A Comparison of Parameter Recovery Using different Computer Programs and R-Packages in Estimating the Graded Response Model

Ou Zhang
Tianshu Pan

Pearson

April 2014

Paper Presented at the annual meeting of American Education Research Association at Philadelphia, PA

A Comparison of Parameter Recovery Using different Computer Programs and R-Packages in Estimating the Graded Response Model

Item response theory (IRT) has been widely used in educational and psychological tests. Due to its mathematical complicity, estimating IRT item and person parameters requires the use of specialized software, e.g., MULTILOG (Thissen, 2003), PARSCALE (Muraki & Bock, 2003), etc. Accurate recovery of model parameters from response data is a central problem in item response theory (IRT). In order to make informed decisions about which software package to use, researchers have conducted varies of studies to compare the item and personal parameter estimates from competing estimation software packages (Mislevy & Stocking, 1989; Yen, 1987; and many others).

There have been substantially less works comparing programs that estimate parameters for polytomous IRT models, e.g., Samejima's (1969) graded response model (GRM). Reise and Yu (1990) discussed parameter recovery in the graded response model using MULTILOG. Demars (2003) compared MULTILOG and PARSCALE on their recovery of item and trait parameters under the graded response and generalized partial credit item response models. Jurich and Goodman (2009) compared performance between PARSCALE and the freeware alternative IRT Command Language (Hanson, 2002) on accuracy of item and person parameters under dichotomous, polytomous, and mixed format conditions. Results show ICL to be equally effective as PARSCALE at parameter estimation under all conditions.

In recent years, several open-source and/or freeware programs for IRT, such as the R-packages 'ltm' (Latent Trait Models under IRT; Rizopoulos, 2006), and 'mirt'(Multidimensional Item Response Therory; Chalmers, 2012) in R language (R Development Core Team, 2007) become more and more popular. The 'ltm' package can be used to estimate the 1PL, 2PL, and

3PL for binary items and the graded-response model for polytomous items with a logit link. The 'mirt' package was created for estimating multiple multidimensional item response theory parameters models for dichotomous and polytomous items using maximum-likelihood methods, but it also can deal with unidimensional IRT models.

Accurate recovery of model parameters from response data is a central problem in item response theory (IRT). A prime concern in applying these IRT software packages is how well these programs can recover item and person parameters. However, little research exists in the literature on comparisons between newly developed IRT software packages (e.g., R-packages 'ltm' and 'mirt') and the present study is intended to fill the gap by conducting a comprehensive simulation study to evaluate the performance of various IRT software packages for the parameter recovery of GRM, and to find whether they can provide parameter estimates which is comparable with the commercial IRT programs, i.e., MULTILOG and PARSCALE. We will provide practical guidelines for choosing the most suitable approach for various practical situations.

Methods

Graded Response Model

The Graded Response Model is one of polytomous IRT models, specifically designed for ordinal manifest variables. This model was first discussed by Samejima (1969) and it is mainly used in cases where the assumption of ordinal levels of response options is plausible. The model is defined as follows:

$$\log\left(\frac{\mathrm{P}_{ik}}{1-\mathrm{P}_{ik}}\right) = a_i(\theta - b_{ik}) ,$$

where P_{ik} denotes the cumulative probability of a response in step/category k^{th} or lower to the i^{th} item, given the latent ability θ, a_i is the discrimination parameters of the i^{th} item, and b_{ik} stands for the k^{th} step parameter of the i^{th} item. However, the model is specified differently in PARSCALE so Childs and Chen's (1999) method is needed to obtain the comparable results across MULTILOG and PARSCALE.

Estimation Algorithm

The expectation-maximization (EM) algorithm is used in R-packages 'ltm', 'mirt', PARSCALE and MULTILOG. The fit of the two models is based on approximate marginal Maximum Likelihood, using the Gauss-Hermite quadrature rule for the approximation of the required integrals.

However, different optimization algorithms are used in these programs. For instance, R-package 'ltm' applied Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm. Meanwhile, PARSCALE use the EM algorithm and Newton-Gauss (Fisher scoring) iterations and MULTILOG uses EM algorithm only.

When estimating person parameters, MULTILOG uses Maximum a posteriori (MAP) and Maximum Likelihood (ML); Expected a posteriori (EAP), ML and weighted ML (Warm, 1989) are implemented in PARSCALE; empirical Bayes (EB), EAP, a multiple imputation like approach (Rizopoulos & Moustaki, 2008) or the Component Scores method (Bartholomew, 1984) are provided in 'ltm' package while 'mirt' package can use EAP, MAP, ML and weighted ML.

Because there are not completely the same estimation algorithms among the four programs, their default methods were applied in the simulation study for convenience.

Simulation

In this study Monte Carlo data are generated to study the capacities of those IRT software packages. The item and person parameters are simulated by following the method in Demars (2003).

Ten items were simulated with 5 response categories. The logs of the discrimination parameters were randomly selected from a normal distribution with a mean of -0.5 and standard deviation of 0.2.

The first category parameters for each item was drawn from a uniform distribution [-2, 1], and successive category parameters in the same item were 0.33 units apart.

Person parameters were drawn respectively from three distributions: normal [0, 1], uniform [-1.732, 1.732], and beta [2, 5.5]. The last one can produce a positively skewed distribution and was standardized to have mean 0 and standard deviation 1.

Three sample sizes, 500, 1000 and 5000, are used. One hundred replications are conducted.

Analyses

The accuracy of parameter estimation was quantified in this study using bias and root mean square error (RMSE). Bias is simply defined as average difference in true and estimated parameters across all people and items. Bias is a measure of any systematic errors in estimation. An estimate of bias is calculated for each replication of each condition, and an average bias for each condition in the simulation.

$$bias_{\lambda} = \frac{\sum_{j=1}^{n} \left(\hat{\lambda}_j - \lambda_j\right)}{n}$$

Where λ is the true value of a item or person parameter, $\hat{\lambda}$ is the estimated value of that parameter using those mentioned software, and n is the total instance of that type of parameter within a replication (i.e. sample size for ability, number of items for discrimination and difficulty, and 5 times the number of items for the category parameters).

RMSE is a measure of absolute accuracy in parameter estimation. RMSE is the square root of the average squared difference between estimated and true parameters.

$$RMSE_{\lambda} = \sqrt{\frac{\sum_{j=1}^{n}\left(\hat{\lambda}_j - \lambda_j\right)^2}{n}}$$

Where terms in the equation are defined as they are with bias.

Results

Tables 1-2 show respectively the bias and RMSE of discrimination, step difficulty parameters estimates of those software packages.

Table 1: Bias & RMSE of Discrimination Parameter Estimates

Sample size	Ability Distribution	Bias				RMSE			
		ltm	mirt	MULTILOG	PARSCALE	ltm	mirt	MULTILOG	PARSCALE
500	Normal	-0.05	-0.05	-0.05	-0.05	0.17	0.17	0.17	0.17
	Uniform	0.11	0.11	0.11	0.11	0.23	0.23	0.23	0.23
	Beta	0.02	0.02	0.02	0.02	0.16	0.16	0.16	0.16
1000	Normal	0.38	0.03	0.04	0.03	0.86	0.12	0.12	0.12
	Uniform	0.72	0.07	0.07	0.07	1.24	0.16	0.16	0.16
	Beta	0.36	-0.01	-0.01	-0.01	0.88	0.12	0.12	0.12
5000	Normal	0.40	0.00	0.00	0.00	0.95	0.05	0.05	0.05
	Uniform	0.64	0.06	0.07	0.06	1.18	0.11	0.11	0.11
	Beta	0.37	-0.02	-0.02	-0.02	0.87	0.07	0.07	0.07

Table 2: Bias & RMSE of Step Parameter Estimates

Sample size	Ability Distribution	Bias				RMSE			
		ltm	mirt	MULTILOG	PARSCALE	ltm	mirt	MULTILOG	PARSCALE
500	Normal	0.06	0.07	0.05	0.08	0.69	0.68	0.32	0.61
	Uniform	0.13	0.13	0.05	0.06	1.89	1.93	0.28	0.36
	Beta	-3.08	0.51	0.02	0.03	164.86	22.18	0.29	0.39
1000	Normal	0.02	0.05	0.05	0.05	1.47	0.19	0.18	0.19
	Uniform	0.11	0.03	0.03	0.04	1.82	0.30	0.18	0.30
	Beta	0.13	0.01	0.01	0.01	5.08	0.17	0.17	0.17
5000	Normal	-0.47	0.00	0.00	0.00	13.31	0.08	0.08	0.08
	Uniform	0.03	0.00	0.00	0.00	1.23	0.08	0.08	0.08

	Beta	0.04	0.01	0.01	0.01	0.89	0.08	0.08	0.08

By those results, it was found that the results of MULTILOG and PARSCALE are very similar, which is consistent with the founding of DeMars (2003), and R-packages 'ltm' and 'mirt' performed worse than NULTILOG and PARSCALE. R-package 'ltm' can provide discrimination parameter estimates similar to what the other three gave, whatever the ability distribution was, when sample size was 500. For step parameter estimates, they are similar only when sample size was 500, and the ability distribution is normal. However, R-package 'mirt' can provide the discrimination parameter estimates consistent with what MULTILOG and PARSCALE obtained. Its step parameter estimates were worse than MULTILOG and PARSCALE's only when sample size was 500 and the ability distribution is skewed (the beta distribution), or uniform-distributed. Generally, the results of 'mirt' are better than 'ltm' between the two R-packages.

By those results, it was found that the results of MULTILOG and PARSCALE are very similar, which is consistent with the founding of DeMars (2003), and R-packages 'ltm' and 'mirt' performed worse than NULTILOG and PARSCALE. R-package 'ltm' can provide discrimination parameter estimates similar to what the other three gave, whatever the ability distribution was, when sample size was 500. For step parameter estimates, they are similar only when sample size was 500, and the ability distribution is normal. However, R-package 'mirt' can provide the discrimination parameter estimates consistent with what MULTILOG and PARSCALE obtained. Its step parameter estimates were worse than MULTILOG and PARSCALE's only when sample size was 500 and the ability distribution is skewed (the beta distribution), or uniform-distributed. Generally, the results of 'mirt' are better than 'ltm' between the two R-packages.

Table 2 displays that some RMSE's of R-package 'ltm' are even larger than 10, which is unacceptable in practice. We checked its parameter estimates, and found that there are some extremely large step parameter estimates under some conditions. Although those weird estimates account for only 0.1%-0.72% of all step parameter estimates, their extremely large values made the overall bias and RMSE unacceptable. We also found that those weird estimates appeared on the items, where examinees' responses concentrated on smallest and largest categories. R-package 'mirt' also had the similar issue, but was not as serious as 'ltm'. If we excluded those weird estimates, the bias and RMSE looks acceptable although 'ltm's results were still larger than others. Table 3 shows the bias and RMSE of those software packages to estimate step parameters after those extreme values were excluded.

Table 3: Bias & RMSE of Step Parameter Estimates

Sample size	Ability Distribution	Bias				RMSE			
		ltm	mirt	MULTILOG	PARSCALE	ltm	mirt	MULTILOG	PARSCALE
500	Normal	-0.05	-0.05	-0.05	-0.07	0.51	0.51	0.32	0.60
	Uniform	-0.06	-0.06	-0.05	-0.06	0.33	0.33	0.27	0.28
	Beta	-0.04	-0.04	-0.02	-0.03	0.43	0.43	0.29	0.35
1000	Normal	-0.04	-0.05	-0.05	-0.05	0.48	0.19	0.18	0.18
	Uniform	-0.09	-0.03	-0.03	-0.04	0.80	0.30	0.17	0.29
	Beta	-0.01	-0.02	-0.01	-0.01	0.59	0.17	0.17	0.17
5000	Normal	0.02	0.00	0.00	0.00	0.73	0.08	0.08	0.08
	Uniform	-0.04	0.00	0.00	0.00	0.64	0.08	0.08	0.08
	Beta	-0.06	-0.01	-0.01	-0.01	0.45	0.08	0.08	0.08

Because R-package 'ltm' was unable to provide good estimates of category parameter of GRM, it also performed worst when estimating person's abilities, which is showed in Table 4. At the same time, however, we also found that R-package 'mirt' can provide ability estimates comparable with what MULTILOG and PARSCALE gave.

Table 4: Bias & RMSE of Person Parameter Estimates

Sample size	Ability Distribution	Bias				RMSE			
		ltm	mirt	MULTILOG	PARSCALE	ltm	mirt	MULTILOG	PARSCALE
500	Normal	-0.95	0.05	0.04	0.05	1.06	0.48	0.47	0.48
	Uniform	-0.96	0.04	0.03	0.04	1.07	0.47	0.48	0.47
	Beta	-1.00	0.00	-0.01	0.00	1.11	0.49	0.49	0.49
1000	Normal	-0.95	0.04	0.03	0.04	1.09	0.50	0.50	0.50
	Uniform	-0.95	0.03	0.02	0.03	1.09	0.47	0.47	0.47
	Beta	-0.98	0.00	-0.01	0.00	1.12	0.50	0.50	0.50
5000	Normal	-1.00	0.00	-0.01	0.00	1.13	0.49	0.49	0.49
	Uniform	-1.00	0.00	-0.02	0.00	1.13	0.47	0.47	0.47
	Beta	-0.97	0.00	-0.01	0.00	1.11	0.50	0.50	0.50

Conclusions

Although the current study provided evidences to show that those commercial and freeware IRT computer programs can give similar parameter estimates under some conditions, it was not able to answer the question which program is most accurate. The accuracy of the IRT program depends on many factors, e.g., estimation and optimization algorithms, the number of quadrature points or iteration cycles, etc. We only want to know whether the freeware IRT computer programs, namely, R-packages ‘ltm’ and ‘mirt’, can provide parameter estimates of GRM which is comparable with the commercial IRT software, e.g., MULTILOG and PARSCALE.

By this study, we do not suggest using R-package ‘ltm’ to fit GRM in practice although there is a study (Pan, 2012) to show that it can estimate the Rasch model as well as BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003). R-package ‘mirt’ is better than ‘ltm’ in estimating GRM, but it also needs to be used practically for GRM with caution, especially when sample size is smaller than 500 or probably 1000 and the ability distribution is not normal. It is because the package also shows some issue to estimate items parameters under this condition when responses concentrate on the lowest and top category of graded responses.

Reference

Childs, R. A., & Chen, W-H. (2011). Obtaining comparable item parameter estimates in MULTILOG and PARSCALE for two polytomous IRT models. *Applied Psychological Measurement*, 23, 371-379.

Chalmers, R. P., (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6).

DeMars, C. (2003). *Recovery of Graded Response and Partial Credit Parameters in MULTILOG and PARSCALE*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Hanson, B. A. (2002). *IRT Command Language (ICL)* [Computer software]. Available at http://www.b-a-h.com/software/irt/icl/index.html

Jurich, D., & Goodman, J. (2009). *A Comparison of IRT Parameter Recovery in Mixed Format Examinations Using PARSCALE and ICL.* Poster presented at the Annual meeting of Northeastern Educational Research Association.

Kieftenbeld, V., & Natesan, P.(2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov Chain Monte Carlo Estimation. *Applied Psychological Measurement*, 36(5), 399-419.

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.

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

Pan, T. (2012, April). *Comparison of Four Maximum Likelihood Methods in Estimating the Rasch Model*. Paper presented at the annual meeting of American Educational Research Association, Vancouver, Canada.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Reise, S. P., &Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement*, 27, 133-144.

Rizopoulos, D. (2006). ltm: an R package for latent variable modeling and Item Response Theory analyses. *Journal of Statistical Software*, 17(5).

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.

Thissen, D. (2003). *MULTILOG* 7: *Multiple categorical item analysis and test scoring using item response theory* [computer program]. Chicago, IL: Scientific Software International.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG 3: Item *analysis and test scoring with binary logistic models*. Chicago, IL: Scientific Software International. [Computer software].