Paper 3376-2015

SAS®-PIRT (SAS® macro) for Estimating Parameters of Polytomous Items Based on Graded Response Model (GRM)

Sung-Hyuck Lee, Hongwook Suh, ACT

ABSTRACT

Polytomous items have been widely used in educational and psychological settings, with increasing demand for statistical programs that estimate the parameters of polytomous items. Samejima (1969) proposed the graded response model (GRM) in which category characteristic curves are characterized by the difference of the two adjacent boundary characteristic curves. In this paper, it is demonstrated how the GRM parameter estimation program SAS®-PIRT (SAS® macro written in SAS® IML) performs in recovering the parameters of polytomous items using simulated data.

ITEM RESPONSE THEORY

In many psychological and educational settings where researchers find themselves in confronting unobservable entities called traits (or constructs), it is assumed that the traits of an examinee can be manifested by the responses of the examinee to the items that are constructed based on a theory to measure the traits. Therefore, what is needed for the researchers is a test theory that relates the performance of an examinee to the items measuring those traits. A test theory is a vehicle that helps researchers to quantify the trait so they can discriminate examinees on the trait continuum.

Unlike classical test theory in which test score depends exclusively on the trait, item response theory (IRT) provides various mathematical models that describe the interaction of item characteristics and the trait in determining the test score. In IRT models, the probability that an examinee responds correctly to an item depends on not only on a person's trait (ability or proficiency level) but also on various item parameters. The probability that an examinee provides a correct response to a particular item increases monotonically as ability increases. However, when the examinee ability is held constant, the probability of a correct response decreases monotonically as the difficulty of an item increases.

TWO-PARAMETER LOGISTIC MODEL FOR A DICHOTOMOUS ITEM

In the commonly used IRT two-parameter logistic model (2PLM), the probability that an examinee responds correctly response to an item is modeled as follows. We assume that the random variable X has only two binary outcomes (e.g., X= 1 for a correct response and X = 0 for an incorrect response), in which case the conditional probability of a correct response given θ is

$$P(X=1|\theta)=\frac{1}{1+\exp(-1.7a(\theta-b))}, \qquad (1)$$

Where,

- a is the discrimination parameter of an item,
- b is the difficulty parameter of an item,
- θ is the ability parameter of an examinee.

ITEM CHARACTERISTIC CURVE FOR A DICHOTOMOUS ITEM

The probability function determined by the interaction between examinees and an item is called the item characteristic curve (ICC). When the 2PLM is used, a dichotomous item can be described by the IRT-*a* and IRT-*b* parameters of an item as shown in Figure 1. Here, θ is on the horizontal axis and the ICC value is on the vertical axis. The difficulty parameter *b* is determined at the location where the probability of a correct response to the item is 0.5. The more difficult an item is the less likely the item is answered correctly at any given θ. The discrimination parameter *a* of an item describes how well an item differentiates examinees at the location on the θ scale where the difficulty parameter of the item is determined. The slope of the ICC is steepest at the item's difficulty location, which is the point of inflection of the item. The slope at the point of inflection is steeper for more discriminating items.

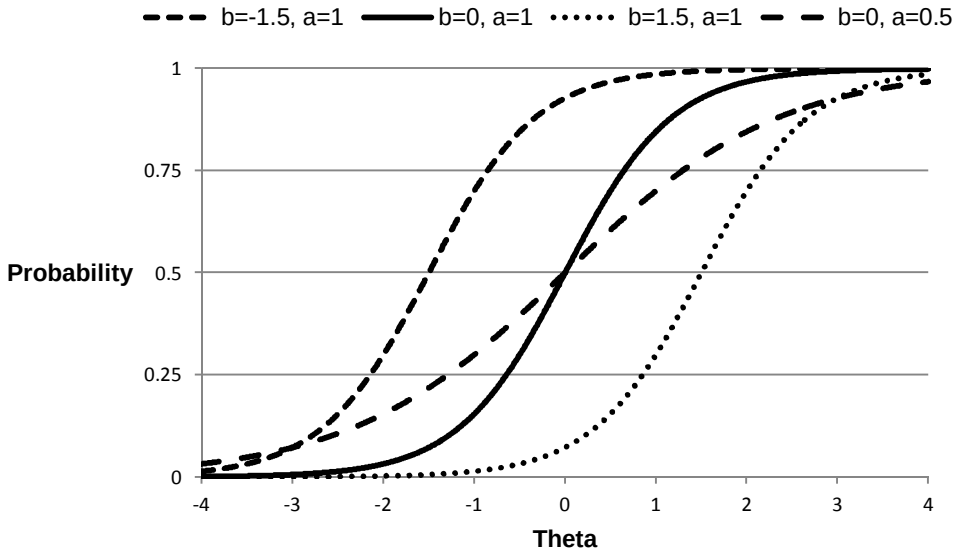

Figure 1. Various Item Characteristic Curves

POLYTOMOUS ITEMS

Polytomous items with more than 2 outcomes have been increasingly used for many practical reasons. First, the response of an examinee to a polytomous item provides more information about the trait measured than the response to a dichotomous item. Polytomous items are designed to indicate how much of the trait an examinee has rather than whether an examinee has the trait or not (e.g., all or nothing). Accordingly, polytomous items may allow researchers to make a more accurate inference about the trait measured than dichotomous items.

Second, a response of an examinee may deserve partial credit when the response is not completely correct but very close to the right answer. This is best reflected when the item score is the number of correct steps that an examinee must go through to get a correct response. For instance, an examinee might go through all the steps necessary to reach the right answer but the last one. In that case, scoring the item polytomously is preferred to scoring it dichotomously since it would allow raters to give a more informative score to an examinee's response compared to scoring the item dichotomously all right or all wrong.

Third, polytomous items can not only increase the accuracy of the estimation of the trait measured but also improve other aspects of a test. For example, the more information obtained from polytomous items can result in a reduction of test length. It means that the same measurement reliability can be achieved for a test

with fewer polytomous items than dichotomous items. A shortened test can improve the motivation of test takers, improve test security, and reduce the cost of test administration.

GRADED RESPONSE MODEL (GRM)

Samejima (1969) proposed the graded response model that can be used for modeling polytomous items whose categorical options are ordered. A noticeable aspect of the graded response model is that it does not directly compute the probability that an examinee chooses a particular category, but instead computes the probability that an examinee chooses the category as the difference between two adjacent boundary characteristic curves.

BOUNDARY CHARACTERISTIC CURVES

The boundary characteristic curves represent the boundaries on the cumulative probabilities of the response categories (Baker, 1992). For example, the first boundary characteristic curve is obtained by assigning 0 (incorrect) to the selection of the first category and 1 (correct) to the selection of the other categories. For the second boundary characteristic curve, 0 is assigned to the selection of either the first or the second category but 1 to the selection of the other categories and so forth. In this way, *m - 1* boundary characteristic curves for a polytomous item with *m* categories can be obtained as shown in Figure 2, in which the category boundaries are -1.5, -0.5, 0.5, and 1.5.

In this paper, the boundary characteristic curve is defined as P_k^* which indicates the probability that an examinee selects category higher than *k* of a polytomous item. Each of the *m - 1* boundary characteristic curves is the same as the item characteristic curve for a dichotomous item with the 2PLM. However, they are used in the graded response model as an intermediate step to obtain the item category characteristic curves for a polytomous item. Additionally, $P_0^* = 1$ and $P_m^* = 0$ should be defined for the operational purpose.

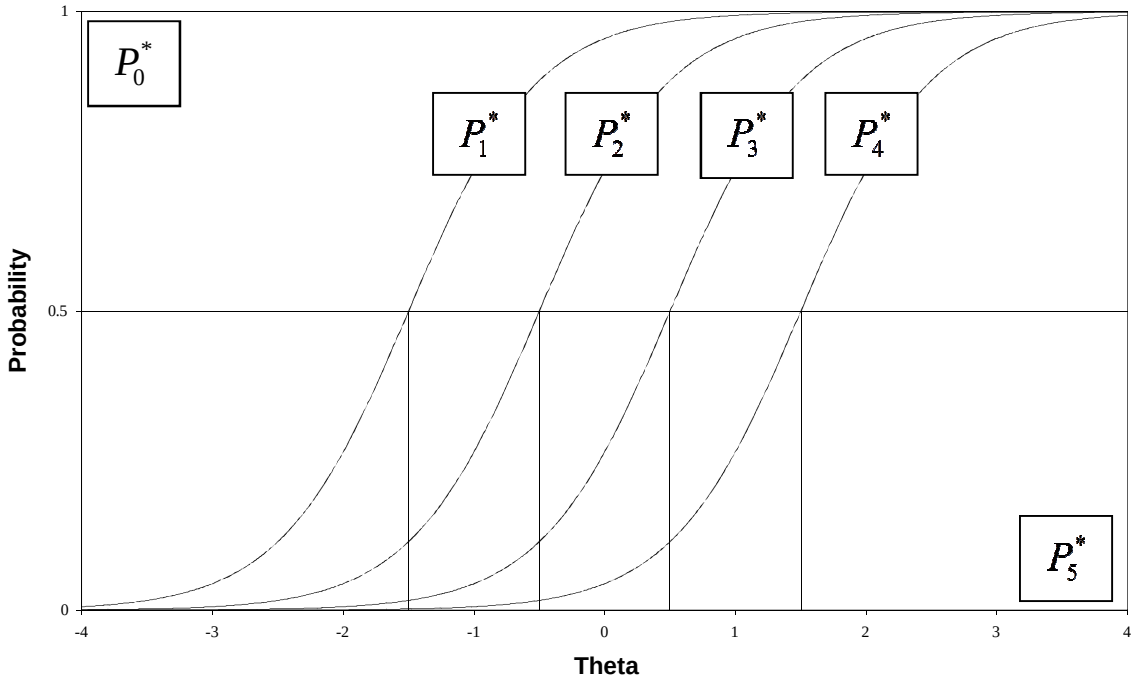

Figure 2. The boundary characteristic curves for a polytomous item with 5 categories

Embretson (2000) pointed out that the boundary difficulty parameters represent the ability level necessary to respond above threshold *k* with probability of 0.5.

ITEM CATEGORY CHARACTERISTIC CURVES

Once the *m -1* boundary characteristic curves are determined, the *m* item category characteristic curves can be computed by the difference between two adjacent boundary characteristic curves, as shown in equation 2. The probability of choosing category *k* (*k = 1,2,3...m*) of a polytomous item given θ is

$$P(x = k \mid \theta) = P^*_{k-1} - P^*_k = \frac{1}{1+e^{-1.7a(\theta - b_{k-1})}} - \frac{1}{1+e^{-1.7a(\theta - b_k)}}, \qquad (2)$$

Where,

- a is the discrimination parameter of a boundary characteristic curve,
- b is the difficulty parameter of a boundary characteristic curve,
- θ is the ability of an examinee.

In the graded response model, a polytomous item with *m* categories is characterized by one slope parameter and *m-1* boundary difficulty parameters. The slope parameter indicates the discriminating power of the polytomous item, and is best shown in Figure 2 above. A large slope parameter is reflected by boundary characteristic curves with a steep slope at the point of inflection (at a boundary difficulty parameter), and whose boundary difficulty parameters are close together.

Figure 3 shows the item category characteristic curves which are computed from the boundary characteristic curves in Figure 2 according to equation 2. Each item category characteristic curve is computed as the probability that an examinee chooses a particular category conditional on his or her ability level. The item category characteristic curve is monotonically decreasing for the 1st option, monotonically increasing for the last option, and unimodal for the middle 3 options. In addition, the boundary difficulty parameters are reflected by the point of inflection for the outer 2 item category characteristic curves and the mode of the middle 3 item category characteristic curves. For example, Figure 3 shows that an examinee whose ability is near 0 is more likely to select category 3 than other categories.

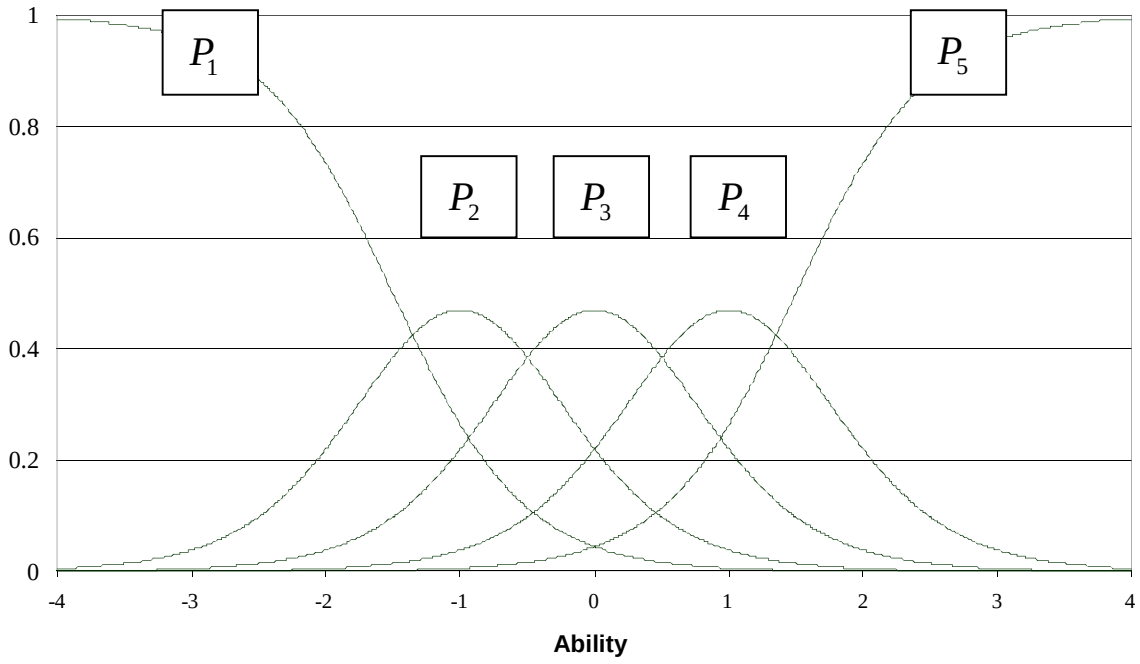

Figure 3. Item Category Characteristic Curves of a Polytomous item with 5 categories

METHODS

A simulation study was conducted to evaluate the accuracy of SAS®-PIRT in recovering the parameters of polytomous items satisfying a GRM model. In this study, two simulation factors are manipulated: the number of examinees and the number of items. For the number of examinees, 1000 and 500 examinees were chosen to represent large and small testing volumes. For the number of items, 30 and 15 items were chosen to represent a long and short test, respectively. Each polytomous item contains 4 categories, reflected by 3 category boundary boundaries. 200 different response data sets were generated for each of the four conditions as reflected by crossing the number of examinees with the number of items. Each data set was analyzed by two different calibration programs: SAS®--PIRT and PARSCALE (Muraki & Bock, 1998).

The behavior of an examinee on polytomous items can be simulated since the probabilities that an examinee selects each of the categories are available when the item parameters of the boundary characteristic curves and the examinee true ability parameters are generated. For example, the GRM curves shown in Figure 3 show that the probability of each response category is 0.0018, 0.0119, 0.0836, 0.3568, and 0.5458 for categories 1 to 5, respectively, when the examinee ability is 1.2, in which case the examinee is most likely to respond with category 5.

GENERATING SIMULATED DATA RESPONSES

The ability parameters of examinees were generated from a normal distribution (θ ~ N[0, 1]). 1000 examinee ability parameters were generated and first 500 were used for the 500 examinee group. Table 1 shows the descriptive statistics of simulated examinee's true ability (theta) for both groups.

Table 1. Descriptive statistics for simulated examinee ability parameter

N	Mean	SD	Skewness	Kurtosis
1000	0.059814	0.988833	-0.04587	-0.01806
500	0.052326	1.017974	-0.01391	0.01650

For item parameter generation, the slope parameters were generated from a log-normal distribution (a ~ lognormal[1, 0.25]). Since polytomous items with 4 categories are used for generating simulation data sets, 3 boundary difficulty parameters need to be generated accordingly from the standard normal distribution (b_k ~ N[0, 1]) assuming they should be ordered ($b_1 < b_2 < b_3$).

Table 2 shows item parameters used to generate examinee responses. 30 item parameters were generated from the distribution as aforementioned. For the 15 item test, the first 15 item parameters were used to generate item responses.

Table 2. Item parameters for simulated response

Item	15 Items				30 Items			
	a	*b1*	*b2*	*b3*	*a*	*b1*	*b2*	*b3*
1	1.059	0.803	0.885	2.435	1.059	0.803	0.885	2.435
2	0.668	-0.561	-0.422	-0.105	0.668	-0.561	-0.422	-0.105
3	2.032	0.175	0.921	1.222	2.032	0.175	0.921	1.222
4	0.922	-0.755	-0.678	1.554	0.922	-0.755	-0.678	1.554
5	0.495	-1.416	-1.119	0.150	0.495	-1.416	-1.119	0.150
6	1.159	0.935	1.021	1.308	1.159	0.935	1.021	1.308
7	1.158	-0.169	0.588	0.673	1.158	-0.169	0.588	0.673
8	0.916	-0.576	0.910	1.289	0.916	-0.576	0.910	1.289
9	1.274	-0.995	-0.863	2.258	1.274	-0.995	-0.863	2.258
10	1.262	-0.488	-0.036	2.074	1.262	-0.488	-0.036	2.074
11	2.51	-2.148	0.416	0.809	2.51	-2.148	0.416	0.809
12	0.841	-1.075	0.061	0.496	0.841	-1.075	0.061	0.496
13	1.639	-1.092	-0.761	-0.069	1.639	-1.092	-0.761	-0.069
14	1.455	-1.068	0.262	0.931	1.455	-1.068	0.262	0.931
15	1.494	0.259	0.796	1.409	1.494	0.259	0.796	1.409
16					1.335	-1.251	-0.951	0.31
17					0.811	-2.624	0.332	1.205
18					0.788	-0.512	0.732	0.976
19					1.503	-1.315	0.605	1.319
20					1.608	-1.21	1.383	1.547
21					0.925	-0.541	-0.314	0.434
22					0.915	0.754	0.894	1.574
23					1.800	0.182	0.338	2.143
24					1.766	0.325	0.863	1.006
25					0.959	-0.21	0.053	2.787
26					1.170	-1.204	-0.197	0.572
27					1.028	-0.441	-0.302	0.701
28					1.908	-0.254	-0.136	-0.066
29					1.039	-1.291	-0.615	-0.501
30					1.254	-0.419	-0.016	0.365
Mean	1.259	-0.545	0.132	1.100	1.256	-0.606	0.155	1.027
SD	0.521	0.834	0.741	0.795	0.448	0.824	0.678	0.812

ESTIMATION METHOD

In estimating the parameters of the boundary characteristic curves of polytomous items, the marginal maximum likelihood estimation (MMLE) method was implemented. Unlike the joint maximum likelihood estimation (JMLE) method, the process of estimating polytomous item parameters is not dependent on the process of estimating ability parameters. Since a hyper distribution for ability (e.g., standard normal distribution) was assumed, the parameters of the polytomous items were estimated by integrating the joint probability function of ability and the likelihood over the assumed ability distribution.

As noted, the EM algorithm consists of two steps. In the expectation step, the expected number of examinees and the expected number of examinees who provide each response to the item is computed based on the posterior distribution. In the maximization step, the item parameters which maximize the marginal likelihood are updated using the Newton-Raphson method. In this study, a maximum of 500 EM

cycles and 3 Newton-Raphson iterations were implemented with 41 quadrature points for examinee ability estimates.

To verify the results from SAS®-PIRT, a commonly used commercial IRT program, PARSCALE was used to compare the outcomes. The same calibration parameters (e.g., maximum EM cycles and the number Newton-Raphson iteration) were used for both programs. Pearson correlation coefficients, mean bias error (MBE) and root mean squared error (RMSE) were calculated for each item parameter as evaluation criteria. MBE and RMSE were calculated as follows;

$$\mathrm{MBE} = \frac{\sum_{i=1}^{n}(X_i - X_t)}{n},$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - X_t)^2}{n}},$$

where x_t is true parameter, x_i is the estimator, and n is the number of replications. Pearson correlation coefficients were computed similarly using a correlation equation. Each evaluation criterion was averaged over the 200 replications in each condition.

RESULTS

All of the data showed successful convergence in SAS®-PIRT program, however, PARSCALE did not reach convergence for some replications in which the sample size is 500. Data sets that showed convergence failure were replaced with different ones and implemented again in both programs. Therefore, the results displayed in this section consisted of outcomes from successfully converged data sets only.

Correlation coefficients between the true and estimated parameters were averaged over the 200 replications to evaluate the performance of SAS®-PIRT and PARSCALE in the recovery of the true parameters. Table 3 shows that estimates from two calibration program outcomes were highly correlated with generating parameter values. Although correlations obtained with PARSCALE were slightly higher than correlations obtained with SAS®-PIRT for most of the conditions and item parameters, the differences between the two estimation programs are almost negligible. All correlations are above 0.97 and the difference between correlation coefficients from the two estimation programs are approximately 0.001 for the discrimination parameter and 0.0001 for the boundary difficulty parameters across the conditions.

Table 3. Mean correlation coefficients for item parameters and estimates

Condition	SAS®-PIRT				PARSCALE			
	a	b1	b2	b3	a	b1	b2	b3
E500 I15	0.98753	0.99736	0.99754	0.99630	**0.98867**	**0.99739**	**0.99755**	**0.99634**
E500 I30	0.97267	0.99422	0.99441	0.99334	**0.97389**	**0.99431**	**0.99447**	**0.99338**
E1000 I15	0.97630	0.99452	0.99499	**0.99286**	**0.97844**	**0.99462**	**0.99505**	0.99285
E1000 I30	0.98537	**0.99735**	**0.99742**	**0.99651**	**0.98577**	0.99723	0.99718	0.99640

* *Note:* 1) E500/1000 indicates the number of examinees.
2) I15/30 indicates the number of items.
3) Bold typed number indicates greater value.

Table 4 shows average mean bias error (MBE) computed over 200 replications. Most comparisons show that the absolute bias is slightly smaller for parameters obtained by PARSCALE than for parameters obtained by SAS®-PIRT, although the differences become very small when the number of examinees and items is large (i.e., the E1000/I30 condition). The bias is in the negative direction for almost all comparisons, indicating that the program underestimates the parameter. Only the discrimination parameter tends to be overestimated.

Table 4. Average mean bias error (MBE) for SAS®-PIRT and PARSCALE

	MBE							
	E 500/I15		E 500/I30		E 1000/I15		E 1000/I30	
	SAS®-PIRT	PARSCALE	SAS®-PIRT	PARSCALE	SAS®-PIRT	PARSCALE	SAS®-PIRT	PARSCALE
a	-0.00344	**-0.00138**	**0.03127**	0.03256	0.04601	**0.04398**	-0.00963	**-0.00962**
b1	-0.06825	**-0.06694**	**-0.03847**	-0.04177	-0.04333	**-0.04259**	-0.07060	**-0.07026**
b2	**-0.06048**	-0.06133	-0.05168	**-0.05076**	-0.05915	**-0.05444**	-0.06083	**-0.05901**
b3	**-0.04780**	-0.04836	**-0.05790**	-0.06166	**-0.06482**	-0.06896	-0.04834	**-0.04468**

* *Note:* 1) E500/1000 indicates the number of examinees.
2) I15/30 indicates the number of items.
3) Bold typed number indicates smaller absolute value.

Table 5 shows average root mean squared error (RMSE) computed over 200 replications. Most comparisons show that the RMSE is slightly smaller for parameters obtained by PARSCALE than for parameters obtained by SAS®-PIRT, although the differences become very small when the number of examinees and items is large (i.e., the E1000/I30 condition). The RMSE values for *b3* are generally smaller for SAS®-PIRT than for PARSCALE, which is shown in MBE values in Table 4 as well.

Table 5. Average root mean squared error (RMSE) for SAS®-PIRT and PARSCALE

	RMSE							
	E 500/I15		E 500/I30		E 1000/I15		E 1000/I30	
	SAS®-PIRT	PARSCALE	SAS®-PIRT	PARSCALE	SAS®-PIRT	PARSCALE	SAS®-PIRT	PARSCALE
a	0.00864	**0.00780**	0.01483	**0.01416**	0.02339	**0.02037**	0.00660	**0.00646**
b1	0.00887	**0.00854**	0.01003	**0.00993**	0.00963	**0.00934**	**0.00904**	0.00921
b2	0.00702	**0.00700**	0.00831	**0.00789**	0.00994	**0.00909**	0.00647	**0.00644**
b3	0.00793	**0.00789**	**0.01295**	0.01302	**0.01481**	0.01517	0.00759	**0.00746**

* *Note:* 1) E500/1000 indicates the number of examinees
2) I15/30 indicates the number of items
3) Bold typed number indicates smaller value.

Both SAS®-PIRT and PARASCALE precisely estimate parameters of polytomous items regardless of the number of examinees and the number of items. Although PARSCALE showed slightly better performance in most of the measurement criteria, again, the difference between two programs are negligible. In addition, considering the convergence failure in the PARSCALE runs in the 500 examinee condition, SAS®-PIRT shows more robust outcomes in polytomous item estimation especially under the small sample size. Further analyses will be required to examine under which conditions both program may show advantages and disadvantages.

There are some limitations for generalizing outcomes from this study. First, only two values were provided for each independent variable (number of examinees and number of items). The estimation programs should be investigated for the number of response categories, the number of EM cycles, and the number of Newton-Raphson iterations. It is recommended to verify performance comparison with real data or simulation data with item parameters from real test data. In a similar venue, performance of the programs may show different outcomes under different conditions such as item parameter values other than the generated distribution. It is also recommended to verify the performance with data generated from different item and examinees parameter distributions.

REFERENCES

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques.* New York: Dekker.

Embretson, S. E. & Reise, S. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum Publishers.

Muraki, E. & Bock, R., D. (1998). *PARSCALE (Version 3.5): IRT item analysis and test scoring for rating-scale data* [Computer program]. Chicago, IL: Scientific Software.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34, 100–114.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Sung-Hyuck Lee
ACT, Inc
319-341-2417
Sung.lee@act.org