Research Memorandum
ETS RM–20–02

A Preliminary Comparison of Five Software Applications to Estimate Unidimensional Item Response Theory Models

Jianbin Fu

May 2020

ETS Research Memorandum Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Memorandum series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

A Preliminary Comparison of Five Software Applications to Estimate Unidimensional Item Response Theory Models

Jianbin Fu
Educational Testing Service, Princeton, New Jersey

May 2020

Corresponding author: J. Fu, E-mail: jfu@ets.org@ets.org

Suggested citation: Fu, J. (2020). *A preliminary comparison of five software applications to estimate unidimensional item response theory models* (Research Memorandum No. RM-20-02). Educational Testing Service.

Action Editor: James Carlson

Reviewer: Xueli Xu and Yanming Jiang

Abstract

Five item response theory (IRT) computer programs, IRTPRO, flexMIRT, PARSCALE, mdltm, and MIRT, are compared in terms of item parameter estimates. The five programs are used to run the one-parameter logistic (1PL)/partial credit model (PCM), two-parameter logistic (2PL)/generalized partial credit model (GPCM), and/or three-parameter logistic (3PL)/GPCM on two real data sets and 30 simulated data sets. For real and simulated data sets, the (mean) correlations, differences, and root mean square differences of item parameter estimates among the five programs are compared, and for the simulated data sets, these statistics with the true parameters are also reported. The advantages and disadvantages of each program are discussed. The flexMIRT program is recommended for calibrating large-scale assessment data. It is further recommended that Educational Testing Service develop shareable in-house IRT software based on mdltm, MIRT, or the National Assessment of Educational Progress version of PARSCALE.

Keywords: IRT software comparison, unidimensional IRT models

Acknowledgments

Thanks are due to James Carlson, Yanming Jiang, and Xueli Xu for their helpful suggestions on and edits to early versions of this report. The author is grateful to Ariela Katz, Kim Fryer, Holly Monteith, and Brenda Whiteman for their editorial assistance on this report. Any remaining errors are mine alone.

Table of Contents

List of Tables

Page

List of Figures

Page

In this research memorandum, the following five computer programs for estimating item response theory (IRT) models are compared: (a) IRTPRO (Cai et al., 2011), (b) flexMIRT (Cai, 2017; Houts & Cai, 2016), (c) PARSCALE (Muraki & Bock, 2003), (d) mdltm (Shin et al., 2016; von Davier & Xu, 2009), and (e) MIRT (Haberman, 2013). All of the programs, except for PARSCALE, are designed for estimating general IRT models that include a variety of unidimensional and multidimensional IRT models. However, in this research memorandum, only the estimations of the following unidimensional models are compared: the one-parameter logistic (1PL) model, two-parameter logistic (2PL) model, three-parameter logistic (3PL) model, partial credit model (PCM), and generalized partial credit model (GPCM). All the aforementioned programs can estimate these unidimensional IRT models; the one exception is that mdltm does not have functionality to estimate the 3PL model. The five IRT programs were used to run 1PL/PCM, 2PL/GPCM, and/or 3PL/GPCM, as applicable, on two real data sets and 30 simulated data sets.

This research memorandum is organized into five sections. First, the formulas for the compared IRT models are provided. Second, the process to generate simulated data sets is described. Third, the comparison results of item parameter estimates are presented; specifically, the correlations, mean differences, and root mean square differences (RMSDs) of item parameter estimates are reported and compared among the five programs for each real data set and each applicable combination of IRT models. For the simulated data, the means of correlations, differences, and RMSDs of item parameter estimates and true values across the 30 simulated data sets from the five IRT programs are compared for each applicable IRT model combination. Fourth, the advantages and disadvantages of the features in each program are discussed. Finally, recommendations as well as the corresponding rationales are provided for selecting IRT software for large-scale assessment data.

Item Response Theory Model Formulations

IRT models are often formulated differently (but equivalently via parameter transformation) in different research papers and software manuals. See the appendix for the default IRT model formulas implemented in the five computer programs. In this research memorandum, the following formulas for IRT models were used in the comparisons. The 3PL model with dichotomous items is specified as

$$P_{ij} = p(x_{ij} = 1 \mid \theta_j, a_i, b_i) = c_i + (1 - c_i)\frac{\exp(a_i\theta_j - b_i)}{1 + \exp(a_i\theta_j - b_i)}, \tag{1}$$

where x_{ij} is test taker *j*'s score on item i, a_i is the discrimination (slope) parameter for item i, b_i is the parameter related to item difficulty and referred to here as the item difficulty parameter, c_i is the lower asymptote (pseudo guessing) parameter, θ_j is test taker *j*'s latent (theta) score, and P_{ij} is test taker *j*'s probability of answering item i correctly according to the model. For the 2PL and 1PL models, c_i is set to 0; additionally, a_i is set to 1 for the 1PL model. For the GPCM with polytomous items,

$$P_{ijm} = p(x_{ij} = m \mid \theta_j, a_i, b_i, \mathbf{d}_i) = \frac{\exp\left[\sum_{h=0}^{m}(a_i\theta_j - b_i + d_{ih})\right]}{\sum_{v=0}^{M_i-1}\exp\left[\sum_{h=0}^{v}(a_i\theta_j - b_i + d_{ih})\right]}, \tag{2}$$

where $a_i\theta_j - b_i + d_{i0} \equiv 0$, M_i is item i's number of score categories, m is item i's possible integer score point ranging from 0 to $M_i - 1$, b_i is the item location parameter on item *i*, d_{ih} is the step parameter on item *i* for score h, $\mathbf{d}_i$ is the vector with elements d_{ih}, P_{ijm} is test taker *j*'s probability of achieving score m on item i according to the model, and the other parameters are defined the same as in Equation 1. The statistic $b_i - d_{ih}$ is referred to as the item category parameter. For PCM, a_i is set to 1 in Equation 2.

Real and Simulated Data

Two real data sets were denoted as Data A and Data B. Data A was problematic in terms of convergence compared to Data B. The numbers of items and test takers in each data set are shown in Table 1.

For each of the three IRT model combinations (i.e., 1PL/PCM, 2PL/GPCM, and 3PL/GPCM), 30 simulated data sets were generated. Each simulated data set included 3,000 test takers, 40 dichotomous items, 10 three-score-category items, and 10 four-score-category items (see Table 1). For polytomous items, the responses were simulated based on the GPCM formula

$$P_{ijm} = p(x_{ij} = m \mid \theta_j, a_i, \mathbf{t}_i) = \frac{\exp\left(t_{im} + m a_i \theta_j\right)}{\sum_{v=0}^{M_i - 1} \exp\left(t_{iv} + v a_i \theta_j\right)}, \tag{3}$$

where t_{im} is the parameter related to the easiness of score m of item i and $\mathbf{t}_i$ is the vector with elements t_{im}. Equations 2 and 3 are equivalent via the transformation of the parameters t_{im} to b_i and d_{ih}. For all simulated data, a_i s were generated from a lognormal distribution with shape parameter equal to −.0196 and location parameter equal to .1980, which corresponded to the mean of a_i s equal to 1 and the standard deviation equal to .2 on the arithmetic scale; θ_j s, b_i s, and t_{im} s were generated from a standard normal distribution. For 3PL data, c_i s were generated from a beta distribution with $\beta = 5$ and $\alpha = 17$. Each simulated data set had the common set of generating item parameters and a unique set of ability (theta) parameters.

Table 1. Data Sets: Number of Items and Sample Size

Data set	Two score categories	Three score categories	Four score categories	Total	Sample size
Data A	171	100	8	279	41,130
Data B	158	80	16	254	40,473
Simulation	40	10	10	60	3,000

Model Estimation Setups

The estimations of the two real data sets in PARSCALE were carefully set up for smooth calibrations, including removing problematic items and selecting prior distributions and appropriate starting values for item parameters. For fair comparisons, the setups of all estimations of real and simulated data sets were kept as similar as possible across the five programs, as described in detail in the following pages.

Item Parameter Priors

Item priors were used in PARSCALE, IRTPRO, and flexMIRT. Specifically, for the two real data sets, a lognormal prior distribution with shape parameter equal to 0 and location parameter equal to .5 were used for the discrimination parameters in the 2PL/GPCM and 3PL/GPCM models, which is the default prior distribution in PARSCALE. For the simulated data sets, a lognormal prior distribution with shape parameter equal to −.0196 and location parameter equal to .1980 were used for the discrimination parameters in the 2PL/GPCM and 3PL/GPCM models, which matched the distribution of the generating discrimination parameters. For the guessing parameters, a beta prior distribution with $\beta = 5$ and $\alpha = 17$ were used in all estimations, which is the default prior distribution for guessing parameters in PARSCALE. For mdltm, there is no option to define prior distributions for item parameters, and for MIRT, the option is difficult to implement because of the formulation and structure of MIRT.

Starting Values

Starting values for all discrimination parameters, as well as some item category parameters and guessing parameters, were used in PARSCALE for estimating the two real data sets. Other than that, starting values were not used in any other calibrations.

Convergence Criteria

The convergence criteria implemented in the programs are not all the same. For PARSCALE, the convergence criterion was the largest item parameter change between two consecutive iterations less than .001 in both E-steps and M-steps. For mdltm, the convergence criteria were (a) the largest item parameter change between two consecutive iterations less than .001 and (b) the log likelihood difference between two consecutive iterations less than .01 in E-steps. For IRTPRO and flexMIRT, the convergence criterion was tolerance less than 10^{-5} in both E-steps and M-steps, where tolerance is defined as (log likelihood at the current iteration − log likelihood at the prior iteration)/absolute log likelihood at the prior iteration. For MIRT, the convergence criterion was tolerance less than 10^{-5} for the two real data sets and 10^{-8} for the simulated data sets. The criteria in all programs led to the log likelihood changes between the final two iterations less than .01, except for the MIRT runs on the two real data sets, for which the log likelihood differences were less than 5 at convergence. The loose criterion (which is the default in MIRT) helped MIRT converge in a reasonable time period on the two real data sets.

Quadrature Points

All programs, except for MIRT, used 41 equal space points between −4 and 4 with normal approximation to approximate the population ability distribution during the maximum marginal likelihood estimation of item parameters. For MIRT, five adaptive Gauss–Hermite quadrature points were used, as recommended by the program manual.

Output

For a fair comparison of estimation times, the following statistics, besides item parameters, were estimated and output in all programs, if available: expected a prior ability (theta), basic item fit statistics for single items, and item information.

Comparison Criterion

In the next section, estimates of discrimination (a_i) and difficulty/location parameters (b_i) from the five programs are compared; for the simulated data, these estimates are also compared to the generating values. In particular, for the real data, the correlations, mean differences, and RMSDs of discrimination and difficulty/location parameter estimates among the five programs are compared. For the simulated data, the mean correlations, differences, and RMSDs (over 30 replicated data sets) of discrimination and difficulty/location parameter estimates among the five programs and generating values are compared. The mean differences and RMSDs are calculated as

$$\text{mean difference} = \frac{1}{RI}\sum_{r=1}^{R}\sum_{i=1}^{I}(\hat{\eta}_{ri} - \eta_{ri}^{*})$$

$$\text{RMSD} = \frac{1}{R}\sum_{r=1}^{R}\sqrt{\frac{1}{I}\sum_{i=1}^{I}(\hat{\eta}_{ri} - \eta_{ri}^{*})^{2}}\ ,$$

where $\hat{\eta}_{ri}$ is the estimated item parameter from one program and η_{ri}^{*} is the estimated item parameter from another program or the generating item parameter of item i in replicated data set r; R is the number of replicated data sets ($R = 1$ for a real data set and 30 for simulated data sets); and I is the total number of items in a replicated data set.

In addition, estimation times are compared among the five programs for each model combination, as applicable.

Comparison of Item Parameter Estimates

As the following comparison results show, all the estimates from flexMIRT and IRTPRO were the same, excluding rounding differences.

One-Parameter Logistic/Partial Credit Model

Tables 2–4 list the correlations, mean differences, and RMSDs of item difficulty/location parameter estimates from the five programs on Data A, respectively; Tables 5–7 list the comparison results for Data B; and Tables 8–10 show the comparison results for the simulated data. Figures 1–3 are the scatterplots of item difficulty/location parameter estimates from the five programs on Data A, Data B, and the simulated data, respectively. One can observe that all correlations were larger than .99. For the real data sets, the mean differences and RMSDs show that PARSCALE's estimates were closer to mdltm than the other programs, and MIRT's estimates had the largest differences from the other programs, especially for Data A, for which MIRT's estimates were generally larger than the other programs' estimates, as visually presented in Figure 1. For the simulated data, the estimates from all programs were very close, and the mean RMSDs with the true values were all .04.

Table 11 lists the running times for Data A and Data B and the mean running times over 30 replicates for simulated data. The mdltm program was the fastest on the two real data sets, while the slowest on the simulated data sets. On the two real data sets, flexMIRT and IRTPRO were the second and third fastest, respectively. For simulated data sets, the fastest programs were flexMIRT, IRTPRO, and PARSCALE. MIRT was the slowest on real data sets and the second slowest on simulated data sets.

Table 2. Data A One-Parameter Logistic/Partial Credit Model: Correlations of Difficulty/Location Parameter Estimates

Software	IRTPRO	flexMIRT	PARSCALE	mdltm
flexMIRT	1.000			
PARSCALE	1.000	1.000		
mdltm	1.000	1.000	1.000	
MIRT	.994	.994	.995	.995

Table 3. Data A One-Parameter Logistic/Partial Credit Model: Mean Differences of Difficulty/Location Parameter Estimates

Software	IRTPRO	flexMIRT	PARSCALE	mdltm
flexMIRT	.000			
PARSCALE	.013	.013		
mdltm	.014	.014	.001	
MIRT	−.018	−.018	−.032	−.033

Note. Difference is calculated as column method minus row method.

Table 4. Data A One-Parameter Logistic/Partial Credit Model: Root Mean Square Differences of Difficulty/Location Parameter Estimates

Software	IRTPRO	flexMIRT	PARSCALE	mdltm
flexMIRT	.000			
PARSCALE	.036	.036		
mdltm	.038	.038	.004	
MIRT	.111	.111	.100	.100

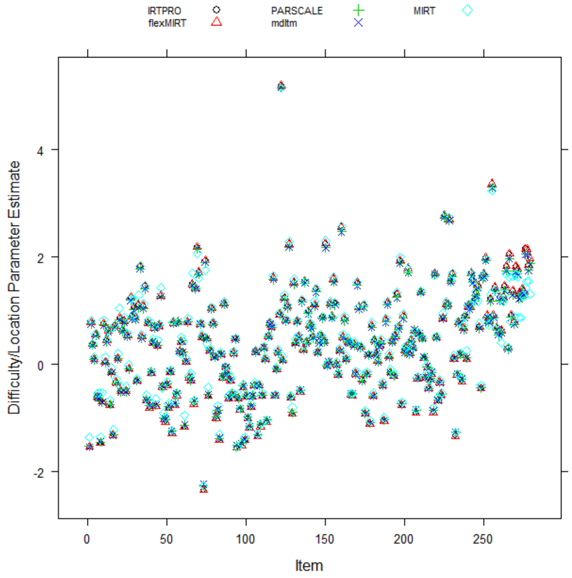

Figure 1. Data A one-parameter logistic/partial credit model: comparisons of difficulty/location parameter estimates.

Table 5. Data B One-Parameter Logistic/Partial Credit Model: Correlations of Difficulty/Location Parameter Estimates

Software	IRTPRO	flexMIRT	PARSCALE	mdltm
flexMIRT	1.000			
PARSCALE	1.000	1.000		
mdltm	1.000	1.000	1.000	
MIRT	1.000	1.000	1.000	1.000

Table 6. Data B One-Parameter Logistic/Partial Credit Model: Mean Differences of Difficulty/Location Parameter Estimates

Software	IRTPRO	flexMIRT	PARSCALE	mdltm
flexMIRT	.000			
PARSCALE	−.002	−.002		
mdltm	−.003	−.003	−.001	
MIRT	−.001	−.001	.001	.002

Note. Difference is calculated as column method minus row method.

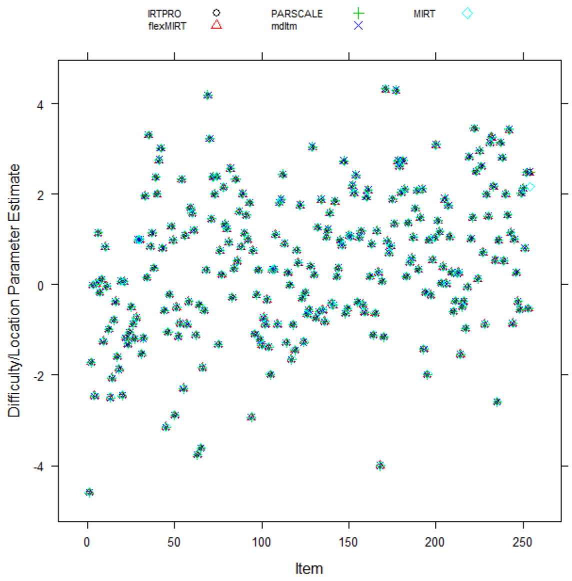

Figure 2. Data B one-parameter logistic/partial credit model: comparisons of difficulty/location parameter estimates.

Table 7. Data B One-Parameter Logistic/Partial Credit Model: Root Mean Square Differences of Difficulty/Location Parameter Estimates

Software	IRTPRO	flexMIRT	PARSCALE	mdltm
flexMIRT	.000			
PARSCALE	.009	.009		
mdltm	.010	.010	.003	
MIRT	.020	.020	.022	.023

Table 8. Simulation One-Parameter Logistic/Partial Credit Model: Mean Correlations of Difficulty/Location Parameter Estimates

Software	IRTPRO	flexMIRT	PARSCALE	mdltm	MIRT
flexMIRT	1.000				
PARSCALE	1.000	1.000			
mdltm	1.000	1.000	1.000		
MIRT	1.000	1.000	1.000	1.000	
True	.999	.999	.999	.999	.999

Table 9. Simulation One-Parameter Logistic/Partial Credit Model: Mean Differences of Difficulty/Location Parameter Estimates

Software	IRTPRO	flexMIRT	PARSCALE	mdltm	MIRT
flexMIRT	.000				
PARSCALE	.002	.002			
mdltm	.001	.001	−.001		
MIRT	.000	.000	−.001	.000	
True	.004	.004	.002	.003	.004

Note. Difference is calculated as column method minus row method.

Table 10. Simulation One-Parameter Logistic/Partial Credit Model: Mean Root Mean Square Differences of Difficulty/Location Parameter Estimates

Software	IRTPRO	flexMIRT	PARSCALE	mdltm	MIRT
flexMIRT	.000				
PARSCALE	.002	.002			
mdltm	.001	.001	.001		
MIRT	.003	.003	.003	.003	
True	.042	.042	.042	.043	.043

Table 11. One-Parameter Logistic/Partial Credit Model: Estimation Time (Seconds)

Software	Data A	Data B	Simulation [a]
IRTPRO	137	133	6
flexMIRT	89	126	5
PARSCALE	178	224	8
mdltm	51	79	14
MIRT	457	231	10

[a] Mean running time over 30 replicates.

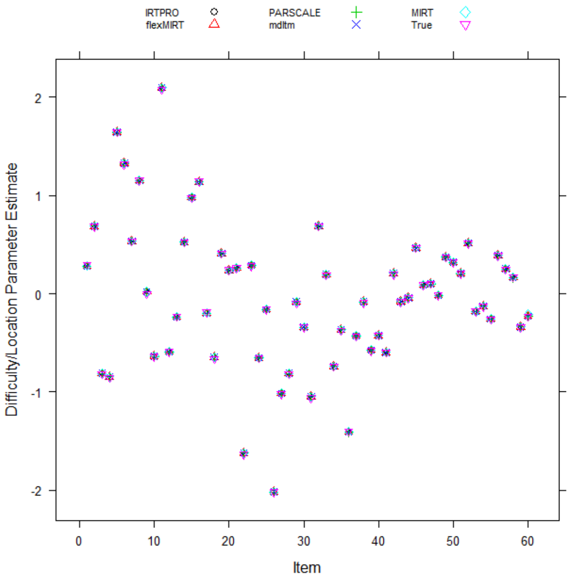

Figure 3. Simulation one-parameter logistic/partial credit model: comparisons of mean difficulty/location parameter estimates.

Two-Parameter Logistic/Generalized Partial Credit Model

Tables 12–17 present the correlations, mean differences, and RMSDs of discrimination and difficulty/location parameter estimates from the five programs on Data A. Three problematic items were removed from calibrations in all five IRT programs to enable better convergence: the three items either had negative discrimination parameters or caused nonconvergent runs. All correlations were larger than .99, except for mdltm and MIRT on discrimination parameters, which were approximately .98 and .94, respectively. The RMSDs show that the discrimination estimates from PARSCALE were close to IRTPRO and flexMIRT, whereas MIRT's estimates on both parameters deviated most from the three programs. Both MIRT and mdltm had a negative discrimination estimate on a common item, while the estimates were positive for the other three programs due to priors and/or starting values applied for discrimination parameters (as mentioned previously, mdltm does not have an option for priors, and for MIRT, they are difficult to set up). Figures 4 and 5 are the scatterplots of discrimination and difficulty/location estimates, respectively, from the five programs on Data A.

Tables 18–23 present the correlations, mean differences, and RMSDs of discrimination and difficulty/location parameter estimates from the five programs on Data B. All estimates resulted in perfect or nearly perfect correlations. In terms of mean differences and RMSDs, the estimates from all programs were also very close (all mean differences in absolute value and RMSDs smaller than .03). This pattern can be visually observed in Figures 6 and 7 for the discrimination and difficulty/location estimates, respectively, on Data B.

Tables 24–29 present the mean correlations, differences, and RMSDs of discrimination and difficulty/location parameter estimates from the five programs on simulated data. All estimates from the five programs were perfectly correlated and closer than those on Data B (all mean differences in absolute value and mean RMSDs smaller than .02). All the mean correlations with the true values were .96 for discrimination estimates and larger than .99 for difficulty/location estimates. All mean differences in absolute value with the true values were smaller than .01, and all mean RMSDs with the true values were between .04 and .05. Figures 8 and 9 are the scatterplots of discrimination and difficulty/location estimates, respectively, from the five programs on simulated data.

Table 30 lists the estimation times for the five programs on the real and simulated data sets. For simulated data sets, flexMIRT, IRTPRO, and PARSCALE were the fastest programs, and mdltm was the slowest. For Data A, flexMIRT ran the fastest, while MIRT ran the slowest. For Data B, mdltm ran the fastest, followed by flexMIRT and PARSCALE, while MIRT was the slowest.

Table 12. Data A Two-Parameter Logistic/Generalized Partial Credit Model: Correlations of Item Parameter Estimates for Discrimination

Software	IRTPRO	flexMIRT	PARSCALE	mdltm
flexMIRT	1.000			
PARSCALE	1.000	1.000		
mdltm	.988	.988	.987	
MIRT	.943	.943	.941	.981

Table 13. Data A Two-Parameter Logistic/Generalized Partial Credit Model: Correlations of Item Parameter Estimates for Difficulty/Location

Software	IRTPRO	flexMIRT	PARSCALE	mdltm
flexMIRT	1.000			
PARSCALE	.998	.998		
mdltm	.998	.998	.996	
MIRT	.993	.993	.990	.998

Table 14. Data A Two-Parameter Logistic/Generalized Partial Credit Model: Mean Differences of Item Parameter Estimates for Discrimination

Software	IRTPRO	flexMIRT	PARSCALE	mdltm
flexMIRT	.000			
PARSCALE	–.010	–.010		
mdltm	.013	.013	.023	
MIRT	.007	.007	.017	-.005

Note. Difference is calculated as column method minus row method.

Table 15. Data A Two-Parameter Logistic/Generalized Partial Credit Model: Mean Differences of Item Parameter Estimates for Difficulty/Location

Software	IRTPRO	flexMIRT	PARSCALE	mdltm
flexMIRT	.000			
PARSCALE	.004	.004		
mdltm	.009	.009	.005	
MIRT	.006	.006	.002	−.003

Note. Difference is calculated as column method minus row method.

Table 16. Data A Two-Parameter Logistic/Generalized Partial Credit Model: Root Mean Square Differences of Item Parameter Estimates for Discrimination

Software	IRTPRO	flexMIRT	PARSCALE	mdltm
flexMIRT	.000			
PARSCALE	.013	.013		
mdltm	.069	.069	.075	
MIRT	.146	.146	.151	.085

Table 17. Data A Two-Parameter Logistic/Generalized Partial Credit Model: Root Mean Square Differences of Item Parameter Estimates for Difficulty/Location

Software	IRTPRO	flexMIRT	PARSCALE	mdltm
flexMIRT	.000			
PARSCALE	.077	.077		
mdltm	.061	.061	.100	
MIRT	.132	.132	.156	.075

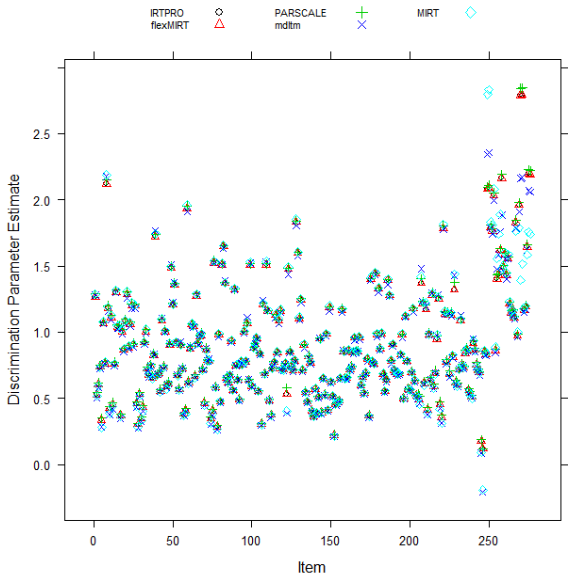

Figure 4. Data A two-parameter logistic/generalized partial credit model: comparisons of discrimination parameter estimates.

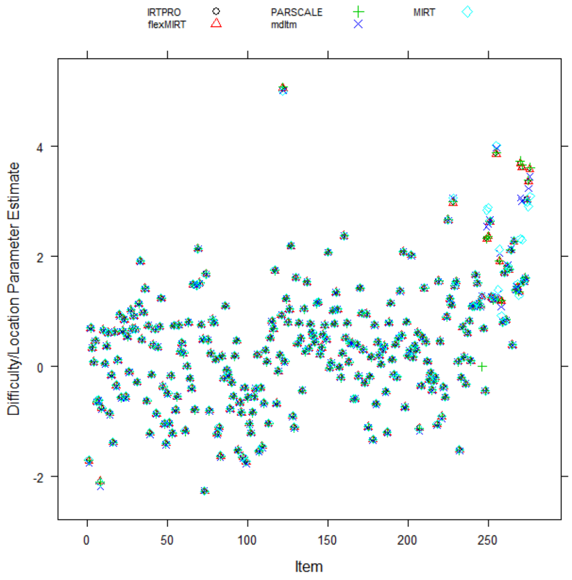

Figure 5. Data A two-parameter logistic/generalized partial credit model: comparisons of difficulty/location parameter estimates.

Table 18. Data B Two-Parameter Logistic/Generalized Partial Credit Model: Correlations of Item Parameter Estimates for Discrimination

Software	IRTPRO	flexMIRT	PARSCALE	mdltm
flexMIRT	1.000			
PARSCALE	1.000	1.000		
mdltm	.999	.999	.999	
MIRT	.999	.999	1.000	1.000

Table 19. Data B Two-Parameter Logistic/Generalized Partial Credit Model: Correlations of Item Parameter Estimates for Difficulty/Location

Software	IRTPRO	flexMIRT	PARSCALE	mdltm
flexMIRT	1.000			
PARSCALE	1.000	1.000		
mdltm	1.000	1.000	1.000	
MIRT	1.000	1.000	1.000	1.000

Table 20. Data B Two-Parameter Logistic/Generalized Partial Credit Model: Mean Differences of Item Parameter Estimates for Discrimination

Software	IRTPRO	flexMIRT	PARSCALE	mdltm
flexMIRT	.000			
PARSCALE	−.009	−.009		
mdltm	−.013	−.013	−.004	
MIRT	−.014	−.014	−.005	−.001

Note. Difference is calculated as column method minus row method.

Table 21. Data B Two-Parameter Logistic/Generalized Partial Credit Model: Mean Differences of Item Parameter Estimates for Difficulty/Location

Software	IRTPRO	flexMIRT	PARSCALE	mdltm
flexMIRT	.000			
PARSCALE	.004	.004		
mdltm	−.004	−.004	−.008	
MIRT	−.004	−.004	−.008	.000

Note. Difference is calculated as column method minus row method.

Table 22. Data B Two-Parameter Logistic/Generalized Partial Credit Model: Root Mean Square Differences of Item Parameter Estimates for Discrimination

Software	IRTPRO	flexMIRT	PARSCALE	mdltm
flexMIRT	.000			
PARSCALE	.011	.011		
mdltm	.024	.024	.016	
MIRT	.022	.022	.013	.008

Table 23. Data B Two-Parameter Logistic/Generalized Partial Credit Model: Root Mean Square Differences of Item Parameter Estimates for Difficulty/Location

Software	IRTPRO	flexMIRT	PARSCALE	mdltm
flexMIRT	.001			
PARSCALE	.006	.006		
mdltm	.018	.018	.017	
MIRT	.014	.014	.013	.007

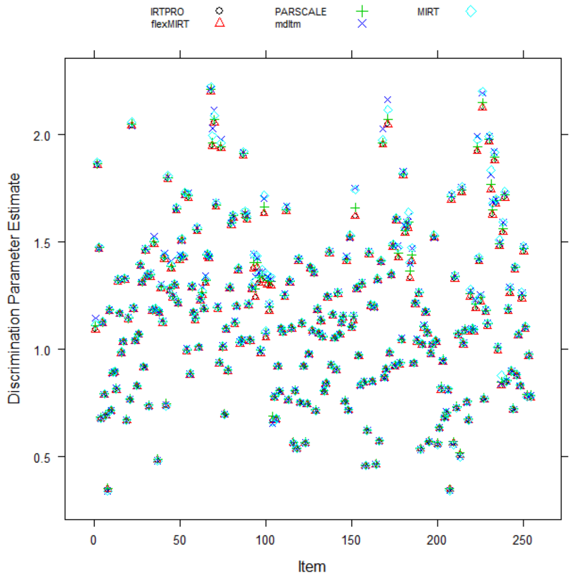

Figure 6. Data B two-parameter logistic/generalized partial credit model: comparisons of discrimination parameter estimates.

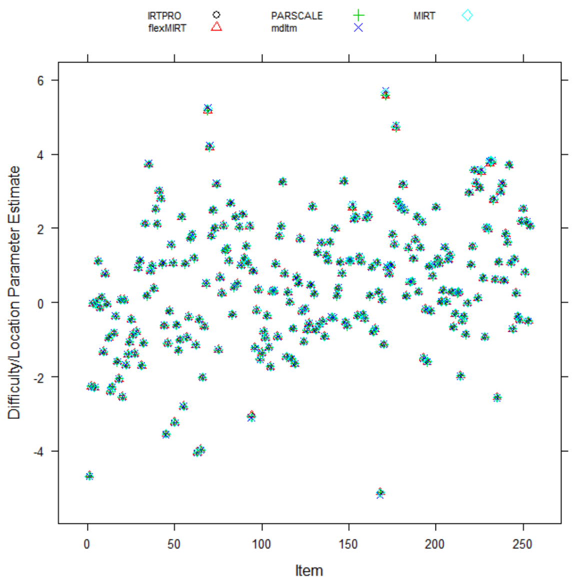

Figure 7. Data B two-parameter logistic/generalized partial credit model: comparisons of difficulty/location parameter estimates.

Table 24. Simulation Two-Parameter Logistic/Generalized Partial Credit Model: Mean Correlations of Item Parameter Estimates for Discrimination

Software	IRTPRO	flexMIRT	PARSCALE	mdltm	MIRT
flexMIRT	1.000				
PARSCALE	1.000	1.000			
mdltm	1.000	1.000	1.000		
MIRT	1.000	1.000	1.000	1.000	
True	.960	.960	.960	.960	.960

Table 25. Simulation Two-Parameter Logistic/Generalized Partial Credit Model: Mean Correlations of Item Parameter Estimates for Difficulty/Location

Software	IRTPRO	flexMIRT	PARSCALE	mdltm	MIRT
flexMIRT	1.000				
PARSCALE	1.000	1.000			
mdltm	1.000	1.000	1.000		
MIRT	1.000	1.000	1.000	1.000	
True	.999	.999	.999	.998	.998

Table 26. Simulation Two-Parameter Logistic/Generalized Partial Credit Model: Mean Differences of Item Parameter Estimates for Discrimination

Software	IRTPRO	flexMIRT	PARSCALE	mdltm	MIRT
flexMIRT	.000				
PARSCALE	−.010	−.010			
mdltm	−.013	−.013	−.003		
MIRT	−.012	−.012	−.002	.001	
True	−.010	−.010	.000	.003	.002

Note. Difference is calculated as column method minus row method.

Table 27. Simulation Two-Parameter Logistic/Generalized Partial Credit Model: Mean Differences of Item Parameter Estimates for Difficulty/Location

Software	IRTPRO	flexMIRT	PARSCALE	mdltm	MIRT
flexMIRT	.000				
PARSCALE	−.007	−.007			
mdltm	.002	.002	.008		
MIRT	.002	.002	.008	.000	
True	.003	.003	.009	.001	.001

Note. Difference is calculated as column method minus row method.

Table 28. Simulation Two-Parameter Logistic/Generalized Partial Credit Model: Mean RMSDs of Item Parameter Estimates for Discrimination

Software	IRTPRO	flexMIRT	PARSCALE	mdltm	MIRT
flexMIRT	.000				
PARSCALE	.010	.010			
mdltm	.018	.018	.012		
MIRT	.017	.017	.012	.003	
True	.047	.047	.046	.049	.049

Table 29. Simulation Two-Parameter Logistic/Generalized Partial Credit Model: Mean RMSDs of Item Parameter Estimates for Difficulty/Location

Software	IRTPRO	flexMIRT	PARSCALE	mdltm	MIRT
flexMIRT	.000				
PARSCALE	.007	.007			
mdltm	.003	.003	.009		
MIRT	.003	.003	.009	.001	
True	.045	.045	.044	.045	.045

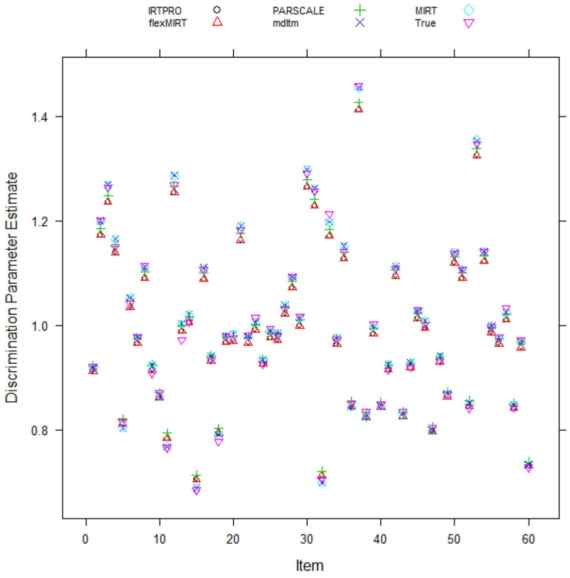

Figure 8. Simulation two-parameter logistic/generalized partial credit model: comparisons of mean discrimination parameter estimates.

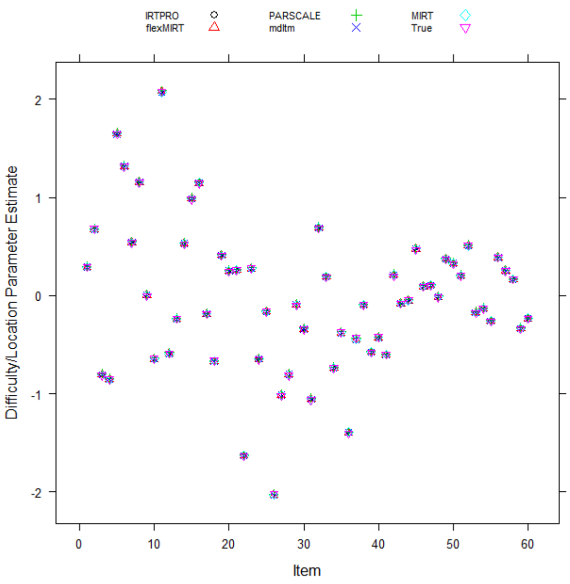

Figure 9. Simulation two-parameter logistic/generalized partial credit model: comparisons of mean difficulty/location parameter estimates.

Table 30. Two-Parameter Logistic/Generalized Partial Credit Model: Estimation Time (Seconds)

Software	Data A	Data B	Simulation [a]
IRTPRO	198	165	7
flexMIRT	147	135	6
PARSCALE	315	137	6
mdltm	275	113	22
MIRT	631	396	17

[a] Mean running time over 30 replicates.

Three-Parameter Logistic/Generalized Partial Credit Model

As mentioned previously, mdltm does not have the 3PL option and thus is not included in this section.

For Data A, the three items excluded in the 2PL/GPCM calibrations were also removed from the 3PL/GPCM calibrations for all programs. Among the 171 dichotomously scored items, 115 multiple choice (MC) items were fitted by 3PL, and the rest were fitted by 2PL. Tables 31–39 list the comparison results among the four programs on Data A. One can see that PARSCALE's estimates were close to those of flexMIRT and IRTPRO, especially on the discrimination and difficulty/location parameters, while MIRT's parameter estimates, especially discrimination and guessing, deviated from the other three programs' estimates, which is evident from the scatterplots (Figures 10–12) of the estimates of the three parameters among the four programs.

Table 31. Data A Three-Parameter Logistic/Generalized Partial Credit Model: Correlations of Item Parameter Estimates for Discrimination

Software	IRTPRO	flexMIRT	PARSCALE
flexMIRT	1.000		
PARSCALE	.998	.998	
MIRT	.700	.700	.712

Table 32. Data A Three-Parameter Logistic/Generalized Partial Credit Model: Correlations of Item Parameter Estimates for Difficulty/Location

Software	IRTPRO	flexMIRT	PARSCALE
flexMIRT	1.000		
PARSCALE	.997	.997	
MIRT	.910	.910	.907

Table 33. Data A Three-Parameter Logistic/Generalized Partial Credit Model: Correlations of Item Parameter Estimates for Guessing

Software	IRTPRO	flexMIRT	PARSCALE
flexMIRT	1.000		
PARSCALE	.936	.936	
MIRT	.582	.581	.504

Table 34. Data A Three-Parameter Logistic/Generalized Partial Credit Model: Mean Differences of Item Parameter Estimates for Discrimination

Software	IRTPRO	flexMIRT	PARSCALE
flexMIRT	.000		
PARSCALE	−.019	−.019	
MIRT	.119	.119	.138

Note. Difference is calculated as column method minus row method.

Table 35. Data A Three-Parameter Logistic/Generalized Partial Credit Model: Mean Differences of Item Parameter Estimates for Difficulty/Location

Software	IRTPRO	flexMIRT	PARSCALE
flexMIRT	.000		
PARSCALE	−.006	−.006	
MIRT	−.029	−.029	−.023

Note. Difference is calculated as column method minus row method.

Table 36. Data A Three-Parameter Logistic/Generalized Partial Credit Model: Mean Differences of Item Parameter Estimates for Guessing

Software	IRTPRO	flexMIRT	PARSCALE
flexMIRT	.000		
PARSCALE	−.003	−.003	
MIRT	−.125	−.125	−.122

Note. Difference is calculated as column method minus row method.

Table 37. Data A Three-Parameter Logistic/Generalized Partial Credit Model: Root Mean Square Differences of Item Parameter Estimates for Discrimination

Software	IRTPRO	flexMIRT	PARSCALE
flexMIRT	.000		
PARSCALE	.037	.037	
MIRT	.386	.386	.392

Table 38. Data A Three-Parameter Logistic/Generalized Partial Credit Model: Root Mean Square Differences of Item Parameter Estimates for Difficulty/Location

Software	IRTPRO	flexMIRT	PARSCALE
flexMIRT	.000		
PARSCALE	.095	.095	
MIRT	.479	.479	.495

Table 39. Data A Three-Parameter Logistic/Generalized Partial Credit Model: Root Mean Square Differences of Item Parameter Estimates for Guessing

Software	IRTPRO	flexMIRT	PARSCALE
flexMIRT	.000		
PARSCALE	.025	.025	
MIRT	.159	.159	.160

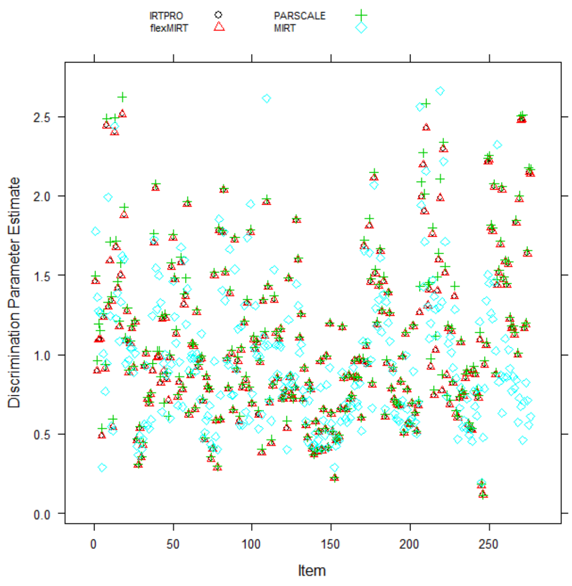

Figure 10. Data A three-parameter logistic/generalized partial credit model: comparisons of discrimination parameter estimates.

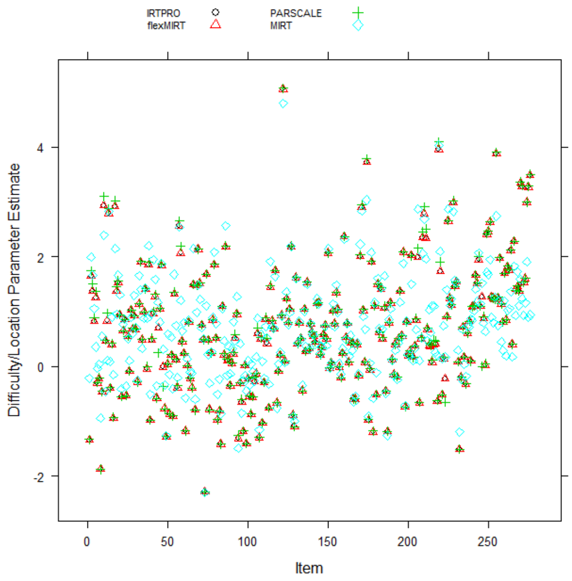

Figure 11. Data A three-parameter logistic/generalized partial credit model: comparisons of difficulty/location parameter estimates.

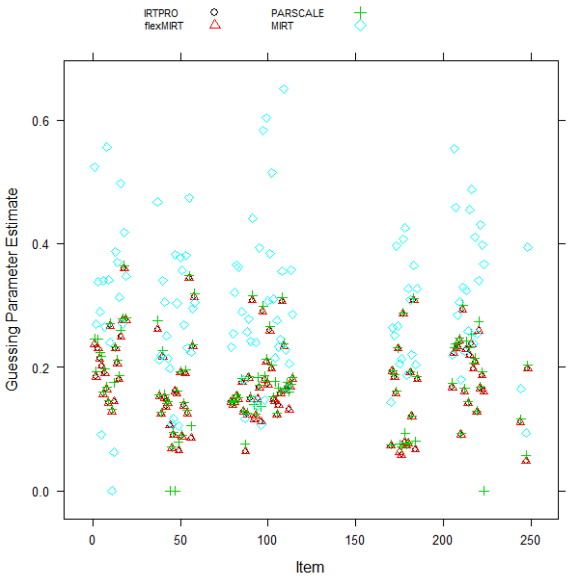

Figure 12. Data A three-parameter logistic/generalized partial credit model: comparisons of guessing parameter estimates.

For Data B, among the 158 dichotomously scored items, 44 MC items were fitted by 3PL, and the rest were fitted by 2PL. Tables 40–48 contain the comparison results, and Figures 13–15 are the scatterplots of the three parameter estimates from the four programs on Data B. PARSCALE's estimates had perfect or nearly perfect correlations with IRTPRO's and flexMIRT's, and their estimates were closer than those on Data A. Like Data A, MIRT's estimates differed from the other three programs' estimates, especially for the guessing parameters, for which MIRT's estimates were further away from the other three programs' estimates compared to those on Data A.

Table 40. Data B Three-Parameter Logistic/Generalized Partial Credit Model: Correlations of Item Parameter Estimates for Discrimination

Software	IRTPRO	flexMIRT	PARSCALE
flexMIRT	1.000		
PARSCALE	1.000	1.000	
MIRT	.712	.712	.716

Table 41. Data B Three-Parameter Logistic/Generalized Partial Credit Model: Correlations of Item Parameter Estimates for Difficulty/Location

Software	IRTPRO	flexMIRT	PARSCALE
flexMIRT	1.000		
PARSCALE	1.000	1.000	
MIRT	.962	.962	.962

Table 42. Data B Three-Parameter Logistic/Generalized Partial Credit Model: Correlations of Item Parameter Estimates for Guessing

Software	IRTPRO	flexMIRT	PARSCALE
flexMIRT	1.000		
PARSCALE	.994	.994	
MIRT	.176	.177	.196

Table 43. Data B Three-Parameter Logistic/Generalized Partial Credit Model: Mean Differences of Item Parameter Estimates for Discrimination

Software	IRTPRO	flexMIRT	PARSCALE
flexMIRT	.000		
PARSCALE	−.011	−.011	
MIRT	.202	.201	.213

Note. Difference is calculated as column method minus row method.

Table 44. Data B Three-Parameter Logistic/Generalized Partial Credit Model: Mean Differences of Item Parameter Estimates for Difficulty/Location

Software	IRTPRO	flexMIRT	PARSCALE
flexMIRT	.000		
PARSCALE	−.011	−.011	
MIRT	.202	.201	.213

Note. Difference is calculated as column method minus row method.

Table 45. Data B Three-Parameter Logistic/Generalized Partial Credit Model: Mean Differences of Item Parameter Estimates for Guessing

Software	IRTPRO	flexMIRT	PARSCALE
flexMIRT	.000		
PARSCALE	−.004	−.004	
MIRT	−.176	−.176	−.172

Note. Difference is calculated as column method minus row method.

Table 46. Data B Three-Parameter Logistic/Generalized Partial Credit Model: Root Mean Square Differences of Item Parameter Estimates for Discrimination

Software	IRTPRO	flexMIRT	PARSCALE
flexMIRT	.001		
PARSCALE	.014	.014	
MIRT	.345	.345	.351

Table 47. Data B Three-Parameter Logistic/Generalized Partial Credit Model: Root Mean Square Differences of Item Parameter Estimates for Difficulty/Location

Software	IRTPRO	flexMIRT	PARSCALE
flexMIRT	.001		
PARSCALE	.010	.010	
MIRT	.500	.500	.501

Table 48. Data B Three-Parameter Logistic/Generalized Partial Credit Model: Root Mean Square Differences of Item Parameter Estimates for Guessing

Software	IRTPRO	flexMIRT	PARSCALE
flexMIRT	.000		
PARSCALE	.009	.009	
MIRT	.218	.218	.214

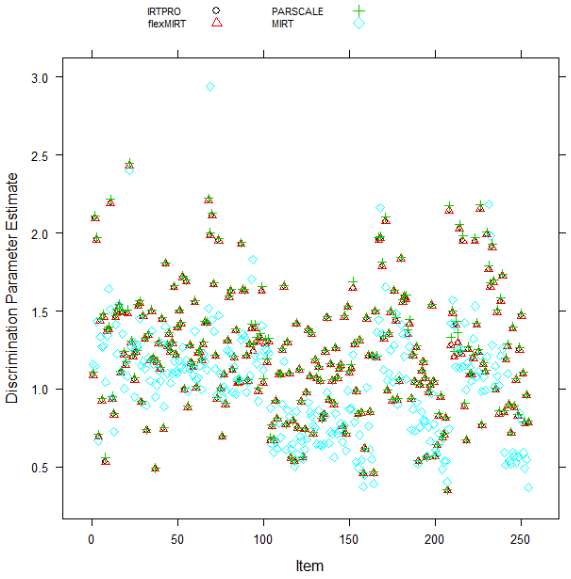

Figure 13. Data B three-parameter logistic/generalized partial credit model: comparisons of discrimination parameter estimates.

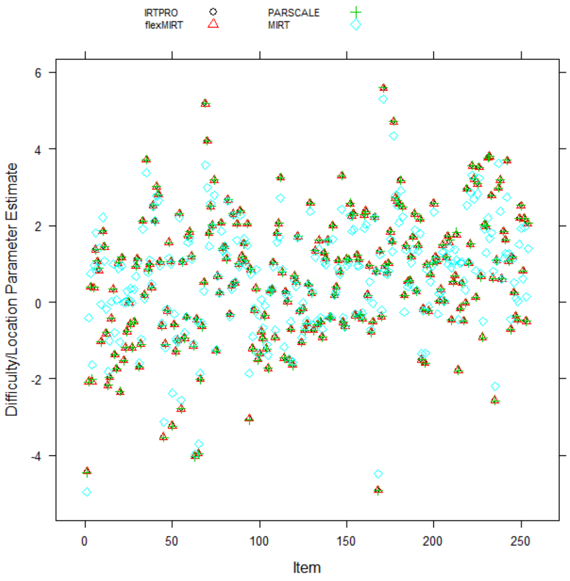

Figure 14. Data B three-parameter logistic/generalized partial credit model: comparisons of difficulty/location parameter estimates.

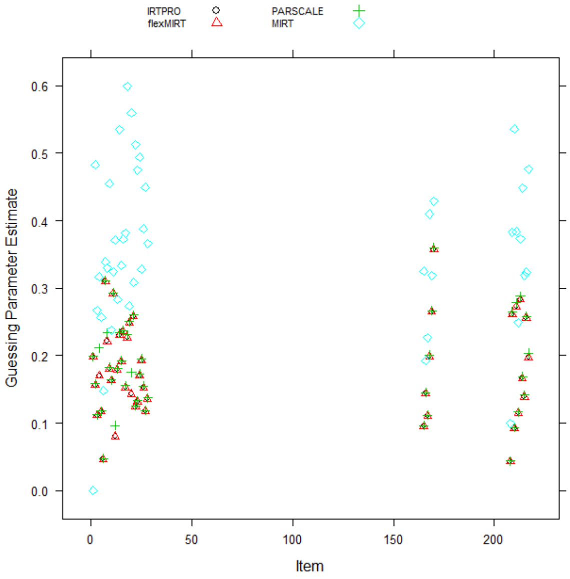

Figure 15. Data B three-parameter logistic/generalized partial credit model: comparisons of guessing parameter estimates.

Tables 49–57 present the mean correlations, differences, and RMSDs, respectively, for the simulated data sets, and Figures 16–18 are the scatterplots of item parameter estimates. The comparison results show the same pattern as those on the real data sets: PARSCALE's estimates were close to IRTPRO's and flexMIRT's, while MIRT's deviated from the other three programs' estimates, especially for the discrimination and guessing estimates. The reason for this may be that the prior distributions for the discrimination and guessing parameters were used in the other three programs but not in MIRT. Compared to the true values, PARSCALE's estimates were a little worse than IRTPRO's and flexMIRT's, while MIRT's were much worse than the other three programs' estimates. For example, the mean correlations with the true values were .85 for the discrimination estimates from the other three programs and .46 from MIRT; .98 for the difficulty/location estimates from the other three programs and .81 from MIRT; and .76 for the guessing estimates from IRTPRO and flexMIRT, .72 from PARSCALE, and .34 from MIRT. In general, all three programs slightly underestimated the true parameters, while MIRT overestimated the true parameters, as shown by the mean biases in Tables 52–54. For all programs, the parameter recoveries in 3PL/GPCM were much worse than for 1PL/PCM and 2PL/GPCM, as shown by mean RMSDs in Tables 55–57, which indicates the estimation issues on 3PL as discussed, for example, by Haberman (2005, 2006).

Table 49. Simulation Three-Parameter Logistic/Generalized Partial Credit Model: Mean Correlations of Item Parameter Estimates for Discrimination

Software	IRTPRO	flexMIRT	PARSCALE	MIRT
flexMIRT	1.000			
PARSCALE	.992	.992		
MIRT	.552	.552	.556	
True	.848	.848	.845	.458

Table 50. Simulation Three-Parameter Logistic/Generalized Partial Credit Model: Mean Correlations of Item Parameter Estimates for Difficulty/Location

Software	IRTPRO	flexMIRT	PARSCALE	MIRT
flexMIRT	1.000			
PARSCALE	.992	.992		
MIRT	.826	.826	.815	
True	.980	.980	.975	.807

Table 51. Simulation Three-Parameter Logistic/Generalized Partial Credit Model: Mean Correlations of Item Parameter Estimates for Guessing

Software	IRTPRO	flexMIRT	PARSCALE	MIRT
flexMIRT	1.000			
PARSCALE	.951	.951		
MIRT	.373	.373	.336	
True	.762	.762	.715	.336

Table 52. Simulation Three-Parameter Logistic/Generalized Partial Credit Model: Mean Differences of Item Parameter Estimates for Discrimination

Software	IRTPRO	flexMIRT	PARSCALE	MIRT
flexMIRT	.000			
PARSCALE	−.009	−.009		
MIRT	−.033	−.033	−.024	
True	−.018	−.018	−.009	.015

Note. Difference is calculated as column method minus row method.

Table 53. Simulation Three-Parameter Logistic/Generalized Partial Credit Model: Mean Differences of Item Parameter Estimates for Difficulty/Location

Software	IRTPRO	flexMIRT	PARSCALE	MIRT
flexMIRT	.002			
PARSCALE	.007	.005		
MIRT	−.171	−.173	−.178	
True	−.020	−.023	−.027	.150

Note. Difference is calculated as column method minus row method.

Table 54. Simulation Three-Parameter Logistic/Generalized Partial Credit Model: Mean Differences of Item Parameter Estimates for Guessing

Software	IRTPRO	flexMIRT	PARSCALE	MIRT
flexMIRT	.000			
PARSCALE	−.001	−.001		
MIRT	−.090	−.090	−.089	
True	−.014	−.014	−.013	.076

Note. Difference is calculated as column method minus row method.

Table 55. Simulation Three-Parameter Logistic/Generalized Partial Credit Model: Mean Root Mean Square Differences of Item Parameter Estimates for Discrimination

Software	IRTPRO	flexMIRT	PARSCALE	MIRT
flexMIRT	.000			
PARSCALE	.019	.019		
MIRT	.249	.249	.248	
True	.088	.088	.088	.265

Table 56. Simulation Three-Parameter Logistic/Generalized Partial Credit Model: Mean Root Mean Square Differences of Item Parameter Estimates for Difficulty/Location

Software	IRTPRO	flexMIRT	PARSCALE	MIRT
flexMIRT	.002			
PARSCALE	.067	.068		
MIRT	.505	.506	.523	
True	.154	.154	.174	.518

Table 57. Simulation Three-Parameter Logistic/Generalized Partial Credit Model: Mean Root Mean Square Differences of Item Parameter Estimates for Guessing

Software	IRTPRO	flexMIRT	PARSCALE	MIRT
flexMIRT	.000			
PARSCALE	.021	.021		
MIRT	.167	.167	.170	
True	.053	.053	.060	.165

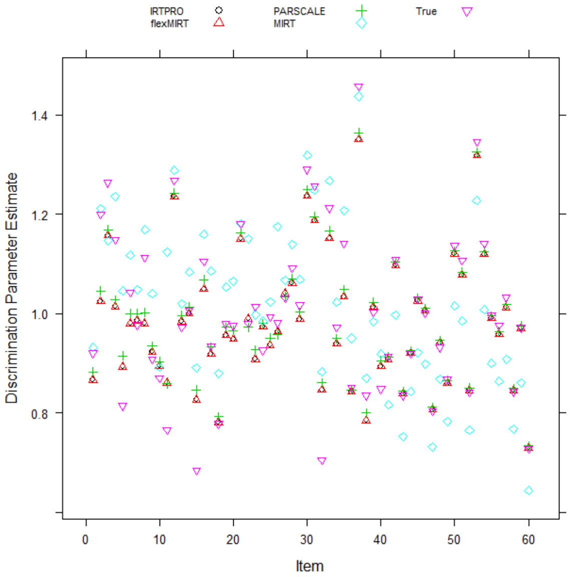

Figure 16. Simulation three-parameter logistic/generalized partial credit model: comparisons of mean discrimination parameter estimates.

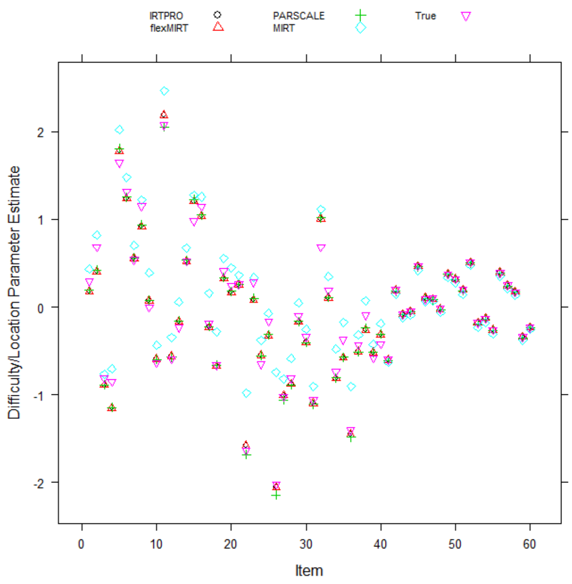

Figure 17. Simulation three-parameter logistic/generalized partial credit model: comparisons of mean difficulty/location parameter estimates.

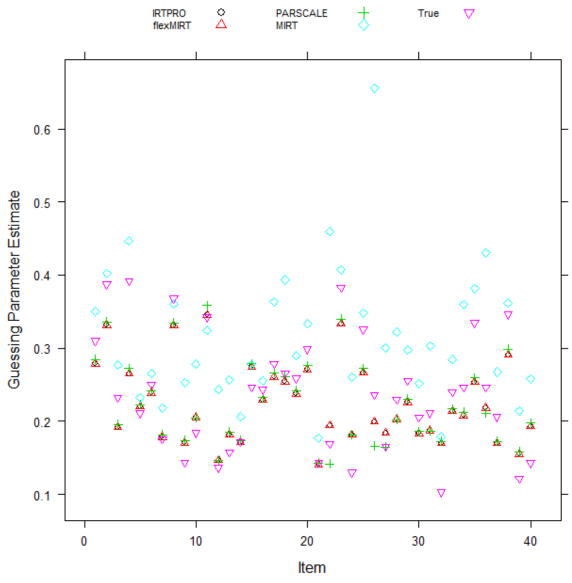

Figure 18. Simulation three-parameter logistic/generalized partial credit model: comparisons of mean guessing parameter estimates.

The estimation times listed in Table 58 show that the other three programs ran much faster than MIRT on simulated data sets. For Data A, flexMIRT ran the fastest; for Data B, PARSCALE and flexMIRT were the two fastest programs. MIRT was the slowest program on both real and simulated data sets.

Table 58. Three-Parameter Logistic/Generalized Partial Credit Model: Estimation Time (Seconds)

Software	Data A	Data B	Simulation [a]
IRTPRO	189	167	8
flexMIRT	171	143	7
PARSCALE	283	140	7
MIRT	540	374	21

[a] Mean running time over 30 replicates.

Summary

On all data sets, flexMIRT and IRTPRO produced the same estimates. It is not surprising considering that the two programs were developed by the same author and the same algorithm was implemented.

For 1PL/PCM and 2PL/GPCM on the two real data sets, MIRT's estimates differed the most from the other four programs' estimates, except for 2PL/GPCM on Data B. For 1PL/PCM and 2PL/GPCM on the simulated data sets, the estimates from all programs were very close and accurate in parameter recovery. Using priors for discrimination parameters in 2PL/GPCM helps achieve a smooth calibration.

For 3PL/GPCM, MIRT's estimates deviated the most from the other three programs' estimates on all data sets (mdltm was not included in the 3PL/GPCM comparisons). One of the reasons was that priors were used in the other three programs for the discrimination and difficulty/location parameters but not in MIRT. This indicates that item parameter priors have big impacts on the 3PL estimation. Although the other three programs had better parameter recovery than MIRT, for all programs, the parameter recovery on 3PL/GPCM, especially on the discrimination and guessing parameters, was much worse than that on 1PL/PCM and 2PL/GPCM. This signifies the estimation difficulty and possible identification issues when using the 3PL model (e.g., Haberman, 2005, 2006).

The mdltm program ran very fast on the real data (i.e., large data) but relatively slowly on simulated data sets (i.e., small data). IRTPRO, flexMIRT, and PARSCALE ran very fast on simulated data sets and also on the real data. MIRT was the slowest program on the real data sets and also quite slow on the simulated data sets.

Advantages and Disadvantages of Item Response Theory Software

Note that the following comments apply only to findings for these programs based on calibrations for 1PL, 2PL, 3PL, PCM, and GPCM and are not related to other capacities of these programs (e.g., estimating multidimensional IRT models):

IRTPRO

Advantages are that it is relatively well documented and user friendly, it has a fast processing time, and it provides many familiar fit statistics. Disadvantages are that it does not provide maximum likelihood estimates (MLEs) for ability parameters.

flexMIRT

Advantages are that it is relatively well documented and user friendly, it has a fast processing time, it provides MLEs for ability parameters, and it implements many familiar fit statistics. It has no obvious disadvantages.

PARSCALE

Advantages are that it is relatively well documented and user friendly, it has a fast processing time, it provides MLEs for ability parameters, and it is one of the most common IRT programs used in the educational measurement community in the past. Disadvantages are that it is no longer supported by the developer; the user needs to choose starting values carefully for calibrations that are hard to converge; likelihood values may jump around rather than monotonically increasing during iterations for difficultly converged calibrations; the user cannot fix discrimination parameters; ability MLEs of “999” do not distinguish between students of high and low ability; and nonextreme cases might be assigned “999” scores for reasons not explained in the manual.

mdltm

Advantages are its fast processing time on large data and that it provides relatively stable and accurate estimates for model parameters. Disadvantages are that the manual is concise, it does not allow for defining priors for item parameters, it does not provide MLEs for ability parameters, likelihood values may jump around rather than monotonically increasing during iterations for difficulty converged calibrations, and it cannot estimate the 3PL model.

MIRT

Advantages are that it has a solid theoretical foundation and provides stable and accurate estimates for model parameters, except for the 3PL model. Its disadvantages are that the manual is difficult to follow even for experienced psychometricians, the underlying models and estimation algorithm are hard to understand, the model fit statistics are unfamiliar, the syntax is very complicated, and the program is not user friendly. Furthermore, it does not provide MLEs for ability parameters; it is hard to define priors for item parameters; it has difficulty estimating the 3PL model; and its processing time is slow, especially when fit statistics for item pairs are requested (i.e., it may take hours for large data sets).

Recommendations for Large-Scale Assessments

When selecting IRT software for calibrating data from large-scale assessments, the following factors should be considered:

1. The program should produce robust and accurate estimations and require minimal effort to adjust the calibration setup for data that do not readily converge.
2. The program should be easily modified to address issues and accommodate new requests.
3. The program should be well documented and user friendly, and the implemented procedures (e.g., fit statistics) should be familiar to psychometricians.
4. The program should provide MLEs of ability, which are preferred for K–12 assessments. Bayesian scores are not preferable for individual student score reporting in K–12 assessments, as the concept of a student’s score being affected by the scores of the other students is not appropriate.

On the basis of the foregoing considerations, the commercially available software flexMIRT appears to meet most of the criteria because this program (a) is relatively user friendly and quickly processes large data sets; (b) is very flexible, implements many fit statistics and estimation methods for model parameters and standard errors of parameter estimates, and has all or most of the functionalities needed for analysis of large-scale assessment data; and (c) is constantly updated, well documented, and actively supported.

While flexMIRT meets most of the stated criteria, we recommend that psychometric and research staff at Educational Testing Service (ETS) develop in-house IRT software for calibration of large-scale assessment data. We also propose that this software, once fully vetted, be made available to groups outside of ETS. Advantages of developing in-house software include the ability to customize the program to meet specific requirements for analyzing large-scale assessment data, timely and in-depth technical support, and capability to promptly address issues and requests for new functionalities, as needed. Good candidates to use as a basis for the in-house IRT software are mdltm, MIRT, and the National Assessment of Educational Progress (NAEP) version of PARSCALE, if NAEP’s PARSCALE can be shared outside of ETS. MIRT is favored because, based on the author’s working experience and theoretical judgment, it can provide more robust estimates for complicated multidimensional IRT models compared to other multidimensional IRT programs.

References

Cai, L. (2017). *flexMIRT R version 3.51: Flexible multilevel multidimensional item analysis and test scoring* [Computer software]. Vector Psychometric Group.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO for Windows* [Computer software]. Scientific Software International.

Haberman, S. J. (2005). *Latent-class item response models* (Research Report No. RR-05-28). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2005.tb02005.x

Haberman, S. J. (2006). *An elementary test of the normal 2PL model against the normal 3PL alternative* (Research Report No. RR-06-14). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2006.tb02020.x

Haberman, S. J. (2013). *A general program for item-response analysis that employs the stabilized Newton–Raphson algorithm* (Research Report No. RR-13-32). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02339.x

Houts, C. R., & Cai, L. (2016). *flexMIRT user's manual version 3.5: Flexible multilevel multidimensional item analysis and test scoring.* Vector Psychometric Group.

Muraki, E., & Bock, R. D. (2003). *PARSCALE 4: IRT item analysis and test scoring for rating-scale data* [Computer software]. Scientific Software International.

Shin, H. J., Khorramdel, L., Xu, X., & von Davier, M. (2016). *Software for multidimensional discrete latent traits models—mdltm (draft manual)*. Educational Testing Service.

von Davier, M., & Xu, X. (2009). *Estimating latent structure models (including diagnostic classification models) with mdltm—A software for multidimensional discrete latent traits models* [Computer software]. Educational Testing Service.

Appendix. Default IRT Model Formulas Used in Computer Programs

Program	3PL/2PL	GPCM	Note
IRTPRO	$c_i + (1-c_i)\frac{\exp[a_i(\theta_j - f_i)]}{1+\exp[a_i(\theta_j - f_i)]}$	$\frac{\exp\left\{\sum_{h=0}^{m}[a_i(\theta_j - f_i + k_{ih})]\right\}}{\sum_{v=0}^{M_i-1}\exp\left\{\sum_{h=0}^{v}[a_i(\theta_j - f_i + k_{ih})]\right\}}$	For 1PL/PCM, a_i is fixed to 1 by the Constraints command. Item parameter estimates are saved in the PRM file.
flexMIRT	$c_i + (1-c_i)\frac{\exp[a_i(\theta_j - f_i)]}{1+\exp[a_i(\theta_j - f_i)]}$	$\frac{\exp\left\{\sum_{h=0}^{m}[a_i(\theta_j - f_i + k_{ih})]\right\}}{\sum_{v=0}^{M_i-1}\exp\left\{\sum_{h=0}^{v}[a_i(\theta_j - f_i + k_{ih})]\right\}}$	For 1PL/PCM, a_i is fixed to 1 by the fix and value commands. Item parameter estimates are provided in the IRT file. The item parameter estimates saved in the prm file correspond to the different formulas (see Houts & Cai, 2016, pp. 187–191, 213–124).
PARSCALE	$c_i + (1-c_i)\frac{\exp[1.7a_i(\theta_j - f_i)]}{1+\exp[1.7a_i(\theta_j - f_i)]}$	$\frac{\exp\left\{\sum_{h=0}^{m}[1.7a_i(\theta_j - f_i + k_{ih})]\right\}}{\sum_{v=0}^{M_i-1}\exp\left\{\sum_{h=0}^{v}[1.7a_i(\theta_j - f_i + k_{ih})]\right\}}$	The scale constant (i.e., 1.7) is set by SCALE option under the CALIB command; in the calibrations for the report the scale constant was set to 1. For 1PL/PCM, a common a_i is estimated across items; a_i cannot be fixed in the program. Item parameter estimates are saved in the PAR file.
mdltm	$\frac{\exp(1.7a_i\theta_j + z_i)}{1+\exp(1.7a_i\theta_j + z_i)}$	$\frac{\exp\left[\sum_{h=0}^{m}(1.7a_i\theta_j + q_{ih})\right]}{\sum_{v=0}^{M_i-1}\exp\left[\sum_{h=0}^{v}(1.7a_i\theta_j + q_{ih})\right]}$	For 1PL/PCM, 1.7a_i is fixed to 1 by the doslopes = false command. Parameter estimates are saved in the parameter output file (the items file).
MIRT	$\exp(g_i)/[1+\exp(g_i)] + \frac{\exp(a_i\theta_j + z_i)}{[1+\exp(g_i)][1+\exp(a_i\theta_j + z_i)]}$	$\frac{\exp\left[\sum_{h=0}^{m}(a_i\theta_j + q_{ih})\right]}{\sum_{v=0}^{M_i-1}\exp\left[\sum_{h=0}^{v}(a_i\theta_j + q_{ih})\right]}$	The logit-guessing parameter is a_i. For 1PL/PCM, s fixed to 1 by the rasch_slope_1 = T option under the &allskillspecs command. Item parameter estimates are saved in the parameter output file (unitparam).

Note. Different notations are used to show the variety of IRT model formulas implemented in the different programs. The 3PL formula is reduced to the 2PL formula by fixing c_i or $\exp(g_i)/[1+\exp(g_i)]$ to 0; and the 2PL/GPCM formula is reduced to the 1PL/PCM formula by fixing a_i or $1.7a_i$ (for mdltm) to 1 and estimating the standard deviation of the population distribution of θ_j or by estimating a common a_i across items and fixing the standard deviation of the population distribution of θ_j (PARSCALE implements the latter only for 1PL/PCM). The two formulations for 1PL/PCM are equivalent via parameter transformation. 1PL = one-parameter logistic, 2PL = two-parameter logistic, 3PL = three-parameter logistic, GPCM = generalized partial credit model, IRT = item response theory, PCM = partial credit model.