Technical Report:
Does It Matter Which IRT Software You Use? Yes.

Joy Wang

University of Minnesota

1/21/2018

ASSESSMENT SYSTEMS
FOR GOOD MEASURE™

Abstract

It is undeniable that psychometrics, like many tech-based industries, is moving in the direction of open-source and free software. In this study, we compare the performance of several software platforms for item response theory (IRT) analysis. A number of such programs are available, with a wide variation in estimation accuracy, documentation quality, and user-friendliness. The software compared includes: Xcalibre (Guyer & Thompson, 2014), the ltm package (Rizopoulos, 2017) and irtoys package (Partchev, Maris, & Hattori, 2017) in R (R core team, 2017), jMetrik (Meyer, 2017), IRTPRO (Cai, Thissen, & du Toit, 2017), flexMIRT (Houts & Cai, 2016), and Mplus (Muthen & Muthen, 2017).

The study utilized a monte-carlo simulation approach, with a parameter recovery evaluation. This allowed us to simulate data in eight experimental conditions, crossing two test lengths (30 and 50 items), two sample sizes (500 and 5000 examinees), and two item response models (3PL and GPCM), in addition to the five software programs.

Results indicate that the quality of the software can vary widely. For example, the root mean squared error of the IRT *b* parameter (item difficulty) ranged in one condition from 0.187 to 1.256. Given that the parameter typically only ranges from -3 to +3, being off target by more than 1 unit on many items is quite concerning. Unsurprisingly, differences and overall error were reduced with large sample size and longer test length.

Lastly, we also compare the requirements and use of the software programs, user manuals, types of output files, and functionality. We encourage the innovation in development of new tools, but caution their use in some cases.

Does it Matter Which IRT Software You Use? Yes

It is undeniable that psychometrics, like many tech-based industries, is moving in the direction of open-source and free software. This represents an important innovation in psychometrics and testing. But the old clichés of "buyer beware" and "you get what you pay for" remain as sound purchasing advice. In this study, we compare the performance of several software platforms for item response theory (IRT) analysis.

IRT is a powerful psychometric paradigm that improves many aspects of assessment, but requires highly specialized software to implement. A number of such programs are available, some of them free, but often with little documentation of quality.

The purpose of this monte-carlo simulation study was to compare the item parameter estimates and person parameter recovery form Xcalibre (version 4.2.2; Guyer & Thompson, 2014) and several other popular IRT software, including: ltm package (Rizopoulos, 2017) and irtoys package (Partchev, Maris, & Hattori, 2017) in R (R core team, 2017), jMetrik (Meyer, 2017), IRTPRO (Cai, Thissen, & du Toit, 2017), flexMIRT (Houts & Cai, 2016), and Mplus (Muthen & Muthen, 2017).

All these programs use marginal maximum likelihood estimation (MMLE; Bock & Aitkin, 1981) to estimate item parameters. Item parameters were then used to obtain ability estimates with one of three methods: maximum likelihood estimate (optimal choice in this study), maximum a posteriori (MAP; in IRTPRO), and empirical Bayesian estimate (in R GPCM scoring). The goal was to compare how well each program's estimates recovered the true item and person parameters. The estimation programs were compared under the three-parameter logistic model (3PL) for dichotomous data and generalized partial credit model (GPCM) for polytomous data.

Method

Eight response files with 2 test length (30 and 50 items), two sample size (500 and 5000 examinees), and two item response models (3PL and GPCM) were analyzed. The true item and person parameters were also provided for these eight responses files. This is an initial exploratory study, so no replication was used inside conditions; this is recommended for any future work.

Dichotomous response matrices were generated with the 3PL using the software WinGen (Han, 2007). The item response function for 3PL is

$$P_{ij} = c_j + \frac{1 - c_j}{1 + exp(-Da_j(\theta_i - b_j))}$$

where, P_{ij} denotes the probability of examinee i correctly answers item j. a_j, b_j, and c_j are item parameters of item j, and θ_i is person parameter for examinee i.

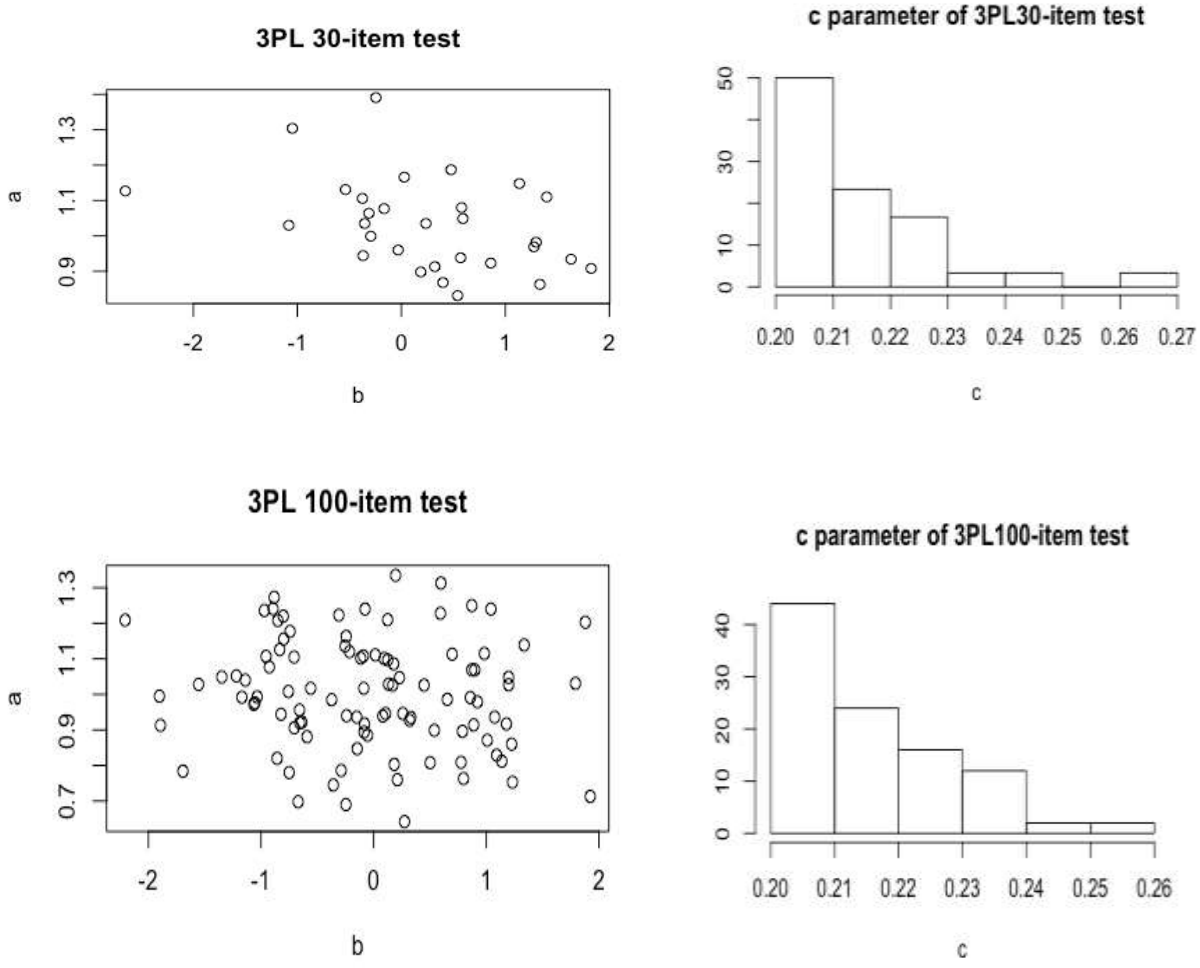
The item response function for GPCM used here is

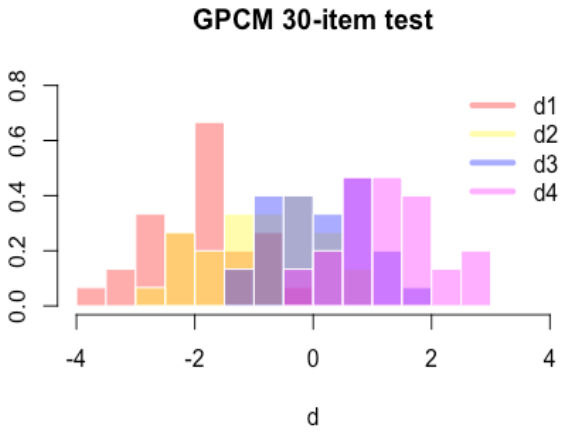
$$P_{ijk} = \frac{exp[\sum_{v=0}^{k} D\, a_j(\theta_i - d_{jv})]}{\sum_{l=0}^{m_j} exp[\sum_{v=0}^{l} D\, a_j(\theta_i - d_{jv})]}$$

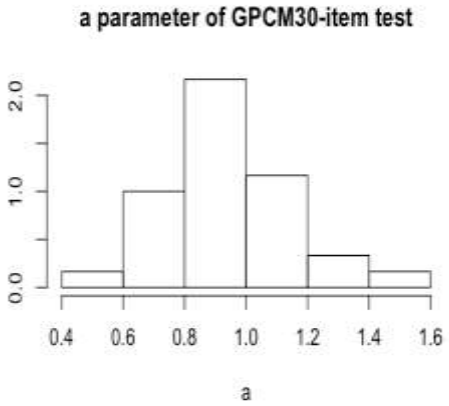
where, P_{ijk} is the probability examinee i endorse option k of item j. There are $m_j + 1$ options in item j, d_{jv} is the threshold of option v. As only m_j threshold parameters can be identified, d_{j0} is fixed to 0 for simplicity.

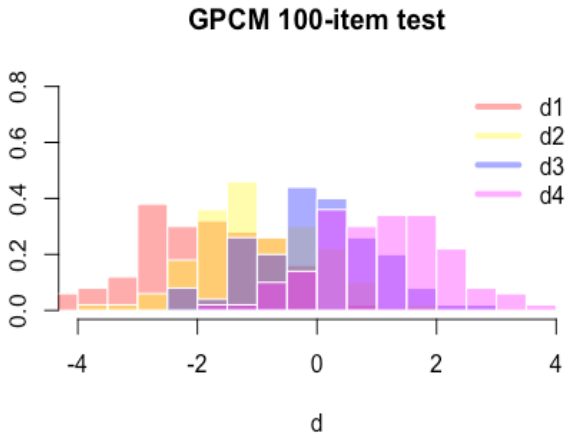
D = 1.702 for both item response models in this study.

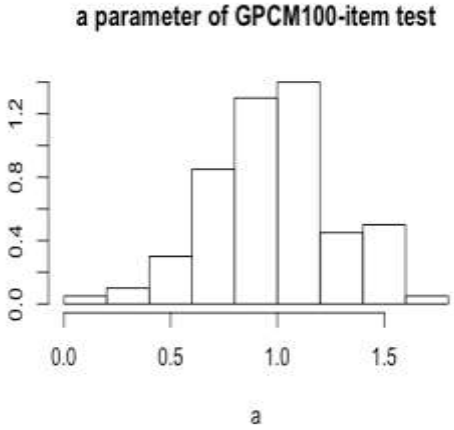
True item parameter distribution







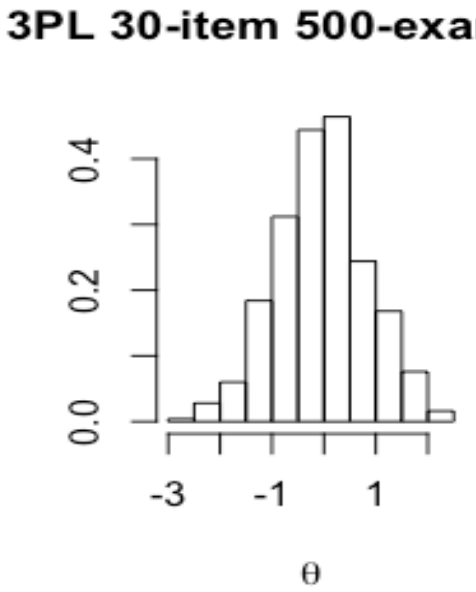


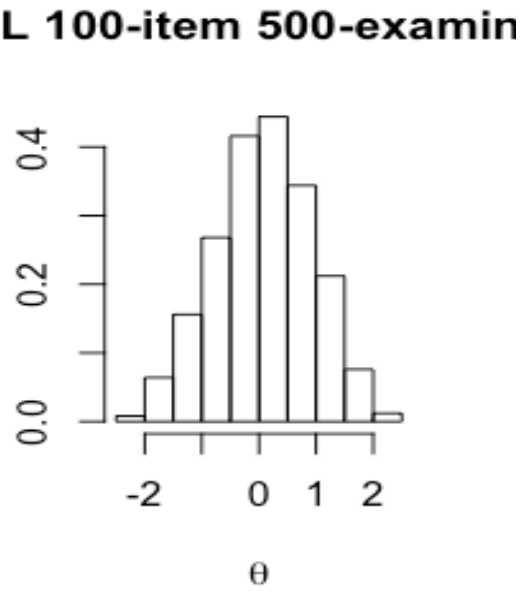


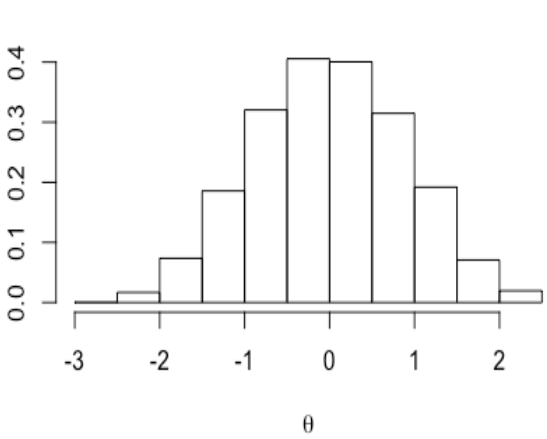
All GPCM items have five response categories, thus four threshold parameters need to be estimated in each item.

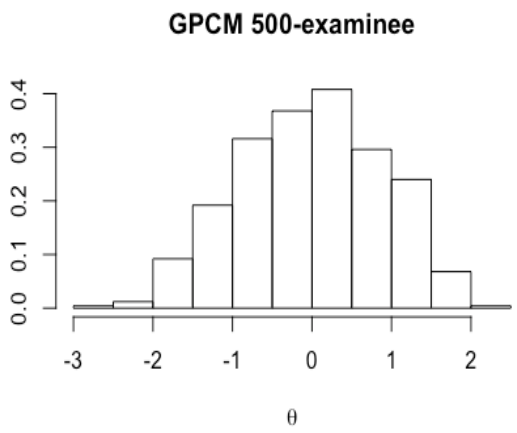
True person parameter distribution

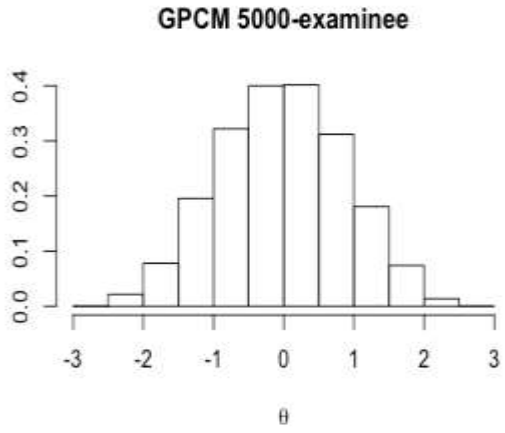
The 3PL 30-item 500-examinee condition does not share the same sample as the the 3PL 100-item 500-examinee condition. Except this pair, all the other pairs with same response model and sample size share the same sample.


3PL 30-item 500-examin
0.0
0.2
0.4
-3
-1
1
θ


3PL 100-item 500-examin
0.0
0.2
0.4
-2
0
1
2
θ


3PL 5000-examinee
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
θ





From these figures we can see, all θ distributions are normally distributed, with mean range from -0.04 to 0.08. and standard deviation from0.85 to 0.9. In this way, the standard normal distribution assumpition of θ used in MMLE holds.

Analysis with R

ltm package was used to do the item parameter and person parameter estimation for both 3PL and GPCM. For some reason, the "ltm" function does not work for 3PL, so "est" function in irtoys package was employed to call "ltm" fuction. The "gpcm" function in ltm package was used to calibrate GPCM data. The "mlebme" function in irtoys package was used to score examinees with the 3PL calibrated result. MLE was used. The "factor.score" function in ltm package was used to score examinees with the GPCM calibrated result. As MLE is not available in "factor.score", empirical bayes was used.

Analysis with jMetrik

"Transform–Advanced Item Scoring" was used to score the items. Then "Analyze–IRT Item Calibration (MMLE)" drop-down memu was used to do the item calibration and person parameter estimateion. MLE was employed in person parameter estimation. jMetrik allows user to choose between D=1 and D=1.7.

Error popped up when analyzing GPCM 30-item 5000-examinee, GPCM 100-item 500-examinee, and GPCM 100-item 5000-examinee. I consulted jMetrik support team, but haven't got reply.

There are a few items in the GPCM 30-item 500-examinee response that have only four options endorsed. Thus, for those items only three threshold parameters are estimable. However, jMetrik estimated all four threshold parameters. The "missing" parameters were assigned extreme values, e.g. -17 for lacking of option 1, and 8 for lacking of option 5.

Analysis with IRTPRO

"Analysis–Unidimensional IRT" was used to calibrate the items and scrore the examinees. As MLE is not available, MAP is used.

Analysis with flexMIRT

In the 3PL item parameter calibration, $\beta(1,4)$ was used as the prior of c, results in a mode of guessing around 0.2.

In the GPCM item parameter calibration, the items with only 4 categories endorsed have to be detected before calibration, and pointed out in the scripts. In the output, those missing parameters need to be suplimented manually with some string or number (0 in this study), otherwise the file can not be read properly by R.

MLE was used to do examinee scoring, which maximum and minimum setting to be 4 and -4.

Analysis with Mplus

In the 3PL item parameter calibration, $N(1.386,1)$ was used as the prior of c.

Mplus 8 can estimate GPCM model directly, and provide item parameters with classical GPCM parameterization. Before that, GPCM is estimated using the nominal model algorithm, then re-parameterize.

MLE was used to do examinee scoring.

Analysis with Xcalibre

The default prior for 3PL (a~N(0.8,0.3), b~N(0,1), c~N(0.25,0.025)) and GPCM (a~N(0.8,0.3)) item parameters were used. Xcalibre also allows user to choose between D=1 and D=1.7. MLE was used to do examinee scoring.

In the GPCM calibration, the control files (mark the items with four categories) as well as the response files (adjust response to 1-4) were required to adjust for the items with only four category endorsement.

Results

Results are collated into four tables: Bias and RMSE for the 3PL, then Bias and RMSE for the GPCM.

Bias and RMSE of 3PL item and person parameters

Table 1 Bias of 3PL parameter estimate

		R	jMetrik	IRTPRO	flexMIRT	Mplus	Xcalibre
30 item 500 examinee	a	-0.057	0.064	0.203	-0.117	-0.099	-0.193
	b	0.036	-0.079	0.072	-0.005	-0.003	-0.118
	c	-0.003	-0.029	0.006	-0.021	-0.022	-0.039
	θ	0.020	0.032	-0.086	0.016	-0.039	-0.013
30	a	0.108	0.107	0.128	0.106	0.107	0.000

item 5000 examinee	b	-0.038	-0.078	-0.028	-0.041	-0.038	-0.024
	c	-0.005	-0.033	-0.003	-0.008	-0.008	-0.023
	ϑ	0.011	0.041	-0.044	0.012	0.002	0.061
100 item 500 examinee	a	0.165	0.063	0.249	0.087	0.084	0.045
	b	0.173	-0.032	0.180	0.111	0.114	0.051
	c	0.017	-0.049	0.021	-0.001	-0.003	-0.030
	ϑ	0.093	0.071	0.075	0.080	0.089	0.098
100 item 5000 examinee	a	0.171	0.059	0.125	0.097	0.096	0.029
	b	0.050	-0.121	0.021	0.002	0.002	-0.034
	c	0.023	-0.045	0.007	-0.004	-0.004	-0.024
	ϑ	-0.006	-0.002	-0.011	0.002	0.003	0.016

Table 2 RMSE of 3PL parameter estimate

		R	jMetrik	IRTPRO	flexMIRT	Mplus	Xcalibre
30 item 500 examinee	a	0.503	0.266	0.306	0.822	0.724	0.672
	b	1.026	0.650	1.256	0.928	0.931	0.187
	c	0.118	0.072	0.043	0.079	0.076	0.047
	ϑ	0.701	0.784	0.386	0.690	0.384	0.637
30 item 5000 examinee	a	0.494	0.136	0.152	0.136	0.137	0.067
	b	0.175	0.158	0.156	0.153	0.152	0.072
	c	0.052	0.097	0.031	0.039	0.039	0.031
	ϑ	0.544	0.742	0.354	0.541	0.361	0.774
100 item 500 examinee	a	0.531	0.218	0.299	0.243	0.241	0.171
	b	0.451	0.270	0.395	0.274	0.273	0.172
	c	0.135	0.129	0.036	0.068	0.065	0.034
	ϑ	0.383	0.426	0.312	0.307	0.257	0.345
100 item 5000 examinee	a	0.520	0.100	0.140	0.119	0.118	0.072
	b	0.237	0.193	0.144	0.126	0.124	0.085
	c	0.061	0.109	0.028	0.046	0.046	0.028
	ϑ	0.316	0.488	0.224	0.291	0.227	0.344

Bias and RMSE of GPCM item and person parameters

Table 3 Bias of GPCM parameter estimate

		R	jMetrik	IRTPRO	flexMIRT	Mplus	Xcalibre
30 item 500 examinee	a	0.217	0.081	0.083	0.083	0.086	0.066
	d1	0.610	1.004	0.235	0.235	0.245	0.198
	d2	0.292	0.079	0.097	0.097	0.103	0.081
	d3	0.015	-0.015	0.002	0.002	0.003	0.002
	d4	-0.293	-0.328	-0.120	-0.120	-0.123	-0.093
	ϑ	0.044	-0.016	0.006	0.003	0.007	0.005
30 item 5000 examinee	a	0.207	NA	0.101	0.101	0.102	0.084
	d1	0.507	NA	0.211	0.209	0.211	0.170
	d2	0.163	NA	0.067	0.066	0.067	0.051
	d3	-0.081	NA	-0.041	-0.042	-0.040	-0.036
	d4	-0.364	NA	-0.176	-0.177	-0.177	-0.148
	ϑ	-0.040	NA	-0.023	-0.027	-0.024	-0.024
100 item 500 examinee	a	0.371	NA	0.080	0.079	0.099	0.098
	d1	1.378	NA	-0.033	-0.033	0.121	0.128
	d2	0.949	NA	-0.061	-0.060	0.078	0.087
	d3	0.417	NA	-0.107	-0.104	0.010	0.027
	d4	-0.222	NA	-0.225	-0.221	-0.132	-0.111
	ϑ	0.371	NA	-0.123	-0.124	-0.009	0.005
100 item 5000 examinee	a	0.361	NA	0.111	0.110	0.100	0.097
	d1	1.159	NA	0.233	0.231	0.209	0.203
	d2	0.711	NA	0.106	0.104	0.092	0.090
	d3	0.147	NA	-0.031	-0.032	-0.031	-0.028
	d4	-0.411	NA	-0.171	-0.172	-0.158	-0.150
	ϑ	0.094	NA	-0.024	-0.027	-0.026	-0.024

Table 4 RMSE of GPCM parameter estimate

		R	jMetrik	IRTPRO	flexMIRT	Mplus	Xcalibre
30 item	a	0.227	0.104	0.106	0.106	0.108	0.092
	d1	0.694	2.875	0.323	0.323	0.331	0.286

500 examinee	d2	0.433	0.178	0.188	0.188	0.194	0.169
	d3	0.237	0.108	0.109	0.109	0.111	0.095
	d4	0.406	1.108	0.198	0.198	0.202	0.171
	ϑ	1.449	0.205	0.181	0.205	0.186	0.193
30 item 5000 examinee	a	0.213	NA	0.106	0.106	0.106	0.089
	d1	0.648	NA	0.264	0.263	0.266	0.224
	d2	0.288	NA	0.128	0.127	0.128	0.104
	d3	0.220	NA	0.104	0.104	0.104	0.087
	d4	0.432	NA	0.218	0.219	0.219	0.185
	ϑ	1.413	NA	0.191	0.217	0.193	0.204
100 item 500 examinee	a	0.389	NA	0.104	0.103	0.119	0.121
	d1	1.517	NA	0.328	0.328	0.355	0.355
	d2	1.149	NA	0.185	0.184	0.204	0.202
	d3	0.716	NA	0.179	0.177	0.159	0.152
	d4	0.652	NA	0.316	0.314	0.270	0.246
	ϑ	1.704	NA	0.170	0.176	0.133	0.138
100 item 5000 examinee	a	0.378	NA	0.117	0.117	0.107	0.104
	d1	1.464	NA	0.295	0.293	0.269	0.262
	d2	1.164	NA	0.215	0.214	0.199	0.193
	d3	0.857	NA	0.160	0.160	0.148	0.140
	d4	0.714	NA	0.233	0.234	0.215	0.204
	ϑ	1.621	NA	0.145	0.156	0.139	0.144

Generally, bias of the item and person parameter estimate is pretty small (under 0.25). The only exception is the d1 and d4 estimate in GPCM using jMetrik. In GPCM, when certain options is not endorsed, the corresponding threshold parameter is not estimable. However, jMetrik assigns extreme estimate to these non-existed GPCM threshold parameters and causes the very large bias in these item parameter estimates. In the 30-item 500-examinee GPCM data, the inaccuracy only happened to some of the first and the last options, thus only d1 and d4 were affected by these extreme estimates.

Discussion

There was a surprising amount of variance in the performance of the software programs. In the 3PL case, the average RMSE for the b parameter was 0.13 for Xcalibre, 0.32 for jMetrik, 0.37 for MPlus and FlexMIRT, 0.47 for R and 0.49 for IRTPRO. For the GPCM, Xcalibre (0.18) and MPlus/FlextMIRT/IRTPRO (0.20) performed similarly for the location parameters, while the mean RMSE for the R package averaged a disturbing 0.68. However, IRTPRO has the smallest RMSE in 3PL person parameter estimate. This may due to the MAP method IRTPRO employs.

While replications are necessary for deeper interpretations, it is clear that all IRT calibration programs are not created equal. Moreover, we only considered parameter recovery accuracy here; the user-friendliness, documentation quality, support levels, and output quality can vary just as much. Some programs have a friendly user interface while others require programming expertise. Some have lengthy manuals with example files, and some do not. Some have on-call support, some have no support at all. Some have Word or HTML output, some are DOS-style text. Therefore, we highly recommend that the selection of an IRT software program for an organization involve investigation and thoughtfulness, rather than picking what appears to be the quickest or least expensive up front.

References

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 40,* 443-459.

Cai, L., Thissen, D., & du Toit, S.H.C. (2017). *IRTPRO 4.2 for Windows [Computer software].* Skokie, IL: Scientific Software International, Inc.

Guyer, R., & Thompson, N. A. (2014). *User's Manual for Xcalibre item response theory calibration software, version 4.2.2 and later.* Woodbury MN: Assessment Systems Corporation.

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement, 31*(5), 457-459.

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

Partchev, I., Maris, G., & Hattori, T. (2017). *irtyoys: A Collection of Functions Related to Item Response Theory (IRT).*

Meyer, J. P. (2017). *jMetrik, version 4.1.0.*

Meyer, J. P. (2014). *Applied measurement with jMetrik.* Routledge.

Mmuthen, L. K. & Muthen, B. O. (1998-2017). *Mplus User's Guide. Eight Edition.* Los Angeles, CA: Muthen & Muthen.

Rizopoulos, D. (2017). *ltm: Latent Trait Models under IRT. R package version 1.1-0.*

R Core Team (2013). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.