A standardization approach to adjusting pretest item statistics

Shun-Wen Chang
National Taiwan Normal University

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

Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, April 2000.

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 presented and evaluated a method of standardization that may be used by test practitioners to standardize classical item statistics when sample sizes are small. The effectiveness of this standardization approach was compared with the 1PL and 3PL models based on the criteria of the Pearson product-moment correlation, the MSE, variance and squared bias.

In light of estimating the item difficulty values, the differences of the performance between the 3PL and standardization methods were small, but the differences between the 1PL and these two methods were large. For the estimation of point biserial correlations, the 3PL model seemed to perform better than the standardization method, and the standardization method performed better than the 1PL model. Although the standardization method did not outperform the 3PL model for the design considered in this study, it could be promising when smaller sample sizes are used. This method may be recommended for use in conjunction with the IRT models for the test development when the pretesting sample sizes are small. By employing the classical measurement framework to obtain pretest item statistics, the problem of inaccurate IRT parameter estimates when limited calibration sample sizes are available can be avoided.

Acknowledgement

The authors wish to thank Qing Yi for preparing the population item parameters and Bor-Yaun Twu for programming for part of the data simulation.

This study is designed to compare the effectiveness of a standardization method with IRT methods for scaling pretest item statistics to be on the same scale. When item responses are obtained from different groups of small sample sizes, employing traditional IRT item calibration or scaling may not be justified due to the large sample size requirement. This study is intended to explore the method of standardization in adjusting the item statistics obtained from various groups in pretesting. Specifically, this study attempts to achieve the following objectives:

1. to examine the effectiveness of a standardization method;
2. to compare the item statistics recovery of the 1PL model, the 3PL model and the standardization method with small sample sizes.

Theoretical Framework

The problem arises 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. To maintain test security each item should be seen by the smallest number of examinees possible while still obtaining good item statistics. To achieve this goal, the pretest items can be included in parallel forms and be administered to different small groups of examinees. It is not easy to use randomly equivalent groups, however, since the groups used for pretesting are often conveniently formed based on school and concerns about item exposure preclude the administration of different forms to the same school using a spiraling process.

Based on specific groups of the 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 different groups are not directly comparable, posing a problem in the test development. 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.

To achieve adequate precision in the item parameter estimates obtained using IRT, the number of examinees used for calibration is required to be moderately large (Hambleton, Jones, & Rogers, 1993; Tsutakawa & Johnson, 1990). The requirement of large sample sizes in practical pretesting situations can be hard to meet. Depending on the specific testing situation (e.g., the number and nature of the items on the test, the distribution of the examinees' abilities) and the

particular IRT model chosen, the minimum number of examinees recommended for accurate item parameter estimation varies (Barnes & Wise, 1991). In general, greater numbers of items and more complex IRT models require larger sample sizes.

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. Because there is only one parameter to be estimated, this model does not require large sample sizes. Previous research suggested that a sample size of as large 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 suggested that at least 1,000 (Reckase, 1979; Skaggs & Lissitz, 1986) to 10,000 (Thissen & Wainer, 1982) examinees would be needed. Estimating IRT item discriminating parameters requires larger sample sizes than estimating item difficulty parameters (Barnes & Wise, 1991).

As an alternative to IRT item calibration with small sample sizes, this study proposes a standardization approach to adjust conventional item statistics which may perform better than IRT methods with small sample sizes. The purpose of using the standardization method is to adjust the item statistics obtained from small nonequivalent samples to more closely represent the item statistics in the population of interest. The idea is similar to that of the *direct standardization* described in Mosteller and Tukey (1977). This method is 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 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 common item score in the total group of examinees can then be obtained. Specifically, the standardization method is described as follows.

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 a subpopulation of examinees, g. The joint distribution of the unique item and common item scores in the subpopulation g can be formulated as

$$\Pr(U_g=u, X_{cg}=x) = \Pr(U_g=u|X_{cg}=x)\Pr(X_{cg}=x),\ g = 1, \ldots,G;\ u = 0, 1;\ x = 0, 1,\ldots, m,$$
$$= 0, \text{ elsewhere.}$$

The notation of $U_g = u$ represents the random variable U_g 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 $X_{cg} = x$ represents the random variable X_{cg} being equal to number correct score x in the subpopulation g, with the values of 0 to the number of common items, m. $\Pr(U_g=u|X_{cg}=x)$ is the conditional distribution of the unique item score given common item score and $\Pr(X_{cg}=x)$ 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

$$\Pr(U_o=u, X_{co}=x) = \Pr(U_o=u|X_{co}=x)\Pr(X_{co}=x),\ u = 0, 1;\ x = 0, 1,\ldots, m,$$
$$= 0, \text{ elsewhere.}$$

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

$$\Pr(U_o=u, X_{co}=x) = \Pr(U_g=u|X_{cg}=x)\Pr(X_{co}=x) \text{ for any subpopulation, g.}$$

where $\Pr(X_{co}=x)$, the distribution of the common item scores for the entire population o, is simply obtained based on the responses of all groups to the set of common items. Using this joint distribution of the 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 $\Pr(U_o=u, X_{co}=x)$ 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 be obtained from this joint distribution of the unique item and common item scores in the population of examinees also.

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).

Method and Data

Simulations were employed to carry out this study. Described below are the data, the simulation procedure, and the criteria used for the data analyses.

The Test

Ten forms were created from ACT Assessment (ACT, 1997) Mathematics items to be as parallel as possible in their content and statistical specifications. Each form consisted of 24 unique items and 12 common items. In each form the common items followed the unique items. Three-parameter logistic model item parameters were calibrated from multiple forms of the ACT Assessment Mathematics Test using BILOG (Mislevy & Bock, 1990). The data used for the calibration were randomly equivalent groups of examinees taking each form of the ACT Assessment, so all item parameter estimates were on a common scale. These calibrated item parameters were treated as the population item parameters in this study.

Table 1 presents the summary statistics of the item parameters a, b and c of the various forms in representing the item discrimination parameter, the item difficulty parameter and the guessing parameter, respectively. The N column lists the number of common or unique items.

The Samples and Population

Ten groups of examinees were generated, with a sample size of 250 examinees in each group to represent a situation where the group size was small. The ability distribution of the examinees in each group was generated based on a normal distribution. The current study was intended to simulate a situation where the various groups vary in their ability averages since randomly equivalent groups cannot be easily obtained using spiraling in real pretesting situations. Meanwhile, the ability mean differences that are expected among the groups could be small because efforts can still be made to mitigate the differences during sampling. In this study, the variances were set to be 1.0 for all groups, and the means were specified to be -0.4, -0.3, …, 0.5 for groups 1 through 10, respectively. That is, the ability of the various groups was distributed as N(-0.4, 1), N(-0.3, 1),…, N(0.5, 1), respectively. The entire sample of all ten groups was the chosen population in this study. Thus, the population ability distribution was a mixture of the ten normal distributions. The density of the mixture distribution at any θ point was a weighted sum of the ten densities using weight .10 for each density.

Based on the ability distributions of the examinees in the groups and the population item parameters, random responses of the items were generated for the various groups. Every examinee in each of the ten groups responded to 24 unique items and the set of 12 common items.

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 is calculated using the item characteristic curve (ICC)

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

where a, b and c are parameters that characterize the item. For the 3PL model, all the three parameters are considered 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 population p-value of an item was derived by approximating this continuous distribution $fo(\theta)$ with a discrete distribution at the θ levels equally spaced over the interval of –5.0 to 5.0 with an increment of .10 (i.e., -5.0, -4.9,…, 5.0), totaling 101 points. That is, the integral was replaced by the sum over these discrete θ points, with $fo(\theta)$ being the discrete probabilities of the θ points.

Because 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. The purpose of this study was to produce the item statistics that were comparable across groups. Point biserial correlations were computed between the unique item scores and common item scores so that all items were correlated with a common variable, making the correlations more comparable across items. The population point biserial correlation was computed based on the joint distribution of the unique item and the common items in the overall population $\Pr(Uo=u, Xco=x)$. This joint distribution $\Pr(Uo=u, Xco=x)$ is given by

$$\Pr(Uo=u, Xco=x)$$
$$=\int \Pr(U=u, Xc|\theta)\, fo(\theta)\, d\theta$$
$$=\int \Pr(U=u|\theta)\, \Pr(Xc|\theta)\, fo(\theta)\, d\theta$$

where Pr(U=1|θ) was calculated using the ICC and Pr(U=0|θ) is 1–Pr(U=1|θ). Pr(Xc|θ), 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 Pr(Xc|θ) for a particular θ, one for each of the common item scores of 0, 1, …, 12.

To derive this joint distribution Pr(Uo=*u*, Xco=*x*), the continuous distribution *fo*(θ) was also approximated with a discrete distribution at the θ levels equally spaced over the interval of –5.0 and 5.0 with an increment of .10. That is, the integral was replaced by the sum over these discrete levels, with *fo*(θ) being the discrete probabilities of the θ points.

The correlation based on values of this bivariate distribution Pr(Uo=*u*, Xco=*x*) was the population point biserial correlation of interest in this study, as

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

where Pr(Uo=*u*, Xco=*x*) is the joint distribution of the unique item and common item scores and Pr(Uo=*u*) and Pr(Xco=*x*) are the marginal distributions of the unique item and common items, respectively.

The Estimated Item Statistics

The estimated item statistics (both the 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 using the nonequivalent groups equating. Then, these item parameters were converted to the conventional p-values and point biserial correlations using the following formulas:

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

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

where $\hat{\Pr}(U=1|\theta)$ is the probability of correctly answering an item conditional on θ, which was estimated based on the 1PL or 3PL model; *fo*(θ) is the ability distribution in the overall population;

and also, $\hat{\Pr}(Uo=u, Xco=x)$, $\hat{\Pr}(Uo=u)$ and $\hat{\Pr}(Xco=x)$ are the respective distributions that were estimated based on the 1PL or 3PL model.

Also based on the response data in each of the ten groups, the classical p-value and point biserial correlation were computed for each unique item. The standardization approach described above was used to estimate the conventional item statistics in the entire population. In this study, the joint distribution of the unique and common items $\Pr(Ug=u, Xcg=x)$ was smoothed using two bivariate polynomial log-linear models (Hanson, 1991; Rosenbaum & Thayer, 1987). The first model used degree 4 for the common item score, degree 1 for the unique item score, and a degree 1 interaction term. The second model used degree 5 for the common item score, degree 1 for the unique item score, and a degree 2 interaction term. The marginal distribution of the common items $\Pr(Xco=x)$ was smoothed using a univariate polynomial log-linear model in Kolen (1991) with polynomial degree 4. This smoothed marginal distribution was used with each of the two smoothed bivarate distributions in each group to produce two estimates of the p-value and point biserial correlation.

Replications

The above process of estimating the item statistics was replicated 500 times for the three methods, respectively.

The Criteria

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 are 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 $r = 1, 2, \ldots, 500$, $\hat{p}_{ir}$ 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 are 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 $r = 1, 2, \ldots, 500$, $\hat{\rho}_{ir}$ 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 are 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

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

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

Results

Described below are the results of the 1PL model, the 3PL model and the standardization method 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. For the standardization approach, the two bivariate smoothing methods produced similar results, so only the results for the model using a fourth degree polynomial for the common item score are presented.

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 2 for the various methods. It can be seen that for the p-values, both the 1PL and standardization methods recovered the population p-values to a similar degree. The average correlation for the 3PL model was only 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 2 shows that all methods performed less well in recovering the point biserial than the p-values. The employment of the 3PL model still resulted in the highest average correlation between the estimated and population item statistics. The standardization method performed slightly worse than the 3PL model in recovering these correlations. It can be seen that the performance of the 1PL model was relatively poor among the three methods.

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

Table 3 shows the summary statistics for p-value MSE, variance and squared bias over the 240 items. It can be seen that the average MSE value was slightly lower for the 3PL model than for the standardization method. However, relative to the error in the estimates due to estimation by simulation with the 500 replications, the difference between these two methods was small. Therefore, the results did not provide a clear indication as to which method was better.

The employment of the 1PL model led to the greatest average MSE over items (see Table 3). The difference between the 1PL model and either of the 3PL and standardization methods was large relative to the standard error. The results showed that the 3PL and standardization methods produced smaller overall error for estimating the p-values than the 1PL method.

With respect to the variance, the average value over items was lower for the standardization method than for the other two methods. The average squared bias for the 3PL model was relatively lower than for both the 1PL and standardization models. That the squared bias was lowest for the 3PL model was not unexpected in that the data were generated with the 3PL model. The average values of the squared bias were similar for the 1PL and standardization approaches.

Displayed in Table 4 are the summary statistics of the MSE, variance and squared bias over the 240 items for the point biserial correlations. There seems to be more differences among the methods with regard to these correlations than the p-values. The average MSE value was the lowest for the 3PL model. The standardization method produced a higher average MSE over the 240 items than the 3PL model. The average MSE was even higher for the 1PL model. The differences among the methods in average MSE were large relative to the standard errors of the average MSE. These findings seem to suggest that the 3PL model performed better overall than the standardization approach, and the standardization approach performed better overall than the 1PL model in terms of estimating the point biserial correlations.

While the average variance over items for the standardization method was the lowest for the p-values (see Table 3), the average variance for the standardization method was the highest for the point biserial correlations in Table 4. The 1PL model had the lowest 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. The squared bias was slightly lower for the 3PL model than the standardization method, which was again, not unexpected since the population item parameters were generated with the 3PL model.

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 population of interest has been employed. 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 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 in light of estimating the p-values, the differences of the performance between the 3PL and standardization methods were small relative to the standard error, but the differences between the 1PL and the other two methods were large relative to the standard error. For the estimation of the population point biserial correlations, the 3PL model seemed to perform better than the standardization method, and the standardization method performed better than the 1PL model.

The standardization method proposed in this study failed to outperform the 3PL model in recovering the population point biserial, but the data were generated using a 3PL model. The relative performance of the standardization method and the 3PL model may differ when the data do not perfectly fit a 3PL model. Also, the ten forms of the test were created using item parameters calibrated based on the operational ACT Assessment Mathematics items of high quality. This could impact the results of this study since variation in the item statistics of prestest items could be greater. In addition, it remains unknown how robust the 3PL model would be to even smaller sample sizes. 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 IRT methods may not be justified and the performance of the 3PL model could be deteriorated. Further studies using smaller sample sizes could investigate this issue.

Also, in this study each form consisted of 24 unique items and 12 common items. The results of this study may not generalize to situations where the number of pretest or common items differs from the values used in this study. Results for which of the methods would perform better can not be simply implied based on the findings of this study. Moreover, while group differences affect the adjustment of the standardization method, it would also affect estimation with the IRT models. The issue of the degree to which the methods are affected by group differences needs more consideration. It is also worthwhile to implement other IRT models such as the 2PL model or a modified model and compare its effectiveness with the standardization approach.

The standardization procedure was carried out based on the assumption that the conditional distributions of unique or noncommon items given common items are the same across all groups of examinees. Factors such as the number and location of common items imbeded, the

representativeness of common items of the entire test, and the item characteristics of these common items might affect the viability of this assumption. It is not known the extent to which the assumption held in the present study.

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 presented and evaluated a method that may be used by test practitioners to standardize classical item statistics when sample sizes are small. Although the standardization method performed slightly less well than the 3PL model for the design considered in the current study, this method of standardization could be promising when smaller sample sizes are used. This method may be recommended for use in combination with the IRT models for the test development when the pretesting sample sizes are small. By employing the classical measurement framework to obtain pretest item statistics, the problem of inaccurate IRT parameter estimates when limited calibration sample sizes are available can be avoided.

References

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

Barnes, L. B., & Wise, S. L. (1991). The utility of a modified one-parameter IRT model with small sample sizes. *Applied Measurement in Education, 4*, 143-157.

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.

Hambleton, R. K., Jones, R. W., & Rogers, H. J. (1993). Influence of item parameter estimation errors in test development. *Journal of Educational Measurement, 30*(2), 143-155.

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.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models*. Chicago, IL: Scientific Software, Inc.

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression: a second course in statistics*. Reading, MA: Addison.

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.

Tsutakawa, R. K., & Johnson, J. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika, 55*(2), 371-390.

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. Summary Statistics of the Population Item Parameters for the Common Items and the Unique Items for the Various Forms

The Common Items

Item Parameters	N	Mean	SD	Minimum	Maximum
a	12	0.9349	0.2658	0.5360	1.3540
b	12	0.0683	0.6553	-0.6980	1.3020
c	12	0.1698	0.0548	0.0960	0.3000

The Unique Items

Form 1

Item Parameters	N	Mean	SD	Minimum	Maximum
a	24	1.0079	0.2837	0.4960	1.6640
b	24	-0.0754	0.9047	-1.8650	1.7440
c	24	0.2005	0.0901	0.0570	0.4390

Form 2

Item Parameters	N	Mean	SD	Minimum	Maximum
a	24	1.1303	0.3664	0.6140	1.8760
b	24	-0.0787	1.1617	-2.5220	1.5140
c	24	0.1805	0.0831	0.0570	0.4750

Form 3

Item Parameters	N	Mean	SD	Minimum	Maximum
a	24	0.9085	0.3338	0.4200	1.7150
b	24	-0.0849	1.2512	-2.7130	2.0650
c	24	0.1651	0.0626	0.0430	0.3100

Form 4

Item Parameters	N	Mean	SD	Minimum	Maximum
a	24	1.0865	0.4130	0.3270	2.1230
b	24	0.0100	1.3322	-3.7180	1.7940
c	24	0.1622	0.0622	0.0410	0.2640

Form 5

Item Parameters	N	Mean	SD	Minimum	Maximum
a	24	1.2420	0.3410	0.5390	1.9630
b	24	-0.0134	0.8635	-2.3470	1.6220
c	24	0.1775	0.0635	0.0710	0.3050

Table 1. (Continued)

Form 6

Item Parameters	N	Mean	SD	Minimum	Maximum
a	24	1.1430	0.3739	0.6320	2.1210
b	24	-0.0845	0.8920	-2.2830	1.5000
c	24	0.1588	0.0563	0.0710	0.2640

Form 7

Item Parameters	N	Mean	SD	Minimum	Maximum
a	24	1.0472	0.3817	0.4960	2.3790
b	24	-0.0241	0.8904	-1.6660	1.7780
c	24	0.1738	0.0840	0.0320	0.4320

Form 8

Item Parameters	N	Mean	SD	Minimum	Maximum
a	24	0.9976	0.3075	0.5230	1.6130
b	24	-0.2165	1.1063	-2.2650	1.6680
c	24	0.1397	0.0530	0.0640	0.2590

Form 9

Item Parameters	N	Mean	SD	Minimum	Maximum
a	24	1.2295	0.3962	0.6160	2.1010
b	24	-0.0744	1.2673	-3.4590	1.3490
c	24	0.1681	0.0760	0.0500	0.3430

Form 10

Item Parameters	N	Mean	SD	Minimum	Maximum
a	24	1.0041	0.3205	0.6150	1.7140
b	24	0.0459	1.2908	-2.2920	2.4940
c	24	0.1744	0.0843	0.0830	0.4910

Table 2. Summary Statistics of the Correlations between the Estimated and Population Item Statistics

p-Values

Method	N	Mean	SD	Minimum	Maximum
1PL	500	0.98990	0.00105	0.98657	0.99236
3PL	500	0.99121	0.00094	0.98802	0.99386
Standardization	500	0.98899	0.00132	0.98440	0.99254

Point Biserial Correlations

Method	N	Mean	SD	Minimum	Maximum
1PL	500	0.61584	0.01212	0.57734	0.64596
3PL	500	0.87582	0.01475	0.80593	0.91449
Standardization	500	0.82997	0.01739	0.77128	0.88271

Table 3. Summary Statistics for the Estimated p-value MSE, Variance and Squared Bias for the Various Methods

MSE

Method	N	Mean	SD	Minimum	Maximum	Standard Error
1PL	240	0.00108	0.00049	0.00000	0.00063	0.00001487143
3PL	240	0.00095	0.00025	0.00000	0.00071	0.00001436668
Standardization	240	0.00096	0.00038	0.00000	0.00125	0.00000525862

Variance

Method	N	Mean	SD	Minimum	Maximum
1PL	240	0.00085	0.00025	0.00015	0.00132
3PL	240	0.00091	0.00023	0.00017	0.00132
Standardization	240	0.00073	0.00019	0.00012	0.00110

Squared Bias

Method	N	Mean	SD	Minimum	Maximum
1PL	240	0.00023	0.00035	0.00000	0.00212
3PL	240	0.00004	0.00006	0.00000	0.00032
Standardization	240	0.00023	0.00027	0.00000	0.00133

Table 4. Summary Statistics for the Estimated Point Biserial MSE, Variance and Squared Bias for the Various Methods

MSE

Method	N	Mean	SD	Minimum	Maximum	Standard Error
1PL	240	0.00649	0.00968	0.00000	0.00308	0.0000609573
3PL	240	0.00253	0.00093	0.00000	0.00242	0.0000343220
Standardization	240	0.00378	0.00169	0.00000	0.00335	0.0000187347

Variance

Method	N	Mean	SD	Minimum	Maximum
1PL	240	0.00048	0.00014	0.00038	0.00134
3PL	240	0.00202	0.00054	0.00104	0.00457
Standardization	240	0.00311	0.00076	0.00165	0.00614

Squared Bias

Method	N	Mean	SD	Minimum	Maximum
1PL	240	0.00602	0.00968	0.00000	0.07004
3PL	240	0.00052	0.00076	0.00000	0.00703
Standardization	240	0.00067	0.00131	0.00000	0.01005