it as the population equating function.

As evaluation indices, mean squared errors (total error; MSE), systematic error (bias), variances (random error), the weighted average mean squared errors (WMSE), the weighted average systematic errors (WBIAS), and the weighted average random errors (WRANDOM) of the estimated equating functions were computed for each of the 30 combinations of equating methods and strains using the 500 replications. The total error is the sum of the random and systematic errors: MSE = random error + systematic error. The total error is equal to the average squared difference between the equated score and the population equated score at each raw score point over the 500 replications. The systematic error or bias is the squared difference between the mean of the equated scores over the 500 replications and the population equated score at each raw score point. The random error is equal to the variance of the 500 equated scores at each raw score point. The following formulas were used to compute the errors for the strain k [where k is the baseline, high, low, mixed1(LH), or mixed2(HL)].

$$MSE(x_i) = \frac{1}{500}\sum_{j=1}^{500}\{\hat{e}_{kj}(x_i) - b_{ase}(x_i)\}^2, \quad (2)$$

$$\text{Systematic Error (Bias)}(x_i) = [\frac{1}{500}\sum_{j=1}^{500}\hat{e}_{kj}(x_i) - b_{ase}(x_i)]^2, \quad (3)$$

$$\text{Random Error}(x_i) = \frac{1}{500}\sum_{j=1}^{500}[\hat{e}_{kj}(x_i) - \frac{1}{500}\sum_{j=1}^{500}\hat{e}_{kj}(x_i)]^2, \quad (4)$$

$$\text{WMSE} = \{\sum_{i=0}^{60}(\frac{f(x_i)}{N})MSE(x_i)\}, \quad (5)$$

$$WRANDOM = \{ \sum_{i=0}^{60} (\frac{f(x_i)}{N}) RandomError(x_i) \}, \quad (6)$$

$$WBIAS = \{ \sum_{i=0}^{60} (\frac{f(x_i)}{N}) SystematicError(x_i) \}, \quad (7)$$

where x_i is a raw score point(x_i=i for i=0,...,60),

j is a replication(j=1,...,500),

k is a strain(low,mixed1(LH),mixed2 (HL), and baseline),

$\hat{e}_{kj}$ is the estimated equating function for strain k on replication j,

$b_{ase}(x_i)$ is a population conversion at each score point,

$f(x_i)$ is the number of simuless at each score point that are expected based on the population distribution,

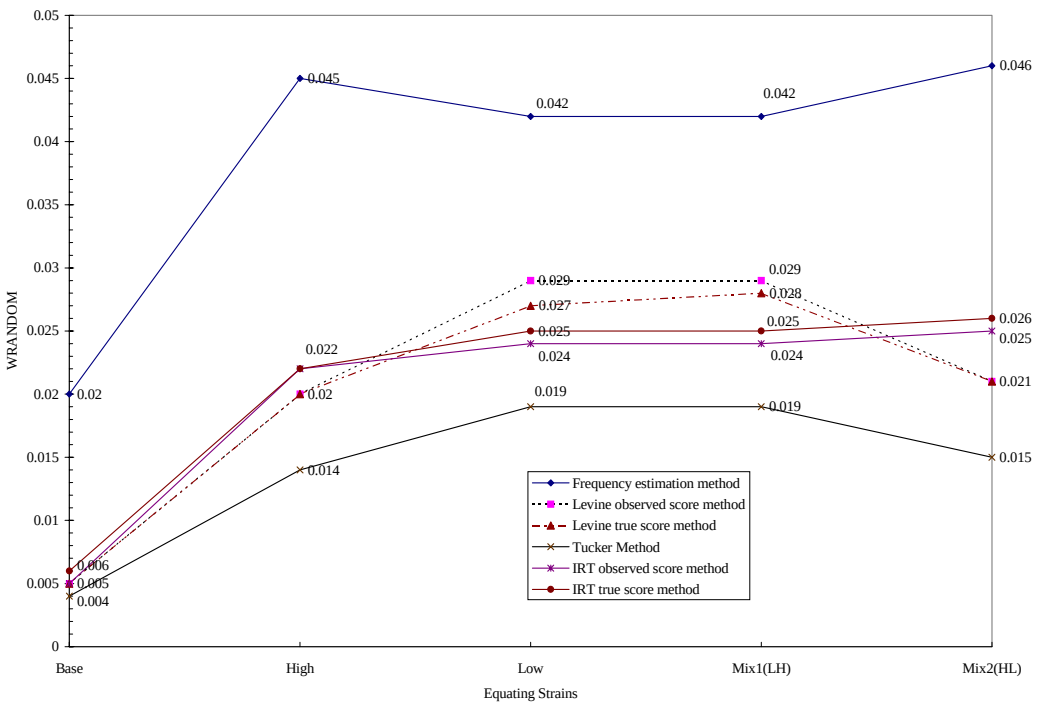
N is the number of simulees(=3,000).

Due to space restriction, only the summary indices WMSE, WBIAS, and WRANDOM are reported. The weighted average error indices are more useful than unweighted error indices because, instead of giving equal weight to all values of the raw score, they give the greatest weight to raw scores that are most likely to occur.

Ⅲ. Results

Figure 2 shows the WRANDOM for all strains using the six equating methods. It shows the frequency estimation method yielded larger WRANDOM for all strains than the other methods. This result is

Figure 2
Weighted Random Errors of Each Strain Using Six Equating Methods



consistent with the general fact that the frequency estimation method needs larger sample sizes than the linear methods (the Levine observed and true score methods and the Tucker method) to maintain the same level of random error (Kolen & Brennan, 1995). The Levine methods resulted in moderate amounts of WRANDOM. The Tucker method produced the least amount of WRANDOM across all strains. The IRT methods usually need relatively larger sample sizes than the conventional equating methods to obtain a similar level of random error. The IRT methods, however, produced a similar level of WRANDOM as the Levine methods and less WRANDOM than the frequency estimation method. The reason is that the IRT methods effectively smooth the data. For example, the IRT observed method is a form of smoothed equipercentile equating, so it should lead to smaller WRANDOM than the unsmoothed frequency estimation method.

For each method, the WRANDOM was similar across all strains except the baseline. It can be seen that different linkage plans did not substantially affect the WRANDOM within each equating method. All methods produced the lowest WRANDOM for the baseline. This makes sense because on the baseline there were no mediating strains and random error was only present for one equating link (Form J to Form F).

The WBIAS and the WMSE are presented in Figures 3 and 4. For most cases, since the WBIAS was the main contributor to the WMSE, Figures 3 and 4 are similar. The

Figure 3
Weighted Systematic Error of Each Strain Using Six Equating Methods

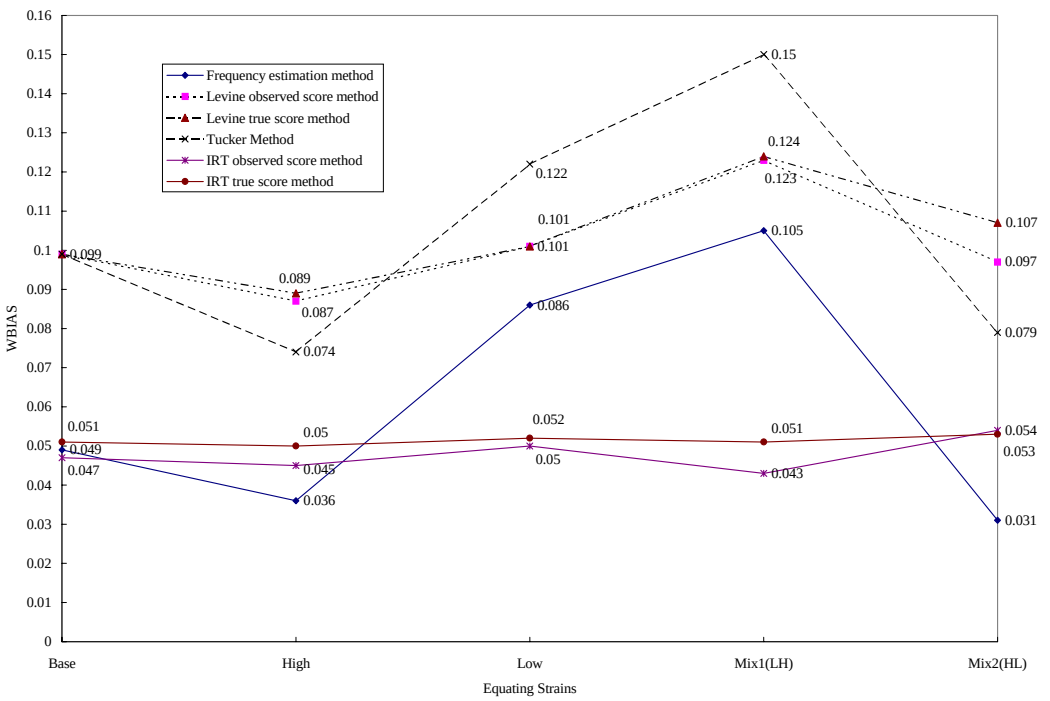
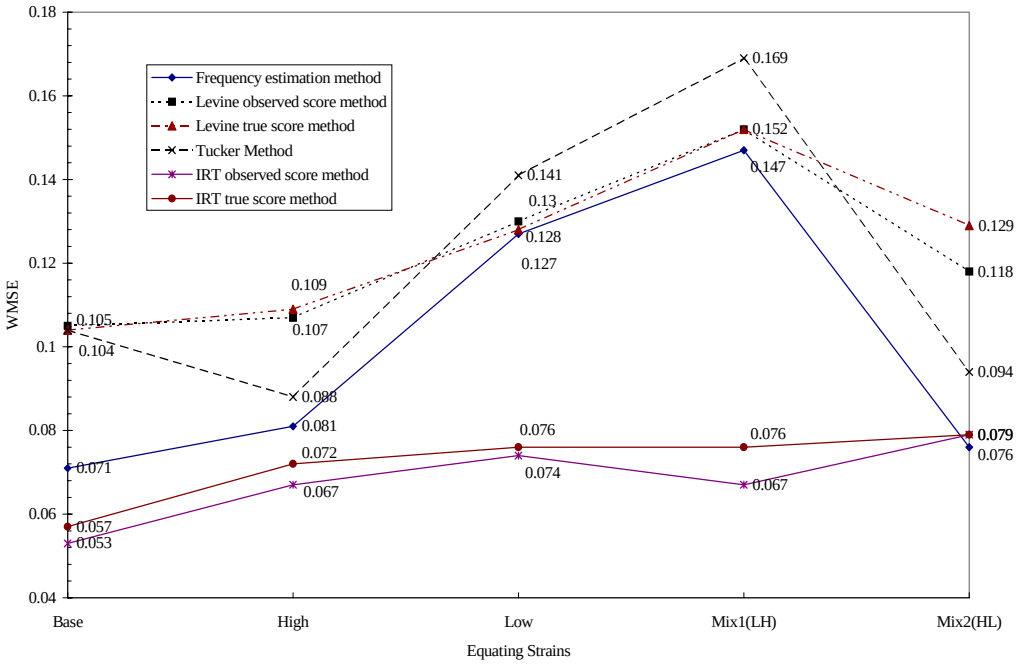


Figure 4
Weighted Mean Squared Errors of Each Strain Using Six Equating Methods



following description will focus on WMSE. Overall, the IRT methods produced the lowest WMSEs for most of strains, the frequency estimation method the next smallest, and the linear methods resulted in the largest WMSEs. The IRT methods also produced lower variability of the WMSEs across strains than the other methods.

Some reasons for better performance of IRT methods than the others may be (1) the simulation data were generated on the basis of IRT model and assumptions of the model were perfectly met by the data, (2) the population equating function, as the criterion, was equipercentile function based on the population number correct score distributions using the IRT model, which is nonlinear, and (3) the assumption that the equating function is linear for linear methods was violated and because the population equating function was nonlinear, the linear methods introduced more systematic errors. The IRT methods took advantage of the characteristics of the simulation data and more accurately decomposed group and form differences, which resulted in lower total errors than for the conventional methods. Since the population equating function was nonlinear, the frequency estimation method yielded smaller WMSEs across all strains than the linear methods.

Of the linear methods, for the high and mixed2(HL) strains, the Tucker method yielded lower WMSEs than the Levine observed and true score methods, while for the low and mixed1(LH) strains the Levine observed and true score methods resulted in lower WMSEs than the Tucker method. It is

not obvious why such results occurred.

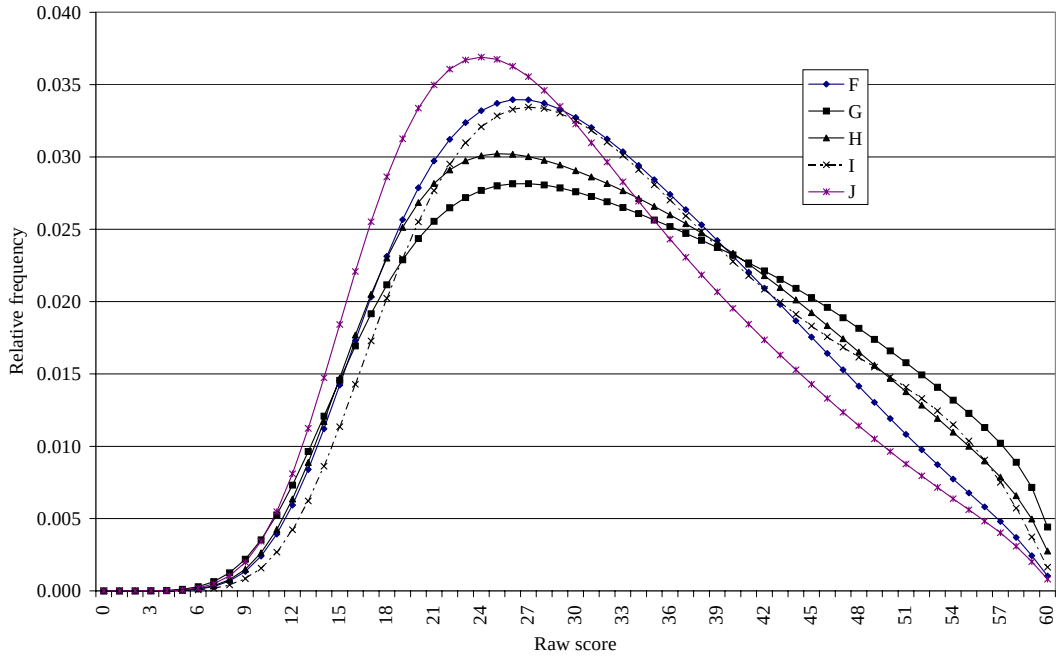
The Tucker method and the frequency estimation method yielded a similar pattern of WMSEs in the high, low, mixed1(LLI), and mixed2(IIL) strains, although the size of the errors was different. Both methods require assumptions about characteristics of the observed relationship between total score and score on the common items being the same for the two populations. These common assumptions may lead to the similar pattern of WMSEs for all strains. For both the Tucker method and the frequency estimation method, the high strain and the mixed2(IIL) strain that included two high means and one low mean had lower WMSEs than the low and the mixed1(LII) strain that included two low means and one high mean (Figure 4). The results indicated either more or less bias depending on whether the newest form, Form J, was equated to the high ability group (as in the high and mixed2(IIL) strains) administered Form I or to the low ability group (as in the low and mixed1(LII) strains) administered Form I.

The more data sets the strains have in common (e.g, the high and mixed2(IIL)), the more similar the WMSEs for the conventional methods are. However, this explanation for the differences in WMSE among strains could be applied only for the Tucker method and the frequency estimation method, not for the Levine methods and the IRT methods. For example, the Levine true score method produced a closer result for the low and the mixed2(IIL) strains than the result for the high and the mixed2(IIL) strains, which have more data in common.

Another tentative explanation for the differences in WMSE among strains involves the fact that since the assumptions for the Tucker methods (e.g., the conditional score variance of unique items given common item scores is the same for new and old form populations for the Tucker method) and for the frequency estimation method were not met, bias can be expected regardless of which equating strains were used. This bias may in part depend on whether the mean differences between groups and forms are in the same direction (the lower performing group takes the harder form and the higher performing group takes the easier form) or the opposite direction (the lower performing group takes the easier form and the higher performing group takes the harder form).

Figure 5 shows population raw score distributions for each form if the population administered Form J had been administered Form F, G, H, and I. Form J stands out as being somewhat more difficult than the other forms. Form J is involved only in the first link to Form I. In the high and mixed2(IIL) strains, because the easier Form I was administered to the better performing group (theta = 0.1498 in Figure 1) and the more difficult Form J was administered to the lower performing group (theta = -0.177) the mean differences between groups and forms were in the same direction. On the other hand, in the low and mixed1(LII) strains, the easier Form I was administered to the lower performing group (theta = -0.517) and the more difficult Form J was administered to the better performing group (theta = 0.177), so the mean differences between groups and forms were in opposite directions. Perhaps the pattern of the relative difference between the group means and form means accounts for some of the difference in bias of the

Figure 5
Population distributions for each form



Tucker method and the frequency estimation method for the low and mixed1(LH) strains versus the high and mixed2(HL) strains.

The Levine observed and true score methods produced a similar pattern of results for most strains. In addition, both methods produced lower variability of WMSEs across strains than the Tucker method and the frequency estimation method. In other words, these methods were less affected by different equating linkages than the frequency estimation method and the Tucker method, although the absolute sizes of WMSEs were larger than those of the frequency estimation method and were larger or smaller than those of the Tucker method. The Levine methods were less affected by group difference than the Tucker and the frequency methods.

The IRT methods decomposed the group difference and form difference more accurately which may be due to the data perfectly meeting the assumptions of the IRT model. This also could account for the IRT methods resulting in a similar level of total errors across all strains.

Ⅳ. Conclusion and Discussion

Equating using the common-item nonequivalent groups design is frequently used in large testing programs that administer several forms on different test dates. Although equating strains or scale drift should be avoided, they are inevitable when a score scale has been in place over a number of years. In this study, using the Tucker and the frequency estimation methods, the high

and mixed2(HL) strains resulted in lower weighted total errors than the low and mixed1(LH) strains. The Levine methods and the IRT methods also produced lower weighted total errors in the high strain. This does not imply that practitioners should use the high strains when planning linking in practice. The results may be specific to given data because, in part, the simulation data violated the assumptions for the conventional methods and equating itself is population dependent. When the IRT methods were used, the weighted total errors were similar across all strains. When comparing equating methods in terms of weighted total error, the IRT methods produced smallest errors across all strains, the frequency estimation method the next smallest errors, and the linear equating methods produced the largest errors.

However, it is not fair to directly compare the IRT methods to the conventional methods with the simulation data because (1) the simulation data were generated using an IRT model, so the assumptions for the IRT model were perfectly met by the data, (2) the population equating function, as the criterion, was equipercentile function using the population number correct score distribution based on an IRT model, which is nonlinear, and (3) the simulated data violated the assumptions for linear methods. If the data were multidimensional or the population equating function was linear, the patterns of bias for the equating methods would likely have changed.

The implication of this study is that using strains can affect equating error, and therefore a linking pattern should be carefully selected. Even with mixed strains, there can be substantial equating error when using a single equating link. The equating error resulting from using a single link may be reduced by using multiple links (Kolen & Brennan, 1995, p. 261; Hanson, Harris, & Kolen, 1997; Mckinley & Schaeffer, 1989). A topic for further research would be to investigate the performance of double linking with equating strains. Also, using simulation data generated from different models, including multidimensional models, increasing the group differences (in this study the difference between low and high groups was about 2 scale score points on the ACT assessment score scale), and increasing the number of links in each strain could be considered for future research.

References

ACT, Inc. (1997). *ACT assessment technical manual.* Iowa City IA: Author.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores.* Princeton NJ: ETS.

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

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement,* 11, 225 244.

Hanson, B. A., Harris, D. J., & Kolen, M. J. (1997). *A comparison of single and multiple linking in equipercentile equating with random groups.* Paper

presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Hanson, B. A., & Zeng, L. (1995a). *ST: A computer program for IRT scale transformation* [Computer program]. Iowa City IA: ACT, Inc.

Hanson, B. A., & Zeng, L. (1995b). *PIE: A computer program for IRT true and observed score equating* [Computer program]. Iowa City IA: ACT, Inc.

Harris, D. J., & Kolen, M. J. (1986). Effect of examinee group on equating relationships. *Applied Psychological Measurement,* 10, 35 43.

Harris, D. J. (1987). *Effect of comparability of examinee groups on equating* (Research Report 87 5). Iowa City IA: ACT, Inc.

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

Kolen, M. J. (1988). Traditional equating methodology. *Educational Measurement: Issues and Practice,* 7, 29 36.

Mckinley, R. L., & Schaeffer, G. A. (1989). *Reducing test form overlap of the GRE subject test in mathematics using IRT triple part equating* (Research Report 89 8). Princeton NJ: ETS.

Mislevy, R. J., & Bock, R. J. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models* (2nd ed.) [Computer program]. Mooresville IN: Scientific Software International.

Petersen, N. S, Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., 221 262). New York: American Council on Education/ Macmillan.

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

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D., (1996). *BILOG MG: Multiple group IRT analysis and test maintenance for binary items* [Computer program]. Chicago: Scientific Software International.

초록

공통문항 비동등집단 설계에서 동등화 연결선(equating strains)이 검사 동등화 오차에 미치는 영향

반재천 · Bradley A. Hanson · Deborah J. Harris
(한국교육과정평가원 · CTB/McGraw-Hill · ACT, Inc.)

검사동등화를 적용하는 대규모 검사프로그램의 경우 여러 가지 난이도를 가진 검사폼(test form)들이 고리처럼 연결되어 검사점수들이 상호 비교 가능하게 된다. 본 연구는 연결되는 모든 검사의 평균점수가 높거나 낮을 때, 혹은 높고 낮은 검사가 섞여 있을 때 생길 수 있는 오차(무선오차와 체계적 오차)의 정도를 탐구하였다. 이 연구에서는 공통문항 비동등집단설계법(common-item nonequivalent groups design)을 이용하였다. 시뮬레이션을 통해 어떤 검사동등화법이 여러 유형의 동등화 연결선 (equating strains)에 영향을 적게 받는지를 탐구했다. High strain은 평균점수가 높은 검사만으로 연결된 것을 말하고, Low strain은 평균점수가 낮은 검사만으로 연결된 것이며, Mixed1(LH)는 평균점수가 낮고 높은 검사들이 교대로 연결된 것을 말하며, 마지막으로 Mixed2(HL)는 평균점수가 높고 낮은 검사들이 교대로 연결된 것을 말한다. Tucker와 빈도추정법(frequency estimation method)을 사용했을 때, Low와 Mixed1(LH)에서 보다 Higher Mixed2(HL)에서 전체오차가 낮았다. Levine방법과 IRT방법을 썼을 때는 High strain에서 전체오차가 작았다. IRT방법은 연구에 사용된 유형 모두에서 전체오차가 비슷했다. 결론적으로 IRT방법이 연구에 사용된 유형의 연결선 모두에서 오차가 가장 적었고, 빈도추정법이 다음으로 적었고, 선형동등화 방법이 가장 오차가 컸다. 그러나 시뮬레이션에 사용된 자료가 IRT모델의 가정을 모두 충족한 반면 다른 방법들의 일부 가정에는 맞지 않았을 수 있다.

Key Words : 검사 동등화, 동등화 연결선, 검사 동등화 오차, 공통문항 · 비동등집단설계법, 빈도추정법, 문항반응이론