GMAC®

Graduate
Management
Admission
Council®
Creating Access to Graduate Business Education®

Scaling Item Difficulty Estimates from Nonequivalent Groups

Fanmin Guo, Lawrence Rudner, and Eileen Talento-Miller

GMAC® Research Reports • RR-09-03 • April 3, 2009

Abstract

By placing item statistics on a common scale, items piloted with groups of test takers who have different ability levels can be combined to yield a test with known characteristics. This project examined two approaches using the classical testing theory and one approach using the Rasch model for scaling item difficulty estimates. A simulation study was conducted to compare the true item difficulties with the scaled difficulties. While bias and error diminished as sample sizes increased, all three approaches were found to be extremely accurate at all tested sample size values. There are slight differences in the approaches in terms of sensitivity to variations on the test-taker ability distributions.

When new items are piloted on separate forms to determine their adequacy for operational use, it is necessary to ensure the item statistics from each form are comparable. This is not a problem if a random equivalent group design is used in the data collection. In that situation, the groups taking the separate forms are equivalent, each representing the test population; therefore, the statistics from different forms are comparable. When random equivalent groups cannot be assured, however, researchers switch to a *nonequivalent groups anchor test* (NEAT) design in which a set of common items is administered to all groups. These common items will provide the statistical adjustments necessary to scale item statistics from each form to a common scale so that they will be comparable.

In this study, we present and evaluate three methods of scaling the difficulty estimates from nonequivalent groups using simulated data.

Two methods are based on classical testing theory; the third is based on *item response theory* (IRT).

The Problem

The initiation of this study coincided with the development of a new test. The challenge was to pre-test 100 new items with three groups of test takers of presumably nonequivalent ability; we allowed each group to take only 40 items. With the NEAT design, 10 items would be selected as anchors and the remaining 90 items would be assigned to three nonoverlapping forms of 30 items each. We chose this design so that impact of ability differences on item difficulty estimates could be statistically removed. The question of how to remove those differences remained, however.

In this study, each group of test takers acquires only one of the three forms of new test items; random assignment of test takers to forms is not feasible. Nothing is known a priori about the items or the examinees' abilities. We randomly selected a common set of anchor items and administered them to all test takers to enable comparison of ability among the groups taking the forms. The remaining items are divided evenly, creating unique sets of items administered to each group, denoted as Form 1, Form 2, or Form 3. All items are dichotomously scored. Table 1 displays the data collection design.

Table 1. Data Collection Design

	Anchor	Form 1	Form 2	Form 3
Group 1	x	x		
Group 2	x		x	
Group 3	x			x

The question, then, is how to use data collected from the groups' performance on the anchor items most effectively to ensure all item statistics are on a common scale. Previous research led to the evaluation of three different methods for this scaling. We conducted a simulation study to determine which of the methods would yield scaled values closest to the true values.

Related Literature

The NEAT design is commonly used in practice for equating test scores across forms. Holland and Dorans (2006) describe several methods for equating scores on forms using a NEAT design. Very little literature is available, however, on how to scale the item statistics with such a design.

Linear transformation method

Thurstone (1925) introduced a method for scaling tests that were designed for different age groups based on the difficulties of common items. Although Thurstone's purpose in that paper is different from our goal, the transformation of item difficulty estimates from one age (ability) group to another can be applied to our study.

In general, item difficulty, p, is defined as the proportion of correct answers. It needs to be converted to at least an interval scale before linear transformations apply. Lord (1980, pp. 213–215) tried to re-scale the p-values estimated from two subgroups to compare them for *differential item functioning* (DIF). He noticed the curvilinear relationship between the two sets of p- values. He suggested that all ps be converted to a z-scale through the inverse normal function before comparing them. Gullikson (1950, pp. 368–369) constructed the College Entrance Examination Board *delta*-scale, which was a linear transformation of the inverse normal deviates for similar purposes. In our study, we will first convert item difficulty estimates from a p-scale to inverse normal deviates, then make the adjustment with linear transformations based on the group differences of the examinee ability, and finally convert the adjusted (scaled) item difficulty estimates back to the p-scale. This is the first method evaluated for scaling item difficulties. Specific steps are outlined in the Methodology section.

Rasch method

When an IRT framework is used in test development, scaling item statistics is a common topic. By design, IRT item statistics are calibrated on an arbitrary scale. In order to put item statistics on an intended scale, either scaling is built into the calibration process or a post hoc scaling is applied to transform the item statistics from an arbitrary scale to an intended scale. Many researchers (Hanson & Beguin, 1999; Kim & Cohen, 1998; and Peterson, Cook, & Stocking, 1983) have evaluated different methods. For the data collection design in our study, we chose the concurrent calibration method. This is the second method evaluated for scaling the item difficulty estimates. We will first calibrate the item statistics and then scale them to our target scale.

Standardized method

In studying DIF, Dorans and Kulick (1983) developed a method called "standardization method." They used this method to scale item difficulty estimates from subgroups to a common scale for comparison. Dorans, Schmitt, and Bleistein (1988) also applied it to assessing differential speediness. In some testing organizations, the standardization method is also applied to scaling item difficulty estimates (Dorans, 2007).

The scaled *p* from this method is the weighted sum of conditional difficulty estimates. The standardization method is especially powerful when there is an existing ability score (such as GMAT® scores), which may replace the anchors when new items are pre-tested with the same test population. The weights are often calculated with the whole population, so the error in defining the population weights is eliminated with this method.

This is our third method in the current study. We will explain it in the next section.

Methodology

We designed and implemented a simulation study to evaluate the performance of the three methods for the NEAT design. In this section, we describe the methodologies used for the simulation and for the three scaling methods.

Simulations

As discussed previously, the groups of the test takers might not have the same ability. One group might be of higher ability than the other two. The first factor we considered in the simulation study was the ability of the groups. The θ for Group 1 had a normal ability distribution with a mean of -0.3; the mean θ for Group 2 was 0; and the mean θ for Group 3 was 0.5. The standard deviation of the θ for each group was 1.

We expect that item difficulty estimates become more stable with larger sample size; therefore, sample size might also interact with the performance of the three scaling methods. In this study, we included three sample sizes of 250, 500, and 1,000 for the difficulty estimations.

To evaluate the efficacy of the three methods and their interactions with test takers' ability and sample size, we built three ability groups and three sample sizes in the simulation design. Table 2 presents the design.

Table 2. Simulation Design

	Sample Size		
Ability Group	**250**	**500**	**1,000**
Group 1 (N~(-0.3, 1))	Anchor + Form 1	Anchor + Form 1	Anchor + Form 1
Group 2 (N~(0, 1))	Anchor + Form 2	Anchor + Form 2	Anchor + Form 2
Group 3 (N~(0.5, 1))	Anchor + Form 3	Anchor + Form 3	Anchor + Form 3

The simulated responses to the items were generated with the three-parameter-logistic (3PL) IRT models. First, 100 items were sampled such that their parameters had distributions of b ~ N(0, 1), a ~ LogN(1, 0.32), and c = .25. These items were randomly divided into four independent sets with 10, 30, 30, and 30 items. The 10-item set became the anchor set and the three 30-item sets became Forms 1, 2, and 3. The means and standard deviations of the *a*- and *b*-parameters of the item sets are displayed in Table 3. The *c*-parameter for all items was fixed at 0.25, so they are excluded from Table 3.

Table 3. Means (Standard Deviations) of the *a*- and *b*-Parameters

	No. of Items	a	b
Anchor Set	10	0.76 (0.18)	0.48 (0.63)
Form 1	30	0.80 (0.22)	-0.04 (0.83)
Form 2	30	0.75 (0.19)	-0.02 (1.09)
Form 3	30	0.75 (0.21)	-0.18 (1.04)
All items	100	0.76 (0.21)	-0.03 (1.00)

Note that the mean *b*-parameter for the anchor set is higher than those of the forms. We decided not to re-sample the anchor items to match the mean *b*-parameters of the forms. These parameters in Table 3 likely represents a common scenario in a new testing program whenever anchor items have to be selected without known statistics.

Responses were generated for each of the conditions (the cells in Table 2). The generated responses provided the basis for evaluating the scaling methods. A total of 30 data sets (29 replications) were generated and used in this study. From each data set of responses, we drew random samples from each group corresponding to the different sample sizes and analyzed them three times—one for each scaling method.

Methods of scaling item difficulty estimates

Since the three groups are not equivalent in ability, nonanchor items will not have comparable difficulty estimates. To compare and use them for future assembly of test forms, item difficulty estimates from each form need to be scaled to a common scale. In this study, we defined the target scale as the difficulty estimates that are calculated with responses on the anchor items of all three groups combined. To facilitate the discussion, we use the following notations:

- p = true item difficulty
- $p*$ = raw or observed item difficulty
- p' = the item difficulty on the target scale, or the scaled item difficulty

The asterisk (*) and prime (') will also be used with other interim statistics, denoting "raw" or "scaled" statistics. The implementation of the three scaling methods is discussed below.

Linear transformation method

We first converted item difficulty estimates from the *p*- to a z-scale through the inverse normal function and then made adjustment with linear transformations based on the group differences of the examinee ability. Finally, we converted the adjusted (scaled) item difficulty estimates back to the *p*-scale.

We followed these steps in implementing the linear transformation method:

1. Calculate the $p*$ with the data from each form for both anchor and nonanchor items
2. Calculate the p' of anchor items with data from all three groups
3. Convert the observed $p*$ of all items to z-scale of normal deviates[1] to derive z_{p*} such that

$$\phi(z_{p*}) = p*$$

4. Convert the p' of the anchor items to z-scale so that

$$\phi(z_{p'}) = p'$$

5. Estimate the scaling parameters using the z_{p*} and $z_{p'}$ of the anchors as

$$A = \frac{\sigma_{z_{p'}}}{\sigma_{z_{p*}}} \text{ and}$$

$$B = \mu(z_{p'}) - A\,\mu(z_{p*})$$

6. Apply the scaling to the nonanchor items by form to derive the $z_{p'}$ of nonanchor items. The linear transformation is

$$z_{\mathrm{p}'} = A(z_{p*}) + B$$

7. Convert all the $z_{p'}$ back to p' such that

$$p' = \phi(z_{\mathrm{p}'})$$

Rasch method

Although 3PL IRT models were used for generating the response data for this study, we chose the Rasch model for the scaling method. Our focus was on the item difficulty and we found that the Rasch model sufficed for this purpose. A concurrent calibration was first run for calibrating the item parameters using a variant of PARAM-3PL (Rudner, 2005). The input data included a single sparse matrix of responses of all

[1] Items difficulty p has a very counterintuitive interpretation. Items with larger p values are easier than items with smaller p-values. In both Lords's z-scale and Gullikson's delta scale, $z_{(1-p)}$ is used for an item with p rather than $z_{(p)}$. Therefore, their zs or deltas have larger values for more difficult items. In our study, we simply used $z_{(p)}$ in the z-conversion. It will not change our results and we will convert the item difficulty back to p after the scaling.

test takers to all items. Responses for items that were not administered to some groups were coded as missing. We fixed the *a*-parameters at 1 and all *c*-parameters at .25. The fit of estimated item models to the data was reasonable.

All the resultant item measures (*b**-parameters) were then on the same scale from the concurrent calibration. Like any other IRT parameter calibrations, however, the θ-scale underlying the item *b**-parameters was an arbitrary scale because no scaling process had been built into the calibration process. Two scaling methods were available to make sure that the item or person parameters were on a specified scale. Both methods require one of two things: (1) anchor items with known item parameters should be administered with the new items, or (2) examinees with known θs must respond to the new items. One method was to use either fixed item parameter or fixed θ calibration. By fixing the anchor items to their known parameter values or fixing the examinees' ability to their known θs in the calibration, all the resultant item parameters and θs should be on the scale represented by the known item parameters or θs. The other method was to first calibrate the item or person parameters on an arbitrary scale and then transform them linearly to the target scale using the known item parameters or θs. In this study, we used the second method for scaling the item parameters.

The following is the procedure we used in our study:

1. Estimate the item measures (*b**-parameters) from a concurrent calibration run
2. Calculate the *p'* of the anchors as the proportions of correct answers of all three groups of simulees
3. Convert all the *p'* of the anchors to Rasch *b*-parameters as

$$b^{'} = \ln(\frac{1-p}{p})$$

4. Estimate the transformation parameters using the two sets of *b* of the anchor items (*b'* from the responses and *b** from the calibration) as

$$A = \frac{\sigma_{b'}}{\sigma_{b*}} \text{ and}$$

$$B = \mu(b') - A\,\mu(b*)$$

5. Apply the transformation to all the *b** of the nonanchor items

$$b' = A(b*) + B$$

6. Convert resultant *b'* to scaled difficulty estimates *p'* using

$$p' = \frac{1}{1+e^{b'}}$$

Standardization method

The standardization method in this study is defined as

$$p' = \sum_k w_k \times p_k^* ,$$

where *k* is the ability groups based on the performance on anchor items, w_k is the proportion of test takers in each ability group, and $p*_k$ is the proportion of correct answers to a nonanchor item conditional on *k*. We implemented the method as follows:

1. Calculate a number right total score on anchor items for each simulee
2. Sort simulees into *k*-ability groups
3. Calculate the conditional weights w_k such that

$$\sum_k w_k = 1$$

4. Calculate the conditional *p** of a nonanchor item
5. Calculate the *p'* for the item as the weighted sum of conditional *p**
6. Repeat 4 and 5 for all other nonanchor items

Analyses

For each of the 30 simulated data sets, we performed nine scalings, one for each method on each sample size. The criteria for evaluating the performance of the three scaling methods were the bias and error of the scaled item difficulty estimates (*p'*). Smaller errors for a method indicated better performance. The errors in the scaled item difficulty were defined as *p – p'*. Since we have the IRT item parameters and simulees' θs for generating the responses, we calculated the true item difficulty *p* for each item as the sum over all simulees of the probabilities of a correct response:

$$p = \frac{\sum_{e \subset s} P(\theta_e)}{N_{e \subset s}},$$

where *e* was simulees; $P(\theta_e)$ was the probability of a correct answer given the item's *a*, *b*, and *c* parameters, and the simulees' θ; *s* was the sample size levels. The sample size level was in the equation because we scaled the item difficulty estimates separately for each sample size level. To avoid scaling of item difficulty estimates across replications for comparisons, we choose to use the true *p* of each item calculated within each replication. We defined the target scale (*p'*) as the item difficulty estimates from all test takers who took any one of the three forms. Accordingly, the true item difficulty was calculated with all the simulees in all three groups for a sample size level within each replication. For example, for the 500 level, the true difficulty *p* of an item is the sum of the $P(\theta_e)$s of the 1,500 simulees.

We also selected the root mean square error as our measure of error and bias. They were calculated by method by ability group and by sample size for each replication. The root mean square error was defined as

$$RMSE = \sqrt{\frac{\sum_i (p - p')^2}{N_i}}$$

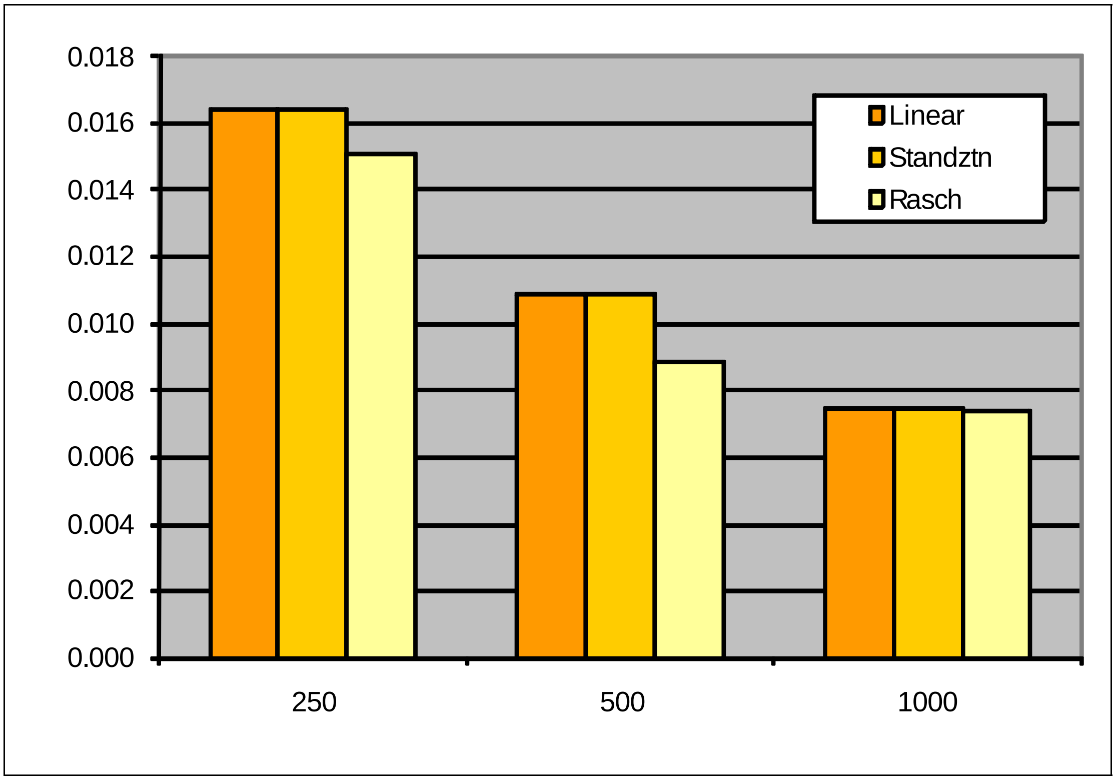
where the summation is over the *i* items.

Bias was calculated as the RMSE between the scaled *p*' and the true *p* using the anchor items; errors were calculated using the nonanchor items. The aggregated errors and biases across replications will be reported. The scaling methods with smaller mean error across replications will be considered better than those with larger mean errors. To get a sense of how big the errors were between the *p*' and *p* in the unit of *p*, we also estimated the average of the differences. The means of the average differences across replications will also be reported in next section.

Results

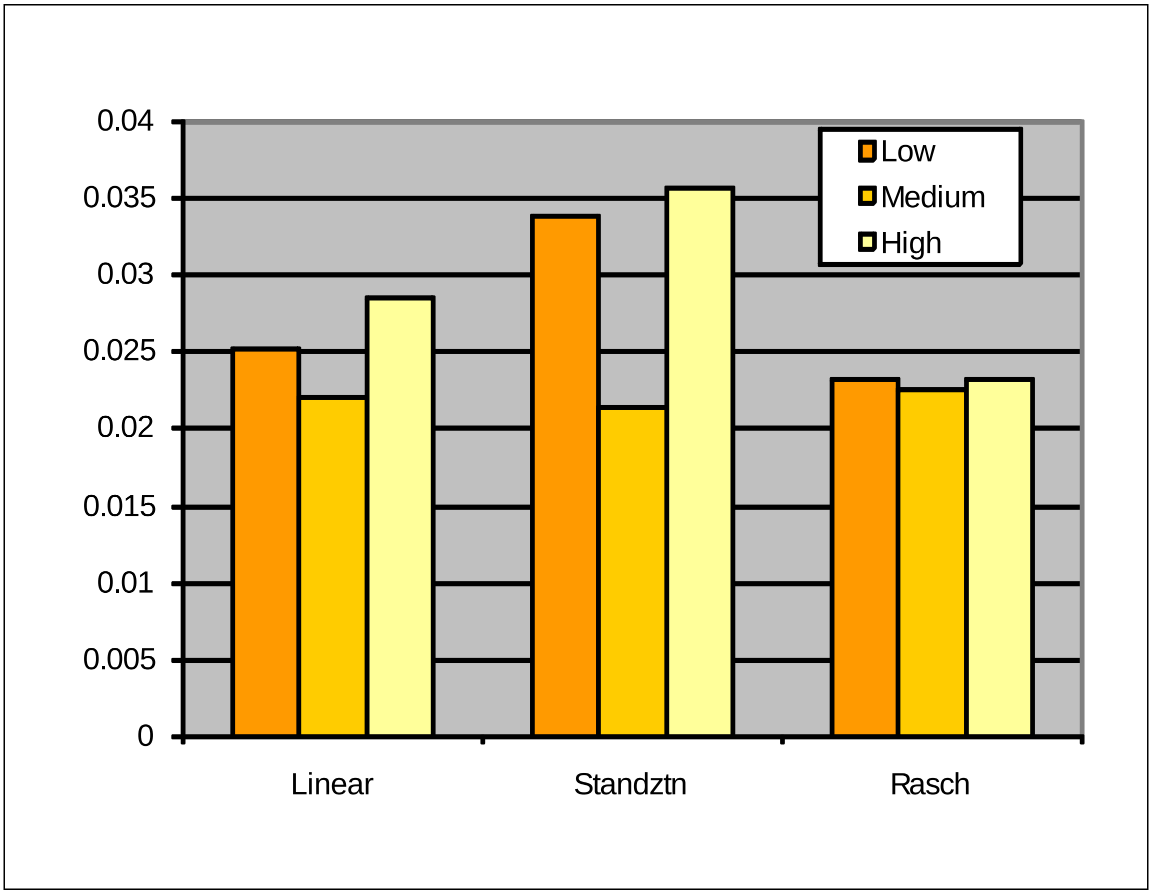
The magnitude of the mean bias was tiny for all methods and all sample sizes. Figure 1 presents the mean bias of the three methods by sample size. It is clear that as the sample size increases, bias decreases for all three methods. At sample sizes equal to 250 and 500, the linear transformation method and the standardization method seem to have similar bias and the Rasch method seems to produce smaller bias in comparison. At sample size of 1,000, however, the three methods show similar bias. Smaller bias indicates better recovery of the item difficulties in the simulated data.

Figure 1. Mean Bias by Sample Size



In the next sections, errors are reported for the three methods. After the item difficulties have been scaled with each method, smaller RMSE indicates better recovery of the true item difficulties by the scaled difficulty estimates. Figure 2 is the mean RMSE by ability groups. It is obvious that the item difficulties estimated from the medium-ability group show smaller errors than those of the low- and high- ability groups. This is consistent for all three methods. The ability distribution of the medium group is similar to that of the population. Since the target scale for the study was based on the three groups combined, the item difficulty estimates calculated from the medium ability group could be very close to their target scale values, thus lest affected by any scaling methods. For the Rasch method, the RMSE for the three ability groups are almost identical. It seems to have performed better than the other two methods for the low- and high-ability groups. The linear transformation method seems to have performed better than the standardization method for the low- and high-ability groups.

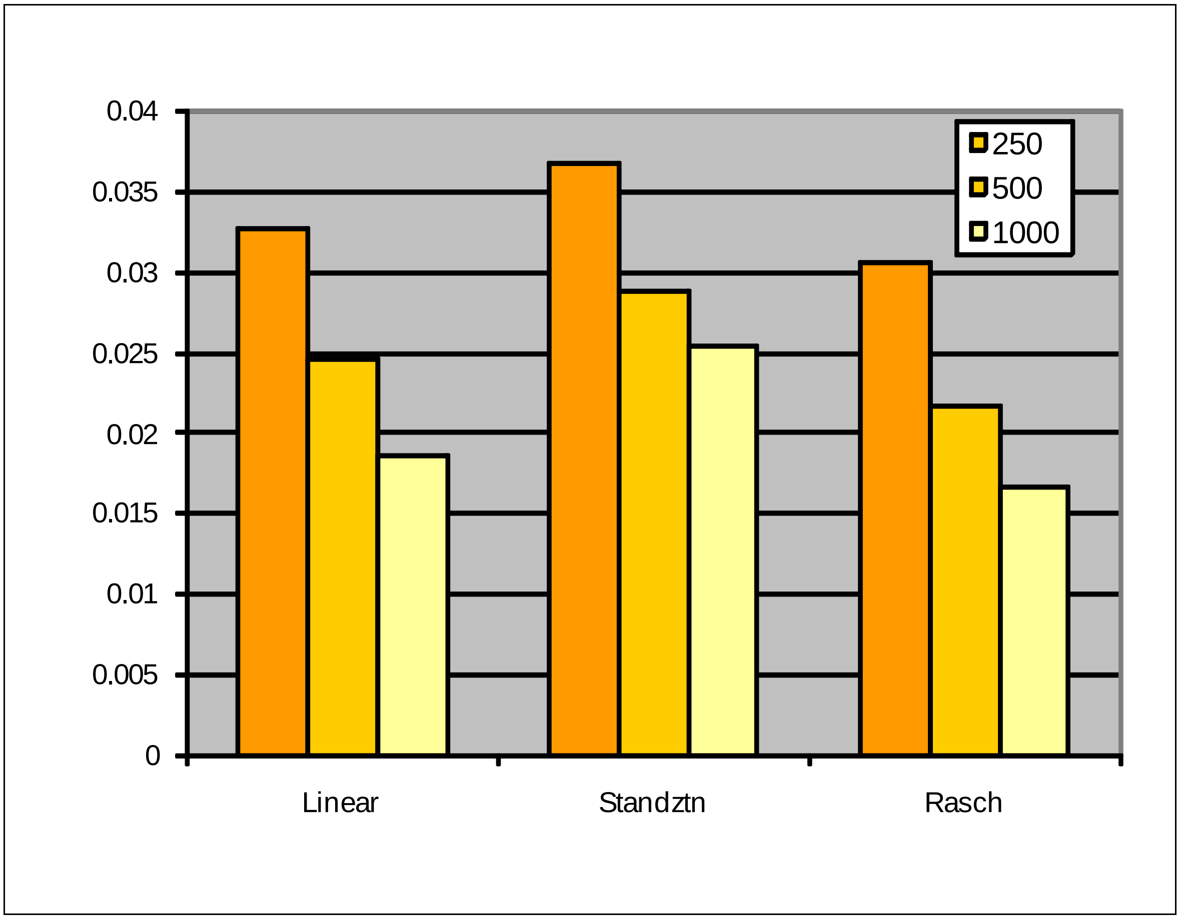
Figure 2. Mean RMSE by Ability



We may conclude that when the sample ability distribution is close to the population ability distribution, all three methods will give similar results. When the sample ability is different from that of the population, however, Rasch method works the best and the linear transformation method might work better than the standardization method.

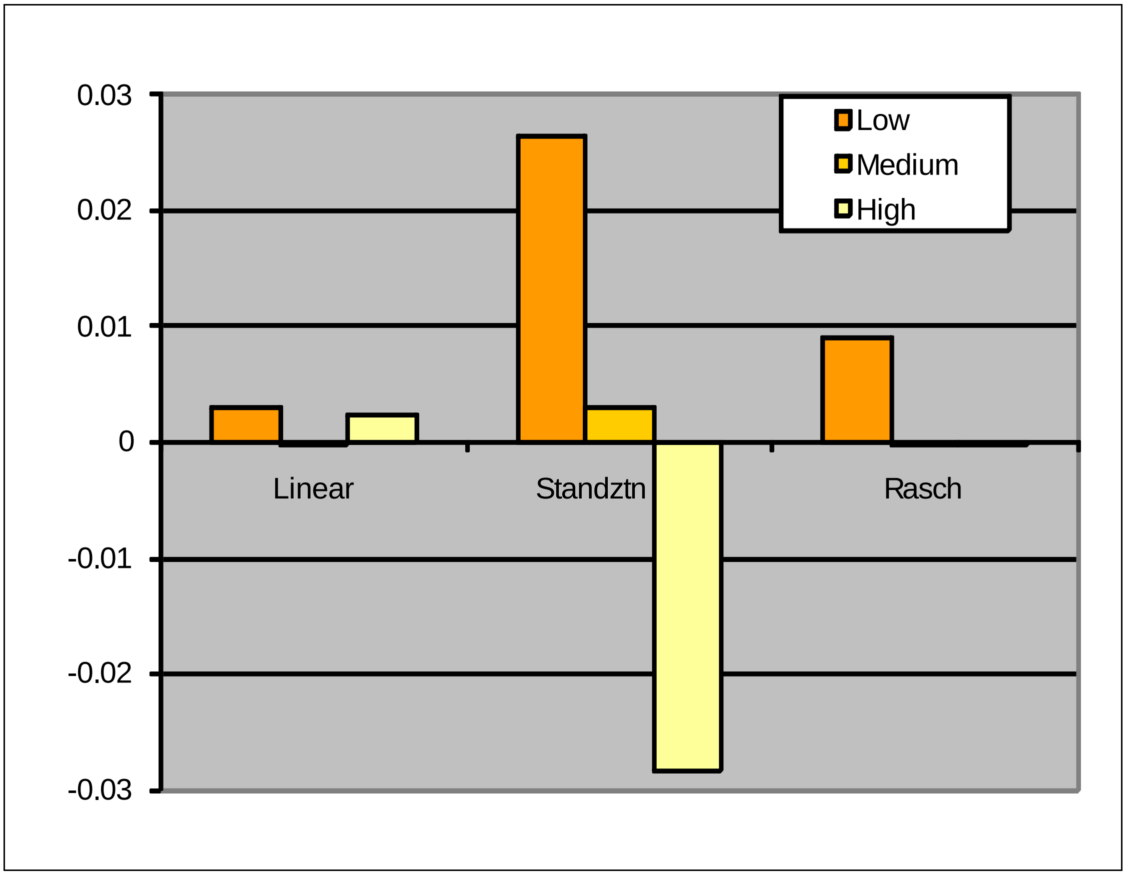
In Figure 3, the mean RMSE are given by sample size. As was the case with the bias, the errors decrease as sample sizes increase.

Figure 3. Mean RMSE by Sample Size



The above errors were calculated as the square roots of mean squared errors. It makes it easy to compare the methods; however, it loses the magnitude of the mean differences on the *p*-scale. Figure 4 shows the mean differences between the true and the scaled difficulties for each method by ability group (*p*–*p*'). While the mean differences are quite small for all methods, the standardization method shows the largest mean differences for the low- and high-ability groups, 0.025 and -0.028. The other two methods seem to have very small mean differences even for the low- and high-ability groups. We conclude that caution may be needed when using the standardization method with samples that are different in ability distributions from the intended population in comparison with the other two methods.

Figure 4. Mean *p*-Value Differences by Ability



Discussion and Conclusion

When item difficulties are placed on a common scale, questions piloted with different, nonequivalent groups can be combined to yield a test with known characteristics. Such scaling has traditionally been touted as a major advantage of item response theory and the Rasch model (see Wright and Stone, 1979). Although not prominent in the research literature, item *p*-value scaling is also possible within the classical measurement theory framework.

Two classical methods and one IRT method for scaling item difficulties estimated from nonequivalent groups based on a common anchor item design were evaluated by comparing the scaled item difficulties with the true item difficulties. One classical approach was based on a linear transformation of inverse normal deviates of the *p*-values. Another was based on weighting observed conditional *p*-values. Data were simulated such that three independent and nonequivalent groups of examinees took three nonoverlapping forms of a 30-item test along with a common set of 10 anchor items. Each of the three scaling approaches was applied to the data using 250, 500, and 1,000 independent simulees per form. Using the same set of generating item and ability parameters, the process was repeated an additional 29 times using different sets of simulees.

As measured by the magnitude of the bias and error and the magnitude of the mean difference between scaled and true item difficulties, all three approaches worked extremely well. The average differences in *p*-values were .03 or less. As expected the mean bias and RMSE decreases as sample size increases. At 1,000 test takers, the mean bias and error values for all methods were less than .008 and .025 respectively. At both 250 and 500, the Rasch method seems to perform slightly better than the linear transformation and the standardization methods. When the examinee groups show ability distributions different from those of the test population, the Rasch method outperformed the linear transformation method, which outperformed the weighted standardization method.

This research demonstrated that classical approaches can accomplish the same item difficulty scaling precision as the Rasch model approach. Because the errors and bias were all small, there are minimal practical differences between approaches. The sensitivity of the weighted standardization approach warrants further investigation, however.

Acknowledgements

We would like to express our thanks to Devina Oliver of GMAC® for running the simulations and analyses for the simulation study.

Contact Information

For questions or comments regarding study findings, methodology, or data, please contact the GMAC Research and Development department at research@gmac.com.

References

Dorans, N. (2007). Personal communications between Fanmin Guo and Neil Dorans.

Dorans, N., & Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1997: An application of the standardization approach*. (RR-83-9). Princeton, NJ: Educational Testing Service.

Dorans, N., Schmitt, A., & Bleistein, C. (1988). *The standardization approach to assessing differential speededness*. (RR-88-31). Princeton, NJ: Educational Testing Service.

Gullikson, H. (1950). *Theory of mental tests*. New York: Wiley.

Hanson, B., & Beguin, A. (1999). *Separate versus concurrent estimation of IRT item parameters in the common item equating design*. (ACT Research Report Series 99-8). Iowa City, Iowa: ACT.

Holland, P., & Dorans, N. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp.187–220). Westport, CT: American Council on Education and Praeger Publishers.

Kim. S., & Cohen, A. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement. 22,* 131–143.

Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillside, NJ: Lawrence Erlbaum.

Peterson, N., Cook, L., & Stocking, M. (1983). IRT versus conventional equating methods: a comparative study of scale stability. *Journal of Educational Statistics. 8*, 137–156.

Rudner, L. M. (2005). *PARAM-3PL Calibration Software for the 3 Parameter Logistic IRT Model (freeware)*. Available: http://edres.org/irt/param

Thurstone, L. L. (1925). A method for scaling psychological and educational tests. *The Journal of Educational Psychology.16*(7), 443–451.

Thurstone, L. L. (1938). Primary mental ability. *Psychometric Monographs*. No. 1.

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