A comparison of the standardization and IRT methods of adjusting pretest item statistics using realistic data

Shun-Wen Chang
National Taiwan Normal University

Bradley A. Hanson and Deborah J. Harris
ACT, Inc.

Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, April 2001.

Abstract

The requirement of large sample sizes for calibrating items based on IRT models is not easily met in many practical pretesting situations. Although classical item statistics could be estimated with much smaller samples, the values may not be comparable across different groups of examinees. This study extended Chang, Hanson, and Harris (2000) by further exploring the standardization method and comparing its effectiveness with the one-parameter (1PL) and three-parameter (3PL) logistic IRT models in adjusting pretest item statistics with small sample sizes, using more realistic data than the previous study.

Based on the realistic data generated from a 50-dimensional MIRT model, the 3PL model performed better than the 1PL or standardization method in recovering both the population p-values and point biserial correlations. The standardization method outperformed the 1PL model in recovering the population point biserial correlations, but not in recovering the population p-values. The performance of the methods was also evaluated using the real pretest data of a high-stakes test. In terms of recovering the p-values and point biserial correlations for the real data, the 1PL model produced the most satisfactory results. The 3PL model performed worst in terms of recovering the p-values for the real data, and the standardization method performed worst in recovering the point biserial correlations for the real data.

Due to the very limited number of conditions studied, one must be cautious about making conclusions about the standardization method relative to IRT methods based on these studies. The standardization method appears to be a viable alternative to IRT methods that may be simpler to implement, although these results do not suggest that it will produce more accurate results.

Acknowledgement

The authors wish to thank Qing Yi for assisting in the parallel form creation and Bor-Yaun Twu for programming part of the data simulation.

When a large number of items are to be pretested, but concerns about test security prevent all pretest items from being administered to the same group of examinees, the pretest items can be included in parallel forms and be administered to different small groups of examinees. The groups used for pretesting could be conveniently formed based on school and concerns about item exposure may preclude the administration of different forms to the same school using a spiraling process. If the groups taking the different forms represent samples from different schools, then the groups are nonequivalent.

Based on specific groups of examinees that are administered the test, classical item statistics such as item difficulty (i.e., the p-value) and the item discrimination (i.e., the biserial or point biserial correlation) are computed. A sample size of 150 to 200 examinees is usually sufficient to obtain stable estimates of these statistics. However, the classical item statistics based on nonequivalent groups are not directly comparable, posing a problem in the test development process. Instead of directly computing classical item statistics, item parameters of IRT models can be estimated using the response data. These parameters in each group can be converted to be on the same scale using IRT scaling or transformation methods. Estimates of the classical item statistics for all items in a particular group can be computed using the IRT parameter estimates.

The Rasch model, or the one-parameter logistic (1PL) model, specifies that the item difficulty is the only item characteristic that varies from item to item, holding the item discrimination values equal for all items. Previous research suggested that a sample size of as small as 200 examinees would be sufficient to accurately estimate item parameters of the 1PL model (Wright & Stone, 1979). However, the 1PL model may not provide a good fit to multiple-choice items where discrimination indices are usually unequal, and examinees are likely to guess on the items. The three-parameter logistic (3PL) model (Birnbaum, 1968) is a more general model where the discriminating power is allowed to vary among items and guessing is allowed to occur for the examinees. However, in order to accurately estimate the 3PL item parameters, previous research suggests that at least 1,000 (Reckase, 1979; Skaggs & Lissitz, 1986) to 10,000 (Thissen & Wainer, 1982) examinees would be needed. The requirement of large sample sizes in practical pretesting situations can be hard to meet.

As an alternative to IRT item calibration with small sample sizes, Chang, Hanson and Harris (2000) proposed a standardization approach to adjust conventional item statistics obtained from small nonequivalent samples to more closely represent the item statistics that would have been obtained if the population of interest has been employed. This method is implemented by incorporating a set of common items across the various forms of a test and assuming that the conditional distributions of unique or noncommon items given the number correct score on the common items are the same across all groups of examinees. A joint distribution of a unique item score and the number correct score on the common items (common item score) in the total group of examinees can then be obtained. An estimate

of the classical p-value in the entire population (i.e., the average probability of correctly answering an item in the total group of examinees) can be obtained from the joint distribution by summing over the common item scores for the correct unique item response. The point biserial correlation between the unique item score and common item score can also be obtained from this joint distribution.

In order to understand the standardization method, let Ug and Xcg be random variables representing the score on a unique item and the number correct score on a set of m common items, respectively, in examinee subpopulation g. The joint distribution of the unique item and common item scores in the subpopulation g can be written as

$$\Pr(Ug=u, Xcg=x) = \Pr(Ug=u|Xcg=x)\Pr(Xcg=x),\ g = 1,...,G;\ u = 0, 1;\ x = 0, 1,...,m,$$
$$= 0, \text{ elsewhere.}$$

The notation of Ug = u represents the event of the random variable Ug taking on the value of u, with the value of 1 for a correct response and 0 for an incorrect response in the subpopulation g, and Xcg = x represents the the event of the random variable Xcg being equal to number correct score x in the subpopulation g, with the values of 0 to the number of common items, m. $\Pr(Ug=u|Xcg=x)$ is the conditional distribution of the unique item score given common item score and $\Pr(Xcg=x)$ is the marginal distribution of the common item score in the subpopulation g.

For the entire population o (i.e., the examinees across all subpopulations), the joint distribution of the unique item and common item scores can be represented as

$$\Pr(Uo=u, Xco=x) = \Pr(Uo=u|Xco=x)\Pr(Xco=x),\ u = 0, 1;\ x = 0, 1,...,m,$$
$$= 0, \text{ elsewhere.}$$

Based on the assumption that the conditional distributions of the unique item response given common item score are the same across all groups, the above joint distribution can be written as

$$\Pr(Uo=u, Xco=x) = \Pr(Ug=u|Xcg=x)\Pr(Xco=x) \text{ for any subpopulation, g,}$$

where $\Pr(Xco=x)$, the distribution of the common item scores for the entire population o, is estimated based on the responses of all groups to the set of common items. Using this joint distribution inthe entire population, estimates of classical item statistics that would have been obtained if the item had been given to a sample from the entire population can be calculated. An estimate of the classical p-value in the entire population (i.e., the average probability of correctly answering an item in the population of

examinees) can be obtained from the joint distribution Pr(Uo=*u*, Xco=*x*) by summing over the common item scores for the correct unique item response. The point biserial correlation between the unique item score and common item score can also be obtained from this joint distribution of the unique item and common item scores in the population of examinees.

Random error in this standardization method can be reduced using smoothing methods. A bivariate polynomial log-linear model analogous to that described in Hanson (1991) and Rosenbaum and Thayer (1987) can be employed to smooth the joint distribution of the unique and common items Pr(Ug=*u*, Xcg=*x*). The marginal distribution of the common items Pr(Xco=*x*) can be smoothed using a univariate polynomial log-linear model described in Kolen (1991).

Chang et al. (2000) showed that both the standardization and the 3PL methods outperformed the 1PL model in recovering the population p-values and point biserial correlations. In estimating the population p-values, the performance of the standardization method was similar to that of the 3PL model. In estimating the point biserial correlations, the standardization method performed slightly worse than the 3PL model. However, the results in Chang et al. (2000) were based on data generated using a 3PL model with a sample size of 250 examinees in each group. When the data do not perfectly fit a 3PL model, the relative performance of the standardization method and the 3PL model may differ. The 3PL model-genertated data might have created a bias in favor of the 3PL model. Also, while the classical item statistics might be stable based on samples as small as of 150 to 200 examinees and the standardization method might still satisfactorily recover the population item statistics, parameter estimation based on the 3PL model may not be justified on so small a sample and its performance could deteriorate. The current study was intended to investigate the comparative performance of the standardization method and both the 1PL and 3PL models with a realsitic data set simulated using a high dimensional MIRT model (Davey, Nering & Thompson, 1997). In addition to the sample size of 250 examinees in each group, a condition of 100 examinees in each group was also used to represent a situation where the group size was very small. The three methods were also applied to a real pretest data set.

The Purpose

This study was designed to compare the effectiveness of the standardization method (Chang et al., 2000) with IRT methods for scaling pretest item statistics to be on the same scale using both realistically simulated data and real pretest data. This study was concerned with the case in which pretest item statistics were obtained using a common item nonequivalent groups design. When the sample sizes for the groups of examinees are small, employing traditional IRT item calibration or scaling may not be justified due to the large sample size requirement. This study examined the standardization method of

adjusting conventional item statistics which may have a less strict sample size requirement. Specifically, this study attempted to achieve the following objectives:

1. to explore the standardization method using realistic simulated data sets;
2. to compare the item statistics recovery of the 1PL model, the 3PL model and the standardization method with sample sizes as small as 100 and 250.

Method and Data

The first part of this research evaluated the performance of the methods using realistic data simulated from a MIRT model. The second part of this research studied the performance of the methods using real pretest data. The plan of the study design, preparation of the data, and methods for data analyses for the two parts of this study are described in separate sections below.

I. The Simulation Study

The Test Forms

Ten test forms were built to be as parallel as possible in their content and statistical specifications. Each form consisted of 24 unique items and 12 common items. Item parameters calibrated from multiple forms of a large scale standardized math test based on a 50-dimensional MIRT model (Davey et al., 1997) using NOHARM (Fraser, 1986) were available for this study. For the purpose of this study ten test forms were created based on these item parameters using the following steps:

1. A 50-dimensional normal ogive MIRT model was fit to a large number of the examinee responses for several forms of the math test using NOHARM to obtain for each item a set of 50 a-parameters, one b-parameter, and one c-parameter.
2. 10,000 examinee responses to all items available were generated based on the 50-dimensional model using the item parameters obtained from step 1. When generating response data from the multidimensional model, the theta vector for each examinee was generated from a 50-variate N($\underline{0}$, I) distribution of ability (i.e., a standard multivariate normal distribution with a mean vector of zero and an identity covariance matrix).
3. Using the10,000 examinee responses generated in step 2, the p-value and the correlation between the item and the total score on all items (i.e., the point biserial correlation) were computed for each item.
4. Based on the content categories and the p-values and point biserial correlations of the items computed in step 3, ten forms each containing 36 items were created to be as parallel as possible. However, the degree of the parallelism may be somewhat limited due to the number of appropriate items available. In each form, the 24 unique items were followed by the 12 common items.

By following the steps above, the forms were built based on the model to be used to generate the

response data of the various examinee groups for the current study. The item parameters of the 50 dimensional MIRT model estimated from NOHARM were treated as the population item parameters.

The Samples and Population

Ten groups of examinees were generated with sample sizes of 1,000 and 2,500 examinees across all groups (approximately 100 and 250 examinees per group), respectively, to represent a situation where the group size was small. Examinee abilities were generated from a 50 dimensional standard multivariate normal distribution. Also, responses to items on an additional form of the math test were generated for each examinee. This form contained items that were different from any of the items used for unique or common items. This form will be referred to as the assignment form. These generated math scores were used for assigning examinees to the ten groups. The score on the assignment form was used to assign examinees to groups due to the difficulty in specifying group differences in terms of the entire 50 dimensional vector of examinee abilities. The rules for assigning examinees to the various groups were constructed so that examinees with lower scores were more likely to be assigned to the lower examinee group than examinees with higher scores, and the group differences were similar in magnitude to group differences observed with the real pretest data in this study.

In order to obtain group differences that were comparable to those in the real pretest data, some trial and error was needed. The first step was to find the quartiles (25th, 50th, 75th percentiles) of the number correct score distribution of the assignment form. The second step was to make and adjust the conditional probabilities of examinees being assigned to each group given an examinee's score of the assignment form was in the 1st, 2nd, 3rd, or 4th quartile of the distribution in order to produce group differences comparable to those in the real data. Third, the examinees were assigned to the various groups based on the conditional probabilities. The conditional probabilities for each quantile of the assignment form score used to assign examinees to groups are given in Table 1. Table 2 gives descriptive statistics of the scores on the 12 common items for 10,000 simulated examinees, who were assigned to groups based on the conditional probabilities presented in Table 1. The number in each group was not exactly 100 or 250 because the group was randomly determined based on the number correct score for the assignment form, and also, due to the fact that the quartiles did not exactly separate the raw score distribution of the assignment form into four even parts. The total sample size across all groups is always equal to 1,000 or 2,500. The assignment rule adopted for this study yielded group differences that were fairly close to those in the real pretest data set. The population ability distribution across all groups was a standard multivariate normal distribution.

Once an examinee was assigned to a group, item responses to the unique items in that group and the common items were generated using the population MIRT item parameters and the generated 50

dimensional ability vector for the examinee.

The Population Item Statistics

The population conventional item statistics (both the p-values and point biserial correlations) were obtained based on the population item parameters and the population ability distribution. The population p-value of an item was computed by evaluating the integral $p = \int \Pr(U=1|\theta)\, fo(\theta) d\theta$, where $\Pr(U=1|\theta)$ is the conditional probability of the correct item response given a particular θ value and $fo(\theta)$ is the ability distribution of the overall population. Monte Carlo numerical integration with 500,000 replications was used to evaluate the 50 dimensional integral. Since the various forms of the test consisted of different unique items and a set of common items, the population point biserial correlation was defined in this study as the correlation between the unique item score and common item score instead of the total test score. This correlation was computed based on the joint distribution of the unique item and the common items in the overall population. The Monte Carlo numerical integration for evaluating the 50 dimensional integral was employed in computing the population point biserial correlations.

The Estimated Item Statistics

The estimated p-values and point biserial correlations were obtained using the 1PL model, the 3PL model and the standardization method, respectively. For both the 1PL and 3PL models, BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) was utilized first to estimate the item parameters for all unique and common items simultaneously using multiple group estimation. Then, these item parameters were converted to the conventional p-values and point biserial correlations. The p-value was computed by evaluating the integral

$$\hat{p} = \int \hat{\Pr}(U=1|\theta)\, fo(\theta) d\theta,$$

where $\hat{\Pr}(U=1|\theta)$ is the conditional probability of the correct item response given a particular θ value and is calculated using the item characteristic curve (ICC)

$$\hat{\Pr}\left(U=1\middle|\theta\right) = c + (1-c)\frac{e^{1.7a(\theta-b)}}{1+e^{1.7a(\theta-b)}},$$

where a, b and c are parameters that characterize the item. For the 3PL model, all the three parameters can vary across items while for the 1PL model, the a parameter is a constant and the c parameter is 0 for

all items. The distribution $fo(\theta)$ is the ability distribution of the overall population. Specifically, the estimated p-value of an item was derived by approximating this continuous distribution $fo(\theta)$ by averaging over the discrete empirical distributions of the ten groups, respectively, from the phase 2 results of the BILOG-MG estimation.

The point biserial correlation was computed based on the joint distribution of the unique item and the common items in the overall population $\hat{\Pr}(Uo=u, Xco=x)$. This joint distribution $\hat{\Pr}(Uo=u, Xco=x)$ is given by

$$\begin{aligned}&\hat{\Pr}(Uo=u, Xco=x)\\ &=\int \hat{\Pr}(U=u, Xc|\theta)\, fo(\theta)\, d\theta\\ &=\int \hat{\Pr}(U=u|\theta)\, \hat{\Pr}(Xc|\theta)\, fo(\theta)\, d\theta\end{aligned}$$

where $\hat{\Pr}(U=1|\theta)$ was calculated using the ICC and $\hat{\Pr}(U=0|\theta)$ is $1-\hat{\Pr}(U=1|\theta)$. $\hat{\Pr}(Xc|\theta)$, the distribution of the number correct common item scores conditioned on a particular θ value, was obtained by the Lord-Wingersky algorithm (Lord & Wingersky, 1984). In this study, there were 13 possible values of $\hat{\Pr}(Xc|\theta)$ for a particular θ, one for each of the common item scores of 0, 1,..., 12.

To derive this joint distribution $\hat{\Pr}(Uo=u, Xco=x)$, the continuous distribution $fo(\theta)$ was also approximated with the discrete empirical distributions of the ten groups from the phase 2 outcomes of the BILOG-MG estimation. The correlation based on values of this bivariate distribution $\hat{\Pr}(Uo=u, Xco=x)$ was the estimated point biserial correlation of interest in this study, which can be written as

$$\hat{\rho} = \frac{\sum ux\,\hat{\Pr}(U_o=u, Xc_o=x) - \sum u\,\hat{\Pr}(U_o=u)\sum x\,\hat{\Pr}(Xc_o=x)}{\sqrt{\sum u^2\,\hat{\Pr}(U_o=u) - \left[\sum u\,\hat{\Pr}(U_o=u)\right]^2}\sqrt{\sum x^2\,\hat{\Pr}(Xc_o=x) - \left[\sum x\,\hat{\Pr}(Xc_o=x)\right]^2}},$$

where $\hat{\Pr}(Uo=u, Xco=x)$ is the joint distribution of the unique item and common item scores and $\hat{\Pr}(Uo=u)$ and $\hat{\Pr}(Xco=x)$ are the marginal distributions of the unique item and common items, respectively, estimated based on the 1PL or 3PL model.

The classical p-value and point biserial correlation were also computed using the standardization method for each unique item based on the response data in each of the ten groups. In this study, the joint distribution of the unique and common items $\Pr(Ug=u, Xcg=x)$ was smoothed using the bivariate polynomial log-linear model (Hanson, 1991; Rosenbaum & Thayer, 1987). The model used polynomial degree 5 for the common item number correct score, degree 1 for the unique item score, and degree 2 for the interaction of the common item number correct score and the item score. The marginal distribution

of the common items Pr(Xco=*x*) across all groups was smoothed using a univariate polynomial log-linear model in Kolen (1991) with polynomial degree 5. This smoothed marginal distribution was used with the smoothed bivarate distribution in each group to produce estimates of the p-values and point biserial correlations.

Replications

The above process of estimating the item statistics was replicated 500 times for both the 1,000 and 2,500 sample size conditions, respectively.

II. The Real Pretest Data Application

The methods were also evaluated using real pretest data from a high-stakes Social Science test currently under development. The concerns about test security prevented all pretest items from being administered to the same group of examinees, so the pretest items were included in parallel forms and administered to different small groups of examinees. In each form of the test, there were 55 unique items and a set of 15 common items which were located at different positions in each form. This study investigated eight forms of the test, a total of 455 items, taken by eight groups of examinees, respectively. The groups used for pretesting were somewhat conveniently formed based on school and the concerns about item exposure precluded the administration of different forms to the same school using a spiraling process. Because the groups taking the different forms represented samples from different schools, the groups were nonequivalent.

The data-collection process was based on a design similar to that used in the first part of this study. However, due to practical limitations during pretesting, there existed in this data set some differences from the simulated realistic data such that the forms may not be as parallel, the extent to which the common items were representative of the total test may not be as great, and the variation in item statistics of the pretest items was greater.

Table 3 presents the summary statistics of the common item scores for the various groups on this Social Science test. The N column lists the number of examinees for each group, which varied slightly among the groups.

The Criteria

For the simulation part of this study, the population p-values and point biserial correlations were used as the baselines for evaluating the accuracy of the estimated p-values and point biserial correlations based on the 1PL model, the 3PL model, and the standardization method. Two indices were used as the criteria. One was the Pearson product-moment correlation coefficient between the estimated and

population item statistics. The other criterion was the mean square error (MSE) over items. The MSE value is the expected squared difference between the estimated and population item statistics and can be decomposed into variance and squared bias. Variance is the average squared difference between the estimated and the expected value of the estimated item statistics across replications. Bias is the difference between the expected value of the estimated item statistics and the population value across replications.

Provided below were the formulas used to compute the MSE, variance and squared bias for the p-value over the 500 replications with respect to each of the unique items *i*.

$$MSE_i = \frac{1}{500}\sum_{r=1}^{500}\left(\hat{p}_{ir} - p_i\right)^2,$$

$$Variance_i = \frac{1}{500}\sum_{r=1}^{500}\left(\hat{p}_{ir} - \bar{p}_i\right)^2, \text{ and}$$

$$Bias_i^2 = \left[\frac{1}{500}\sum_{r=1}^{500}\left(\hat{p}_{ir} - p_i\right)\right]^2,$$

where $r = 1, 2,..., 500$, $\hat{p}_{ir}$ is the estimated p-value of the unique item i for the rth replication, $\bar{p}_i$ is the mean of the estimated p-values across the 500 replications, and p_i is the population p-value of the unique item i.

For the point biserial correlations, the following formulas were used to compute the MSE, variance and squared bias over the 500 replications with respect to each of the unique items i.

$$MSE_i = \frac{1}{500}\sum_{r=1}^{500}\left(\hat{\rho}_{ir} - \rho_i\right)^2,$$

$$Variance_i = \frac{1}{500}\sum_{r=1}^{500}\left(\hat{\rho}_{ir} - \bar{\rho}_i\right)^2, \text{ and}$$

$$Bias_i^2 = \left[\frac{1}{500}\sum_{r=1}^{500}\left(\hat{\rho}_{ir} - \rho_i\right)\right]^2,$$

where $r = 1, 2,..., 500$, $\hat{\rho}_{ir}$ is the estimated point biserial correlation of the unique item i for the rth replication, $\bar{\rho}_i$ is the mean of the estimated point biserial correlations across the 500 replications, and ρ_i is the population point biserial correlation of the unique item i.

The average values of the MSE, variance and squared bias over items were computed as the criteria for the comparisons among the various methods. To facilitate the comparison among the average MSE for the various methods, the standard errors of the mean MSE over items were provided to indicate whether the differences among the average MSE values for the various methods were large relative to the errors introduced by estimating these averages by simulation using 500 replications. The standard error of the mean MSE over items (i.e., the variability over the 500 replications of the average MSE over items) was computed by

$$\sqrt{\frac{\sum_{r=1}^{500}\left(MSE_r^i - \overline{MSE^i}\right)^2}{500}} \Big/ \sqrt{500}\ .$$

For the rth replication, $MSE_r^i = \frac{1}{240}\sum_{i=1}^{240}(\hat{p}_{ri} - p_i)^2$ for the p-value and $MSE_r^i = \frac{1}{240}\sum_{i=1}^{240}(\hat{\rho}_{ri} - \rho_i)^2$ for the point biserial correlation, where i = 1, 2,..., 240 is the number of items. $\overline{MSE^i}$ is the average of the MSE_r^i values across the 500 replications.

For the real pretest data application, the performance of the various approaches was evaluated by treating one of the 15 common items as a non-common item. That is, that item was treated as a unique item in each form even though it was actually the same item. The common item chosen was the one whose item statistics differed more among the groups than any other 14 common items. In each group, estimates of the item p-value and point biserial correlation for that common item were computed employing the 1PL, 3PL and standardization methods, respectively. These estimates were compared to the population item statistics for that common item using data from all forms. The population p-value was the proportion correct of that common item over all groups of the examinees. The population point biserial correlation was the correlation of that item and the score on the other 14 common items using all the data.

Results

The first part of this section presents the results of the 1PL model, the 3PL model and the standardization method using the simulated realistic data with respect to each of the criteria: the Pearson product-moment correlation, the MSE, variance and squared bias for both the p-values and point biserial correlations. The second part of this section contains outcomes of the real pretest data applications for the three methods.

I. The Simulation Study

The Results in Terms of the Pearson Product-Moment Correlation

Summary statistics for the Pearson product-moment correlation coefficients between the estimated and population item statistics are displayed in Table 4 for the various methods. Since the sample size per form was not exactly 100 and 250 due to the way the data were generated in the current study, the two samples were labeled as 1,000 or 2,500, representing the total sample size across all forms. The N column indicates the number of replications.

It can be seen that for the p-values, all three methods recovered the population p-values to a similar degree under the two sample size conditions. The average correlation for the 3PL model was slightly higher than that for the 1PL or standardization method. In fact, it may be concluded that these three methods performed equally well in recovering the population p-values.

Table 4 reveals that all methods performed less well in recovering the point biserial correlations than the p-values. The correlations were higher for all three methods when the sample size was 2,500 than when the sample size was 1,000. For either of the two sample sizes, the 3PL model yielded the highest average correlation between the estimated and population point biserial correlations. The standardization method performed less well than the 3PL model in recovering these correlations and the difference was greater for a sample size of 1,000 than for a sample size of 2,500. The results indicated that the performance of the 1PL model was very poor for both sample sizes.

The Results in Terms of the MSE, Variance and Squared Bias

Table 5 shows the summary statistics for p-value MSE, variance and squared bias over the 240 items for the two sample sizes, respectively. The N column shows the number of the unique items. It can be seen that for either of the two sample sizes, the average MSE value over items was the lowest for the 3PL method. The average MSE was lower for the 1PL method than the standardization method. When the sample size was 2,500, the differences among these three methods became smaller. Relative to the errors in the estimates due to estimation by simulation with the 500 replications, the differences among these three methods were large for both sample size conditions. The results showed that the 3PL method yielded the smallest overall error for estimating the p-values, followed by the 1PL method. The standardization method had the largest average MSE across items.

With respect to the variance, the average value over items was lower for the 3PL method than for the other two methods. Their differences were smaller when the sample size was 2,500. The average squared bias for the 3PL model was lower than that for the 1PL model. The average values of the squared bias were slightly lower for the 1PL model than the standardization approach.

Displayed in Table 6 are the summary statistics of the MSE, variance and squared bias over the 240 items for the point biserial correlations. There seems to be greater differences among the methods with regard to the point biseral correlations than there were for the p-values. The average MSE over the 240 items was still the lowest for the 3PL model for both sample sizes. When the sample size of 1,000 was used, the 1PL model produced a lower average MSE value than the standardization method, but when the sample size of 2,500 was employed, the average MSE was lower for the standardization method than for the 1PL model. Relative to the standard errors of the average MSE introduced by simulation using 500 replications, differences among the methods were quite large. These findings seem to suggest that the 3PL model performed better overall than the 1PL or standardization approach in terms of estimating the point biserial correlations. Employing the larger sample size of 2,500 seems to substantially lower the MSE over items for the standardization method in estimating the point biserial correlations.

While the average variance over items for the 3PL method was the lowest for the p-values (see Table 5), the average variance for the 1PL method was the lowest for the point biserial correlations in Table 6. The standardization method had the hightest average value of the variance. For the squared bias, it can be seen that the average values for both the 3PL and standardization methods were substantially lower than that for the 1PL model, no matter what sample size was used. The squared bias was slightly lower for the standardization method than the 3PL model when the sample size was 1,000, but the squared bias was lower for the 3PL method when the sample size was 2,500.

II. The Real Pretest Data Application

In the real pretest data applications, problems occurred when the BILOG-MG 3PL model was fit to the response data. Eight unique items were not being calibrated for their b-parameters due to very low or negative point-biserial correlations. These items were discarded from the analyses of this study, so 447 items remained. For the standardization method, a fifth degree polynomial was used for the common item score and a second degree polynomial was used for the interaction of the common item number correct score and the item score. Across the eight groups, the proportion correct of the common item being treated as a unique item (i.e., the population p-value) was .50161 and the correlation between that item and the total score of the other 14 common items (i.e., the population point biserial correlation) was .39877.

Results in Table 7 are the item statistics of the common item as a unique item estimated by the 1PL, 3PL and standardization methods, respectively, in each of the eight groups. The unadjusted item statistics in each group are also reported in Table 7. The variation of the estimated point biserial correlations of the 3PL and standardization methods were somewhat large. The reason might be that the

common item was chosen whose item statistics differed more among the groups than any of the other 14 common items.

The differences between the estimated and population p-values and point biserial correlations are contained in Table 8, along with the average of the absolute differences across groups. The average absolute p-value difference for the 1PL model was the smallest and the average absolute difference for the 3PL model was the largest. In terms of the average absolute point biserial correlation difference, the standardization method produced the largest value and the 1PL model resulted in the smallest value. The 1PL method estimated the point biserial correlations that were the closest to the population value. Compared with its performance in the realistic data simulated using the 50-dimensional MIRT model, the 3PL model seemed not to work as well with the real pretest data in making the adjustments of the item statistics. The 1PL model seemed to perform better with the real pretest data than the simulated. All the adjustment methods worked better on average than not performing any adjustment.

Conclusions

Based on the response data of the various examinee groups, IRT models can be employed to estimate item parameters and convert these parameters to be on the same scale using IRT scaling or transformation methods. However, when the examinee groups are small, employing traditional IRT item calibration or scaling may not be justified due to the large sample size requirement. This study explored the standardization approach to adjust conventional item statistics which has a less strict sample size requirement. The purpose of using the standardization method was to adjust the item statistics obtained from small nonequivalent samples to more closely represent the item statistics that would have been obtained if the item had been administered in all groups. This method was implemented by incorporating a set of common items across the various forms of a test and using the assumption that the conditional distributions of unique or noncommon items given common items were the same across all groups of examinees.

The investigations in the current study were proceeded by simulating realistic data from a 50-dimensional MIRT model. Ten forms were created to be as parallel as possible in their content and statistical specifications. The examinees' abilities were generated from a standard multivariate normal distribution. The examinees were then assigned to the various groups using the generated observed scores of an additional form of the math test. The rule used to assign examinees to groups resulted in examinees with lower scores being more likely to be assigned to the lower examinee groups than examinees with higher scores. The group differences obtained were similar in magnitude to the group differences observed with the real pretest data in this study. The two samples of 1,000 and 2,500 were generated for the ten groups of examinees, with sizes of about 100 and 250 in each group, respectively.

The effectiveness of the standardization approach was compared with that of the 1PL and 3PL models using the Pearson product-moment correlation, the MSE, variance and squared bias as the criteria for evaluation. The results showed that with respect to estimating the p-values, the 3PL model produced the smallest overall error, followed by the 1PL model, for both sample size conditions. The overall error for the standardization method was slightly higher than the 1PL model. For the estimation of the population point biserial correlations, the 3PL model still performed better than the 1PL or standardization method for either of the two sample sizes. When the sample size was 1,000, the 1PL model outperformed the standardization method, but when the sample size was 2,500, the standardization method performed better. Employing the larger sample size of 2,500 substantially lowered the MSE over items for the standardization method.

It was somewhat surprising that the 3PL model continued to perform fairly well, even with a sample size of about 100 per unique item. These results might be in part due to the groups being formed based on the score on another math form. Consequently, the groups differed on the same basic composite trait being measured by the items. If the groups had been based on another score such as English and Reading, the IRT models might not have been able to adjust the p-values and point biserial correlations as well. Further studies based on some other criterion for the assignment into groups could investigate this issue. Perhaps the concern about the inaccuracy in parameter estimation using the BILOG-MG 3PL model when calibration sample sizes are small could be possibly lessened when the parameter estimates are used to estimate p-values and point biserial correlations.

The performance of the three methods was also evaluated using the real pretest data of a Social Science test. The results were not consistent with those of the realistic data simulated based on the 50-dimensional MIRT model. In estimating the population p-value of the common item chosen to be a unique item in each form, the 1PL model performed better than the standardization method, and the standardization method performed better than the 3PL model. In estimating the population point biserial correlation of that item, the 1PL model produced the best outcome, followed by the 3PL model. The standardization method resulted in the poorest estimation. However, it is important to recognize that the results with the real pretest data were only based on one common item whose item statistics differed more among the groups than any other common item. It might be worthwhile to examine the relative performance of the methods on the adjustments of the p-values and point biserial correlations for other common items.

The results showed that the relative performance of the 1PL, 3PL and standardization methods was not consistent between the simulated and real pretest data sets. When the simulated realistic data were employed, the standardization method proposed in this study failed to outperform the 3PL model in recovering the population p-values and point biserial correlations. But when the real pretest data were

used, the standardization method seemed to perform better than the 3PL model in recovering the population p-value, although not in recovering the population point biserial correlation. Results for which of the methods performed better could not be suggested based on the findings of this study. It is unknown what accounted for the differences in the adjustments. Factors such as the number and location of the common items, the representativeness of common items to the entire test, the item characteristics of both unique and common items, the variation in item statistics, and the degree of form parallelism might have contributed in part to the differences of the results. Also, the standardization procedure was carried out based on the assumption that the conditional distributions of unique items given common items are the same across all groups of examinees. It is not known the extent to which the assumption held for either of the simulated and real pretest data. Further investigations might be attempted to look at these effects on the performance of the various methods.

Moreover, it might be beneficial to examine the performance of the various approaches when group differences are greater. In the current study, the amount of group differences for both simulated and real pretest data was fairly small, so it may not have required large adjustments on the p-value and point biserial correlations for the various methods. The issue of the degree to which the methods are affected by group differences needs more study.

Unadjusted p-values and points biserial correlations were reported for the real data along with the adjusted values produced by the three adjustment methods. On average, each of the adjustment methods produced more accurate results than using the unadjusted values. While the adjusted values were closer to the overall value in some cases, the adjusted values were just about as far away from the overall value as the unadjusted values were. This raises the question of whether any of the adjusted values, even though they are better than the unadjusted values, is providing item statistics that are accurate enough to rely on for constructing test forms.

The requirement of large sample sizes for calibrating items based on IRT models is not easily met in many practical pretesting situations. Although classical item statistics could be estimated with much smaller samples, the values may not be comparable across different groups of examinees. This study further explored the standardization method and compared its effectiveness with the IRT methods in adjusting pretest item statistics with small sample sizes using more realistic simulated data than used in Chang et al. (2000). One must be cautious about making conclusions about the standardization method relative to IRT methods based on these studies due to the very limited number of conditions studied. Still, in most cases in both studies the IRT method produced more accurate results than the standardization method, although the difference in performance between the standardization method and the better performing IRT method was not large. There was also some inconsistency with regard to which IRT method performed better than the standardization method. The standardization method

appears to be a viable alternative to IRT methods that may be simpler to implement, although these results do not suggest that it will produce more accurate results.

References

Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Chang, S. W., Hanson, B. A., & Harris, D. J. (2000, April). *A standardization approach to adjusting pretest item statistics.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Davey, T., Nering, M., & Thompson, T. D. (1997). *Realistic simulation of item response data* (Research Report 97-4). Iowa City, IA: ACT, Inc.

Fraser, C. (1986). *NOHARM: A computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory* [Computer Program]. Center for Behavioral Studies, The University of New England, Armidale, New South Wales, Australia.

Hanson, B. A. (1991). A comparison of bivariate smoothing methods in common-item equipercentile equating. *Applied Psychological Measurement, 15*, 391-408.

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

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

Reckase, M. D. (1979). Unifactor latent trait models applied to multi-factor tests: Results and implications. *Journal of Educational Statistics, 4*, 207-230.

Rosenbaum, P. R., & Thayer, D. T. (1987). Smoothing the joint and marginal distributions of scored two-way contingency tables in test equating. *British Journal of Mathematical and Statistical Psychology, 40*, 43-49.

Skaggs, G., & Lissitz, R. W. (1986). *The effect of examinee ability on test equating invariance*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47*, 397-412.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

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

Table 1. Conditional Probabilities of Being Assigned to Groups

	Quantile of Score on Additional Math Form			
	1st	2nd	3rd	4th
Group 1	0.15000	0.12500	0.07500	0.05000
Group 2	0.14000	0.12500	0.07500	0.06000
Group 3	0.12500	0.12500	0.07500	0.07500
Group 4	0.12500	0.10000	0.10000	0.07500
Group 5	0.10000	0.10000	0.10000	0.10000
Group 6	0.10000	0.10000	0.10000	0.10000
Group 7	0.07500	0.10000	0.10000	0.12500
Group 8	0.07500	0.07500	0.12500	0.12500
Group 9	0.06000	0.07500	0.12500	0.14000
Group 10	0.05000	0.07500	0.12500	0.15000

Table 2. Moments of Common Item Score for Simulated Sample of 10000

	N	Mean	SD	Skewness	Kurtosis
Group 1	1038	5.39692	2.44722	0.39510	-0.33676
Group 2	1013	5.48075	2.52632	0.38292	-0.45293
Group 3	945	5.67196	2.66218	0.36747	-0.49959
Group 4	1023	5.65689	2.57235	0.24174	-0.62087
Group 5	996	6.16767	2.71623	0.13235	-0.79604
Group 6	1000	5.99900	2.65745	0.16854	-0.71347
Group 7	1033	6.37754	2.70994	0.13243	-0.73396
Group 8	1022	6.54305	2.72429	0.05894	-0.81276
Group 9	993	6.82578	2.62180	-0.00167	-0.76619
Group 10	937	6.97225	2.62818	-0.08643	-0.67459

Table 3. Summary Statistics of the Common Item Scores of the Social Science Pretest Data for the Various Groups

	N	Mean	SD	Skewness	Kurtosis	Minimum	Maximum
Group 1	236	9.38136	3.05807	-0.24872	-0.78745	2	15
Group 2	233	9.77253	2.98195	-0.20143	-0.69539	3	15
Group 3	231	10.21212	3.28012	-0.38077	-0.88370	2	15
Group 4	238	10.23529	3.29609	-0.60110	-0.53485	1	15
Group 5	235	10.42553	3.03220	-0.49267	-0.58135	3	15
Group 6	232	10.48276	2.95415	-0.53891	-0.32423	1	15
Group 7	234	10.53419	3.24945	-0.51341	-0.67732	2	15
Group 8	225	10.70667	3.36250	-0.66154	-0.47533	1	15

Table 4. Summary Statistics of the Correlations between the Estimated and Population Item Statistics

P-Values

1000 Sample Sizes

Method	N	Mean	SD	Minimum	Maximum
1PL	500	0.97458	0.00254	0.96528	0.98181
3PL	500	0.97599	0.00241	0.96833	0.98292
Standardization	500	0.97347	0.00276	0.96525	0.98151

2500 Sample Sizes

Method	N	Mean	SD	Minimum	Maximum
1PL	500	0.98846	0.00115	0.98430	0.99121
3PL	500	0.99001	0.00095	0.98652	0.99326
Standardization	500	0.98791	0.00131	0.98200	0.99074

Point Biserial Correlations

1000 Sample Sizes

Method	N	Mean	SD	Minimum	Maximum
1PL	500	0.19047	0.03750	0.08625	0.30472
3PL	500	0.72769	0.02528	0.64924	0.79472
Standardization	500	0.63362	0.03814	0.49444	0.74145

2500 Sample Sizes

Method	N	Mean	SD	Minimum	Maximum
1PL	500	0.19315	0.02507	0.12369	0.28060
3PL	500	0.80883	0.01750	0.75717	0.86445
Standardization	500	0.75609	0.02486	0.68911	0.81391

Table 5. Summary Statistics of the MSE, Variance and Squared Bias for the P-Values for the Various Methods

1000 Sample Sizes

MSE

Method	N	Mean	SD	Minimum	Maximum	Standard Error
1PL	240	0.00210	0.00049	0.00059	0.00475	0.000009482
3PL	240	0.00191	0.00042	0.00052	0.00293	0.000008696
Standardization	240	0.00218	0.00052	0.00064	0.00475	0.000010285

Variance

Method	N	Mean	SD	Minimum	Maximum
1PL	240	0.00194	0.00040	0.00053	0.00278
3PL	240	0.00186	0.00042	0.00049	0.00288
Standardization	240	0.00201	0.00040	0.00058	0.00285

Squared Bias

Method	N	Mean	SD	Minimum	Maximum
1PL	240	0.00015	0.00027	0.00000	0.00251
3PL	240	0.00005	0.00007	0.00000	0.00045
Standardization	240	0.00017	0.00025	0.00000	0.00223

2500 Sample Sizes

MSE

Method	N	Mean	SD	Minimum	Maximum	Standard Error
1PL	240	0.00093	0.00032	0.00027	0.00314	0.000004155
3PL	240	0.00079	0.00017	0.00022	0.00117	0.000003458
Standardization	240	0.00097	0.00033	0.00029	0.00299	0.000004758

Variance

Method	N	Mean	SD	Minimum	Maximum
1PL	240	0.00078	0.00016	0.00023	0.00113
3PL	240	0.00077	0.00016	0.00022	0.00115
Standardization	240	0.00080	0.00016	0.00024	0.00115

Squared Bias

Method	N	Mean	SD	Minimum	Maximum
1PL	240	0.00015	0.00026	0.00000	0.00221
3PL	240	0.00002	0.00003	0.00000	0.00019
Standardization	240	0.00018	0.00025	0.00000	0.00202

Table 6. Summary Statistics of the MSE, Variance and Squared Bias for the Point Biserial Correlations for the Various Methods

1000 Sample Sizes

MSE

Method	N	Mean	SD	Minimum	Maximum	Standard Error
1PL	240	0.00833	0.01130	0.00011	0.06186	0.000019858
3PL	240	0.00470	0.00396	0.00182	0.02788	0.000023462
Standardization	240	0.01070	0.00649	0.00512	0.05531	0.000050773

Variance

Method	N	Mean	SD	Minimum	Maximum
1PL	240	0.00026	0.00016	0.00011	0.00106
3PL	240	0.00263	0.00072	0.00119	0.00603
Standardization	240	0.00867	0.00212	0.00484	0.01785

Squared Bias

Method	N	Mean	SD	Minimum	Maximum
1PL	240	0.00807	0.01128	0.00000	0.06144
3PL	240	0.00207	0.00386	0.00000	0.02260
Standardization	240	0.00203	0.00584	0.00000	0.04311

2500 Sample Sizes

MSE

Method	N	Mean	SD	Minimum	Maximum	Standard Error
1PL	240	0.00815	0.01123	0.00004	0.06183	0.000011344
3PL	240	0.00326	0.00456	0.00092	0.02738	0.000014745
Standardization	240	0.00537	0.00607	0.00224	0.04633	0.000024171

Variance

Method	N	Mean	SD	Minimum	Maximum
1PL	240	0.00010	0.00006	0.00004	0.00037
3PL	240	0.00155	0.00044	0.00067	0.00342
Standardization	240	0.00335	0.00077	0.00198	0.00631

Squared Bias

Method	N	Mean	SD	Minimum	Maximum
1PL	240	0.00805	0.01122	0.00000	0.06167
3PL	240	0.00171	0.00447	0.00000	0.02430
Standardization	240	0.00202	0.00591	0.00000	0.04173

Table 7. Results of the Estimated Item Statistics of the Common Item as a Unique Item by Method and Group

	Estimated P-Values				Estimated Point Biserial Correlations			
	Unadjusted	1PL	3PL	Standardization	Unadjusted	1PL	3PL	Standardization
Group 1	0.36864	0.43066	0.45737	0.41407	0.31868	0.36717	0.42806	0.34504
Group 2	0.47210	0.51273	0.52953	0.50512	0.31853	0.37521	0.39379	0.33898
Group 3	0.53247	0.53535	0.54754	0.53485	0.54047	0.37580	0.46054	0.52815
Group 4	0.53782	0.53673	0.53271	0.53549	0.36957	0.37581	0.37769	0.36332
Group 5	0.49362	0.47589	0.48308	0.48213	0.36989	0.37274	0.34426	0.37631
Group 6	0.51724	0.49969	0.49940	0.50752	0.27434	0.37455	0.28965	0.28130
Group 7	0.58120	0.55865	0.56663	0.55921	0.50663	0.37564	0.43942	0.49266
Group 8	0.51111	0.46573	0.44465	0.46096	0.43649	0.37174	0.38609	0.41275
Mean	0.50177	0.50193	0.50761	0.49992	0.39183	0.37358	0.38994	0.39231
SD	0.06273	0.04265	0.04366	0.04674	0.09470	0.00300	0.05504	0.08232

Table 8. Differences of the Estimated and Population Item Statistics of the Common Item as a Unique Item by Method and Group

	P-Value Difference				Point Biserial Correlation Difference			
	Unadjusted	1PL	3PL	Standardization	Unadjusted	1PL	3PL	Standardization
Group 1	-0.13297	-0.07095	-0.04424	-0.08754	-0.08009	-0.03160	0.02929	-0.05373
Group 2	-0.02951	0.01112	0.02792	0.00351	-0.08024	-0.02356	-0.00498	-0.05979
Group 3	0.03086	0.03374	0.04593	0.03324	0.14170	-0.02298	0.06177	0.12938
Group 4	0.03621	0.03512	0.03110	0.03388	-0.02920	-0.02296	-0.02108	-0.03545
Group 5	-0.00799	-0.02573	-0.01853	-0.01948	-0.02888	-0.02603	-0.05451	-0.02246
Group 6	0.01563	-0.00192	-0.00221	0.00591	-0.12443	-0.02422	-0.10912	-0.11747
Group 7	0.07959	0.05704	0.06502	0.05760	0.10786	-0.02313	0.04065	0.09389
Group 8	0.00950	-0.03588	-0.05696	-0.04065	0.03772	-0.02703	-0.01268	0.01398
Mean Abs	0.04278	0.03394	0.03649	0.03523	0.07877	0.02519	0.04176	0.06577

Note. The difference was the value of the estimated item statistic minus the population item statistic. The population p-value is .50161 and the population point biserial correlation is .39877.