2PL Model: Compare Generalized Linear Mixed Model with Latent Variable Model based IRT framework

Jihong Zhang, Terry Ackerman

University of Iowa

Yurou Wang

University of Alabama

2PL Model: Compare Generalized Linear Mixed Model with Latent Variable Model based IRT framework

Abstract

Fitting item response theory (IRT) models using the generalized mixed logistic regression model (GLMM) has become more popular in large-scale assessment because GLMM allows to combine complicated multilevel structure (i.e., students are nested in classrooms which are nested in schools) with IRT measurement models. However, the estimation accuracy of item parameters between these two models is not well examined. This study aimed to compare the estimation results of the GLMM based 2PL model (using the PLmixed R package) with the traditional IRT model (using flexMIRT software) under different sample sizes (N= 500, 1000, 5000) and test length (J = 15, 21) conditions. The simulation results showed that for both the GLMM-based method and the traditional method, item threshold's estimates had lower bias than item discrimination parameters. We also found that according to the simulation study, GLMM estimates via PLmixed had lower accuracy than traditional IRT modeling via flexMIRT for items with high discrimination.

Introduction

The link between the generalized mixed logistic regression model (GLMM) for binary data and latent variable model (LVM) have been well established for many years (Skrondal & Rabe-Hesketh, 2004). Many IRT models, among them the "standard" IRT models, fit nicely into GLMM framework. Some software packages such as SAS or R packages (i.e. lme4) allow users to estimate complex IRT models with an arbitrary number of nested or crossed random effects, making it useful for fitting, for example, the 1PL multilevel IRT (Doran et al., 2007) and random item IRT models (De Boeck et al., 2011). Fitting standard IRT models in GLMM framework can also be easily extended and adapted to more complicated scenarios. However, from our current experience, the application studies using GLMM-based IRT in measurement field have been limited for several reasons. The first is that few software programs allow researchers to estimate complex standard IRT models (2PL/3PL) within GLMM framework until Jeon and Rabe-Hesketh (2012) proposed the profile-likelihood estimation method which could fit the 2PL models or the 3PL models using *PLmixed* package. Second, the estimation accuracy of GLMM-based IRT have not been well examined. Standardized testing practitioners may face challenges when applying the GLMM estimation procedures to measurement data because it is still unclear whether GLMM-based estimates have better estimation accuracy of item parameters than IRT. Thus, this study is aimed to fill the void by performing a simulation study to compare the item parameter estimates generated by *PLmixed* and by *flexMIRT* software. Specifically, the main purpose of the current study is to compare the performance of the GLMM-based 2PL modeling with LVM-based traditional 2PL IRT modeling software (*flexMIRT*) with the measures of parameter estimation accuracy and standard errors. The results of this study could also help extend traditional psychometric framework to a broad multilevel frameworks which contributes to the applications in real large-scale scenario. This article was organized as followings. First, we discussed the association between GLMM with the standard 2PL IRT model. Second, we conducted a simulation study to examine the performance of GLMM based 2PL model in terms of parameters accuracy and recovery rate. Next, we present the comparison of estimates between GLMM 2PL model using the *PLmixed* package and the standard 2PL modeling using *flexMIRT*. Finally, the advantages and limitations of GLMM IRT modeling will be discussed. Some advice for future research will be provided.

Generalized Linear Mixed Model for 2PL

IRT are strongly associated with generalized linear mixed model in term of statistical form. One-parameter item response models (1PL-IRT) can be viewed (Eq.1) as two-level logistic regression models for a binary response Y_{ip} to item i by person p, nested in person, where the person abilities are considered as a random intercepts and item difficulties are considered as the regression coefficients of item dummy variables.

$$logit[Pr(Y_{ip}) = 1|\theta_{ip}] = \sum_{r=1}^{I} \boldsymbol{\beta_r d_r i} + \theta_p \tag{1}$$

Here d_{ri} is a dummy variable of item i with the value 1 when r = i and 0 otherwise, the diagnal of the matrix $\boldsymbol{-\beta_r}$ represents item difficulties, and θ_p represents person abilities. θ_p is a random effect (or latent variable in factor models) with $\theta_p \sim N(0, \sigma_p^2)$.

In addition, two-parameter item response models (2PL-IRT) models could be viewed (Eq. 2) as an extended version of generalized linear mixed models with the item discrimination parameters as a fixed part multiplied by the latent abilities as random part. The multi-level version of two-parameter model could be written as

$$logit[Pr(Y_{ip}) = 1|\theta_{ip}] = \sum_{r=1}^{I} \boldsymbol{\beta_r d_{ri}} + \theta_p \boldsymbol{\alpha_i} \tag{2}$$

Similar to the conventional factor model, for model identification, a factor loading (α_i) for one item is typically constrained to one or the variance of the latent variable (σ_p^2) is constrained to one.

Method

Simulation Study

To compare 2PL results in GLMM framework with IRT, a simulation study was conducted with three different levels of sample size (N = 500, 1,000, 5,000) and two different test length (J = 15, 21). The data generation process was conducted using the *mirt* package in R where discrimination values were randomly sampled from a low level (.4-.8), a middle level (.8-1.2) and a high level (1.2-1.6). The discrimination parameters were then multiplied by 1.702 for switching from ogive link to logit link. Three items were selected at the intercept values [-1.5, -.5, 0, .5, 1.5] for 15-item scale and [-2, -1.5, -.5, 0, .5, 1.5, 2] for 21-item scale. For each condition, 100 repetitions were performed.

Computer System

The analyses of profile-likelihood based generalized linear mixed model were performed using the PLmixed package in R under MacOS and the conventional 2PL model were performed using *flexMIRT* (Cai, L. ,2017) under Windows 10.

Comparisons

Estimates and true parameters values were compared using root mean squared differences (RMSE, see Eq.3) and local bias. Item parameters estimates of PLmixed and *flexMIRT* were compared in term of mean of estimates (item slopes and item intercepts) across all repetitions, standardized deviation of estimates and average standard errors of item parameters estimates. The RMSE of an estimator $\hat{\theta}$ with respect to the true parameter θ is defined as the square root of the mean square error:

$$RMSE(\hat{\theta}) = \sqrt{E((\hat{\theta} - \theta)^2)} \tag{3}$$

Results

Estimation Consistency of item parameters

Table 1 and Table 2 presented average estimates and standard deviations for item discrimination and threshold accordingly. Figure 1 summarized the pattern of average standard errors (SEs) and standard deviations (SDs) of item discrimination estimates for item 2, 8, 9 and 14 (red line: SE; blue line: SD; dashed line: PLmixed; solid line: flexMIRT). It showed that first, as the sample sizes increased, the standard deviations of item discrimination estimates between PLmixed and flexMIRT got closer. Second, across all conditions the item discrimination estimates by PLmixed had higher SDs than the estimates of flexMIRT which indicates that the flexMIRT's estimation process was more stable. Third, average standard errors for both estimates decrease when sample size got larger.

Figure 2 summarized the pattern of average standard errors (SEs) and standard deviations (SDs) of item threshold estimates for item 2, 8, 9 and 14. Generally speaking, item threshold estimates had similar standard deviation for PLmixed and flexMIRT. The items' average standard error suggested that PLmixed estimates (.054- .113 for N=5000) had relatively higher standard errors than flexMIRT's estimates (.040-.078 for N=5,000). As Table 2 shown, unlike item discrimination estimates, the estimates from PLmixed (.091-.192 for N=500) have very similar standard deviation (SD) or slightly smaller than the estimates from flexMIRT (.09-.243 for N=500) even for relatively smaller sample size (N=500). Standard errors (SE) of estimates by PLmixed (.0306-.0590 for N=5000) were slightly smaller than SEs of flexMIRT (.0311-.0711 for N=5000).

Estimation Accuracy of item parameters

Table 3 showed the RMSEs and Bias of item discrimination estimates by *PLmixed* and *flexMIRT*. The results suggested that as sample sizes increased, the RMSE of item discrimination estimates by PLmixed decreased (from .255-.711 in N=500 condition to

.154-.325 in N=1000, .053-.137 in N=5000). The flexMIRT's RMSEs decreased more than PLmixed as sample size increased (from .114-.275 in N=500 condition to .079-.168 in N=1000, .039-.081 in N=5000). In addition, higher item discrimination values resulted in higher Bias anf RMSEs for both estimation programs. Figure 3 and figure 4 summarized the bias and RMSEs of item discrimination estimations for item 2, 8, 9 and 14.

Table 4 showed the estimation accuracy measures of item threshold estimates by *PLmixed* and *flexMIRT*. It turned out that the estimates by *PLmixed* were very similar to those for flexMIRT except two items: item 11 and item 15. The pattern was the same for bias. As for the bias of item threshold estimates, because of the shrinkage, item 11's item threshold was underestimated but item 15's was overestimated

Conclusion

In summary, we found four main conclusions for this study: first, for both PLmixed package and flexMIRT, item thresholds' estimates were always more accurate than the estimated item discrimination parameters; second, for both mixed model and flexMIRT, the estimation error decreased as sample size increased; third, generally speaking, flexMIRT had better performance of estimating both item thresholds and item discrimination than PLmixed; last, for 2PL models, profile-likelihood based linear mixed modeling had lower estimation accuracy when estimating high-discrimination or low-discrimination items.

There are still some limitations in this study: first, the test length for this study was fixed to 15 and 21. More items in the test may also play an important role in the estimation process. It is also important to note that the two models were estimated by two different software. The differences of performance may come from the software rather than modeling. For future research, the estimation program should be controlled when estimating different models.

References

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, *39*(12), 1–28.

Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model: With the lme4 package. *Journal of Statistical Software*, *20*(2), 1–18.

Jeon, M., & Rabe-Hesketh, S. (2012). Profile-Likelihood Approach for Estimating Generalized Linear Mixed Models With Factor Structures. *Journal of Educational and Behavioral Statistics*, *37*(4), 518–542. https://doi.org/10.3102/1076998611417628

Skrondal, A., & Rabe-Hesketh, S. (2004, May 11). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. CRC Press.

	500 Sample Sizes						1000 Sample Sizes						5000 Sample Sizes					
	Mean		SD		Mean of SE		Mean		SD		Mean of SE		Mean		SD		Mean of SE	
Item	P	F	P	F	P	F	P	F	P	F	P	F	P	F	P	F	P	F
Item 1	0.45	0.44	-	0.12	-	0.12	0.45	0.44	-	0.08	-	0.09	0.45	0.45	-	0.04	-	0.04
Item 2	0.71	0.65	0.3296	0.14	0.2546	0.13	0.69	0.67	0.1561	0.09	0.1449	0.09	0.65	0.65	0.0550	0.04	0.0550	0.04
Item 3	0.68	0.63	0.2534	0.12	0.2367	0.13	0.66	0.64	0.1578	0.09	0.1403	0.09	0.64	0.65	0.0532	0.04	0.0545	0.04
Item 4	0.74	0.68	0.2900	0.11	0.2584	0.13	0.66	0.64	0.1542	0.09	0.1410	0.09	0.64	0.64	0.0615	0.04	0.0546	0.04
Item 5	0.82	0.75	0.3647	0.15	0.2953	0.15	0.76	0.74	0.1803	0.10	0.1628	0.10	0.74	0.75	0.0642	0.05	0.0627	0.05
Item 6	1.19	1.09	0.5739	0.19	0.4149	0.19	1.10	1.08	0.2095	0.12	0.2203	0.13	1.05	1.06	0.0787	0.06	0.0830	0.06
Item 7	0.86	0.79	0.3584	0.16	0.2929	0.14	0.82	0.80	0.1604	0.09	0.1655	0.10	0.79	0.80	0.0601	0.04	0.0634	0.04
Item 8	0.99	0.92	0.3878	0.14	0.3304	0.15	0.93	0.91	0.2014	0.11	0.1824	0.10	0.88	0.89	0.0710	0.05	0.0685	0.04
Item 9	1.17	1.08	0.4839	0.15	0.3926	0.17	1.09	1.07	0.2090	0.11	0.2106	0.11	1.05	1.07	0.0765	0.05	0.0787	0.05
Item 10	1.11	1.02	0.4648	0.17	0.3830	0.18	1.04	1.02	0.2202	0.12	0.2100	0.12	1.00	1.01	0.0775	0.05	0.0794	0.05
Item 11	1.64	1.55	0.6761	0.27	0.5556	0.27	1.52	1.52	0.3046	0.16	0.3036	0.18	1.44	1.49	0.1113	0.07	0.1126	0.08
Item 12	1.55	1.45	0.7017	0.19	0.5156	0.21	1.42	1.42	0.2768	0.14	0.2695	0.14	1.36	1.41	0.0865	0.06	0.0998	0.06
Item 13	1.44	1.36	0.5499	0.19	0.4704	0.19	1.32	1.32	0.2367	0.13	0.2485	0.13	1.27	1.31	0.1051	0.06	0.0932	0.06
Item 14	1.70	1.61	0.6983	0.23	0.5566	0.23	1.57	1.58	0.3271	0.16	0.3000	0.16	1.50	1.56	0.1197	0.08	0.1109	0.07
Item 15	1.44	1.34	0.5770	0.22	0.4871	0.23	1.35	1.35	0.2523	0.16	0.2679	0.16	1.27	1.31	0.1007	0.07	0.0999	0.07

Table 1

Comparisons of Estimated Item Discrimination a

	500 Sample Sizes						1000 Sample Sizes						5000 Sample Sizes					
	Mean		SD		Mean of SE		Mean		SD		Mean of SE		Mean		SD		Mean of SE	
Item	P	F	P	F	P	F	P	F	P	F	P	F	P	F	P	F	P	F
Item 1	0.66	0.67	0.09	0.09	0.10	0.10	0.67	0.68	0.08	0.08	0.07	0.07	0.66	0.67	0.03	0.03	0.03	0.03
Item 2	0.31	0.32	0.10	0.11	0.10	0.10	0.32	0.32	0.06	0.06	0.07	0.07	0.32	0.33	0.03	0.03	0.03	0.03
Item 3	0.00	0.00	0.10	0.10	0.10	0.10	0.01	0.01	0.06	0.07	0.07	0.07	0.00	0.00	0.03	0.03	0.03	0.03
Item 4	-0.31	-0.32	0.10	0.10	0.10	0.10	-0.32	-0.33	0.07	0.07	0.07	0.07	-0.32	-0.32	0.03	0.03	0.03	0.03
Item 5	-1.09	-1.12	0.11	0.12	0.11	0.12	-1.09	-1.12	0.07	0.08	0.08	0.09	-1.08	-1.11	0.03	0.04	0.04	0.04
Item 6	1.53	1.62	0.12	0.14	0.14	0.16	1.50	1.59	0.10	0.11	0.10	0.11	1.50	1.59	0.04	0.04	0.04	0.05
Item 7	0.40	0.41	0.10	0.11	0.10	0.11	0.38	0.40	0.07	0.07	0.07	0.07	0.39	0.41	0.03	0.03	0.03	0.03
Item 8	0.01	0.01	0.09	0.10	0.10	0.11	0.00	0.00	0.07	0.07	0.07	0.08	0.00	0.00	0.03	0.03	0.03	0.03
Item 9	-0.51	-0.54	0.12	0.13	0.11	0.12	-0.51	-0.54	0.07	0.07	0.08	0.08	-0.50	-0.53	0.04	0.04	0.04	0.04
Item 10	-1.47	-1.54	0.13	0.15	0.14	0.15	-1.43	-1.50	0.09	0.10	0.09	0.10	-1.43	-1.51	0.04	0.04	0.04	0.05
Item 11	2.08	2.28	0.19	0.25	0.19	0.24	2.06	2.25	0.12	0.15	0.13	0.16	2.05	2.24	0.06	0.07	0.06	0.07
Item 12	0.65	0.71	0.12	0.13	0.13	0.14	0.66	0.72	0.09	0.10	0.09	0.10	0.65	0.71	0.04	0.04	0.04	0.04
Item 13	-0.02	-0.02	0.11	0.12	0.12	0.12	0.01	0.02	0.09	0.09	0.08	0.09	0.01	0.01	0.04	0.04	0.04	0.04
Item 14	-0.73	-0.80	0.12	0.13	0.13	0.15	-0.72	-0.79	0.08	0.09	0.09	0.10	-0.71	-0.78	0.04	0.04	0.04	0.04
Item 15	-1.86	-2.00	0.16	0.18	0.17	0.20	-1.87	-2.02	0.12	0.14	0.12	0.14	-1.83	-1.97	0.05	0.06	0.05	0.06

Table 2

Comparisons of Estimated Item Threshold d

	500 Sample Sizes				1000 Sample Sizes				5000 Sample Sizes			
	RMSE		Bias		RMSE		Bias		RMSE		Bias	
Item	P	F	P	F	P	F	P	F	P	F	P	F
Item 1		0.12		-0.01		0.08		-0.01		0.04		0.00
Item 2	0.33	0.14	0.06	-0.00	0.16	0.09	0.04	0.02	0.06	0.04	-0.00	0.00
Item 3	0.25	0.12	0.04	-0.01	0.16	0.09	0.01	-0.01	0.05	0.04	-0.00	0.00
Item 4	0.30	0.11	0.09	0.03	0.15	0.09	0.01	-0.01	0.06	0.04	-0.01	-0.01
Item 5	0.37	0.15	0.08	0.01	0.18	0.10	0.02	-0.01	0.06	0.05	0.00	0.00
Item 6	0.59	0.19	0.13	0.03	0.21	0.12	0.04	0.02	0.08	0.06	-0.01	0.01
Item 7	0.36	0.16	0.06	-0.01	0.16	0.09	0.02	-0.00	0.06	0.04	-0.01	-0.00
Item 8	0.40	0.14	0.10	0.03	0.20	0.11	0.03	0.01	0.07	0.05	-0.02	-0.00
Item 9	0.49	0.15	0.10	0.02	0.21	0.11	0.02	0.01	0.08	0.05	-0.02	0.00
Item 10	0.47	0.17	0.10	0.02	0.22	0.12	0.04	0.01	0.08	0.05	-0.01	0.00
Item 11	0.69	0.28	0.16	0.07	0.31	0.17	0.04	0.04	0.12	0.07	-0.04	0.02
Item 12	0.71	0.19	0.14	0.03	0.28	0.14	0.01	0.00	0.10	0.06	-0.06	-0.01
Item 13	0.56	0.20	0.13	0.04	0.24	0.13	0.01	0.00	0.11	0.06	-0.04	-0.01
Item 14	0.71	0.23	0.13	0.04	0.33	0.16	0.00	0.01	0.14	0.08	-0.07	-0.01
Item 15	0.59	0.22	0.12	0.03	0.25	0.16	0.04	0.03	0.11	0.07	-0.04	-0.01

Table 3

Comparisons of Estimates and True Item Discrimination

	500 Sample Sizes				1000 Sample Sizes				5000 Sample Sizes			
	RMSE		Bias		RMSE		Bias		RMSE		Bias	
Item	P	F	P	F	P	F	P	F	P	F	P	F
Item 1	0.09	0.09	-0.01	0.00	0.08	0.08	0.01	0.01	0.03	0.03	-0.01	0.00
Item 2	0.10	0.11	-0.01	-0.00	0.06	0.06	-0.01	0.00	0.03	0.03	-0.01	0.00
Item 3	0.10	0.10	0.00	0.00	0.06	0.07	0.01	0.01	0.03	0.03	0.00	0.00
Item 4	0.10	0.10	0.01	0.00	0.07	0.07	0.00	-0.01	0.03	0.03	0.01	0.00
Item 5	0.11	0.11	0.03	-0.01	0.08	0.08	0.03	-0.01	0.05	0.04	0.04	0.00
Item 6	0.13	0.14	-0.05	0.03	0.13	0.11	-0.08	0.00	0.09	0.04	-0.08	0.00
Item 7	0.10	0.11	-0.00	0.01	0.07	0.07	-0.02	-0.00	0.03	0.03	-0.01	0.00
Item 8	0.09	0.10	0.01	0.01	0.07	0.07	0.00	0.00	0.03	0.03	0.00	0.00
Item 9	0.12	0.13	0.03	-0.00	0.07	0.07	0.03	-0.00	0.04	0.04	0.03	0.00
Item 10	0.14	0.15	0.04	-0.03	0.12	0.10	0.08	0.01	0.09	0.04	0.07	0.00
Item 11	0.23	0.26	-0.13	0.07	0.20	0.15	-0.16	0.03	0.17	0.07	-0.16	0.03
Item 12	0.13	0.13	-0.06	-0.00	0.10	0.10	-0.05	0.01	0.07	0.04	-0.06	-0.00
Item 13	0.11	0.12	-0.02	-0.02	0.09	0.09	0.01	0.02	0.04	0.04	0.01	0.01
Item 14	0.13	0.13	0.05	-0.02	0.10	0.09	0.07	-0.01	0.08	0.04	0.07	0.00
Item 15	0.19	0.18	0.12	-0.03	0.16	0.14	0.11	-0.04	0.15	0.06	0.15	0.01

Table 4

Comparisons of Estimates and True Item Threshold

Figure 1

Average SE/SD for Item Discrimination

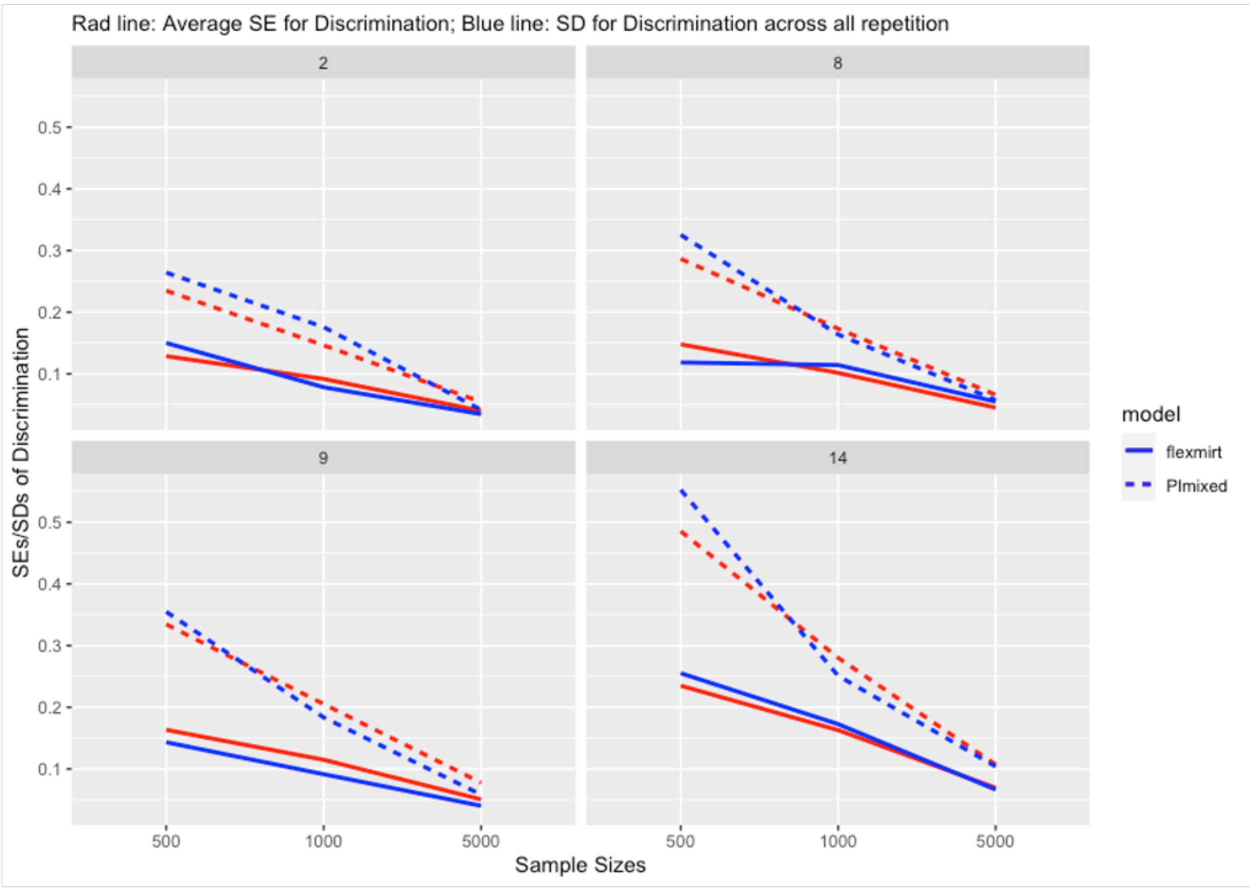


Figure 2

Average SE/SD for Item Threshold

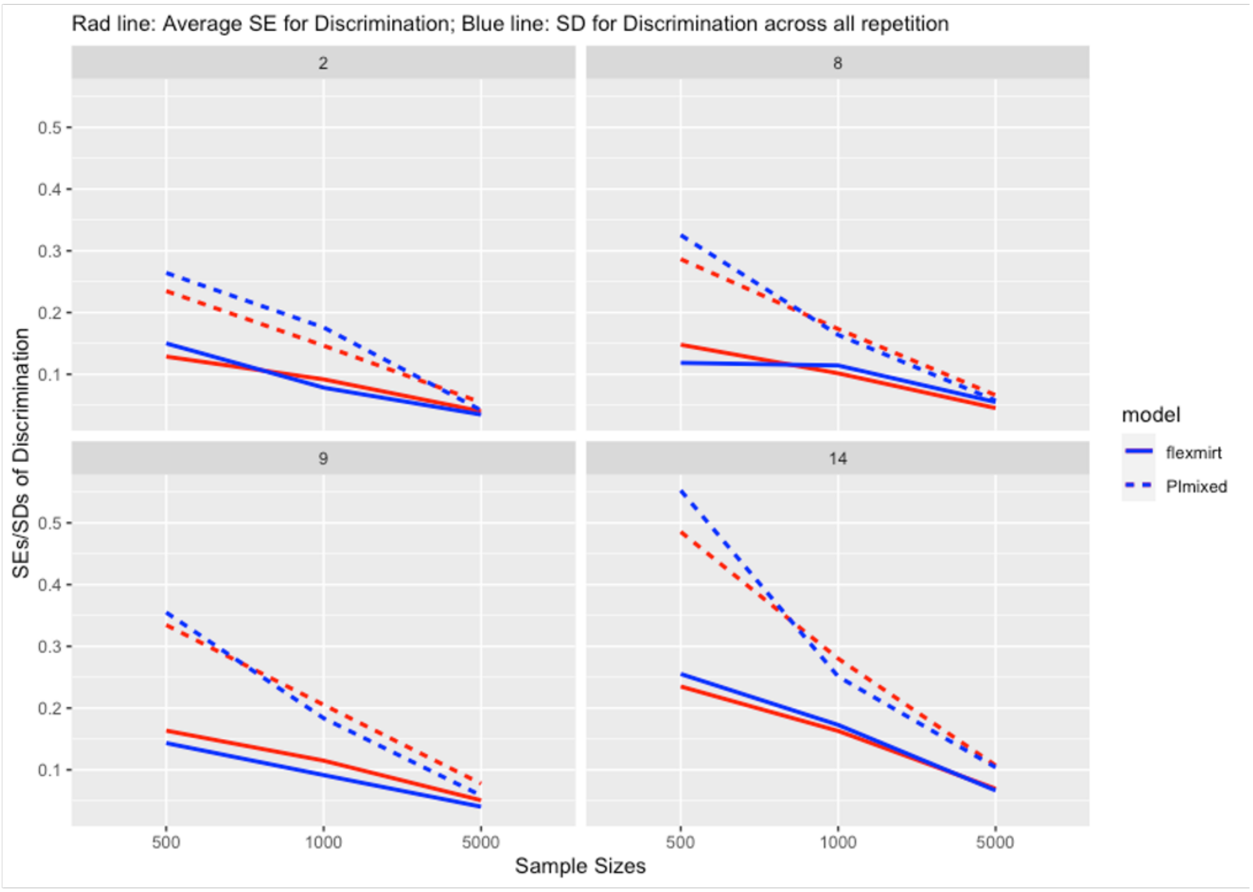


Figure 3

Bias for Item Discrimination

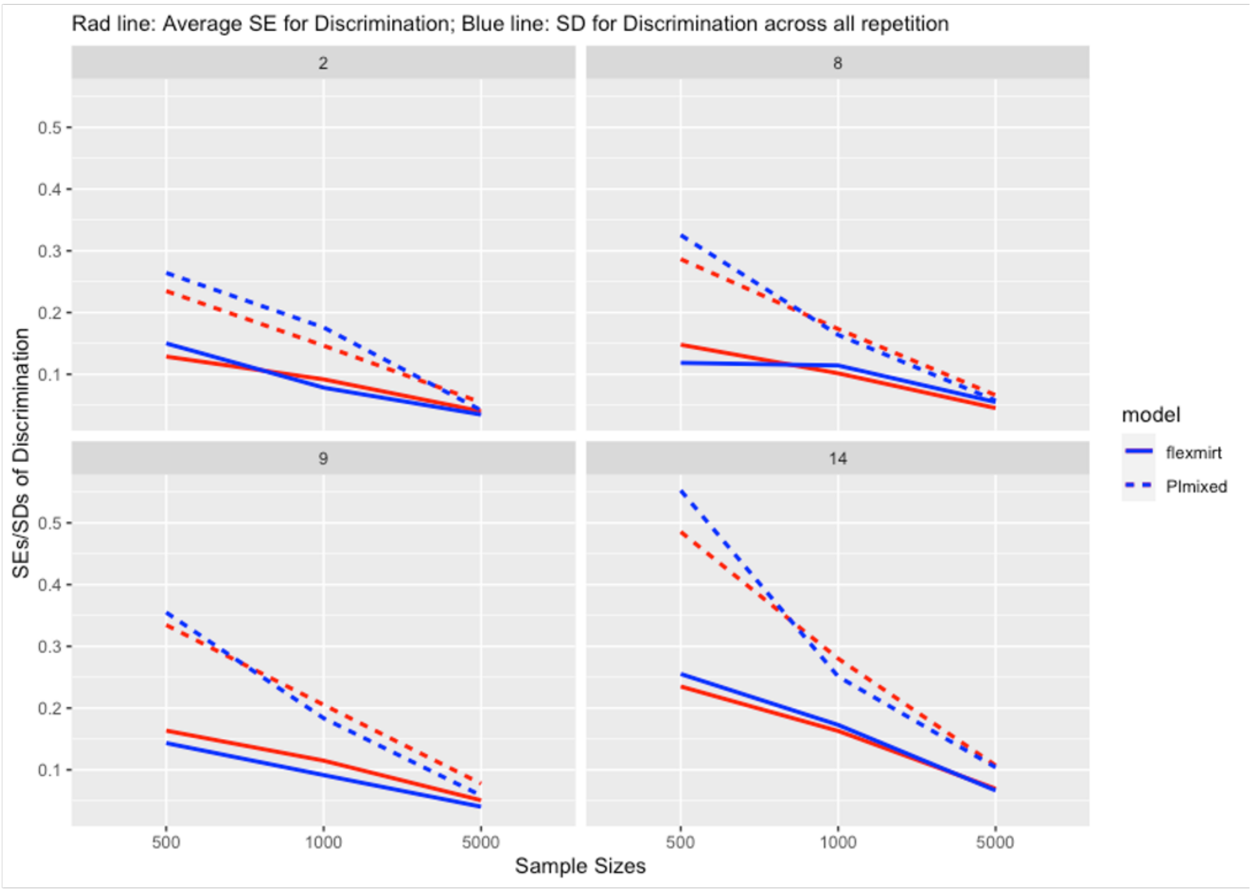
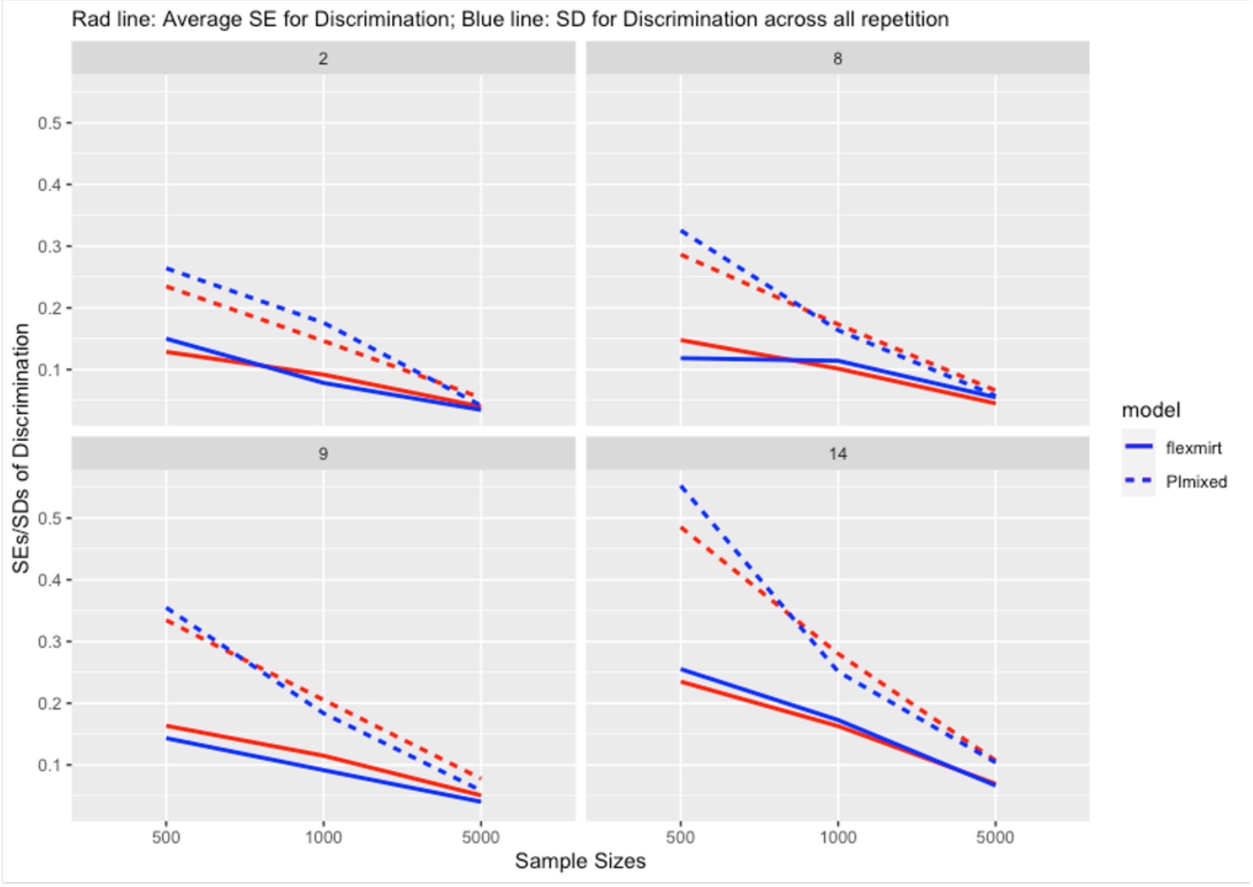


Figure 4

Bias for Item Threshold


Rad line: Average SE for Discrimination; Blue line: SD for Discrimination across all repetition
SEs/SDs of Discrimination
Sample Sizes
model
flexmirt
Plmixed