# A Comparison of IRT Parameter Recovery in Mixed Format Examinations Using PARSCALE and ICL

Daniel Jurich

Joshua Goodman

James Madison University

Poster presented at the Annual meeting of Northeastern Educational Research Association, October, 2009

# Abstract

The increased focus on educational measurement specifically through Item Response Theory (IRT) models has led to demand for accurate and reliable measurement practices. IRT models typically require specialized software for model estimation for which there are several commercial software packages available. However, these packages may be prohibitively expensive in certain situations where Hanson's IRT Command Language (ICL), a freeware IRT estimation package, may be more suitable. This study compared performance between PARSCALE and the freeware alternative ICL on accuracy of item and person parameters under dichotomous, polytomous, and mixed format conditions. Results show ICL to be equally as effective as PARSCALE at parameter estimation under all conditions.

## A Comparison of IRT Parameter Recovery of Mixed Format Examinations in PARSCALE and ICL

For the better part of a century, educational institutions, credentialing associations, and certification boards have relied, at least in part, on large-scale assessments to render high-stakes decisions about which students are promoted or accepted to colleges and which professionals are allowed to practice their chosen occupations. These tests can be composed of highly standardized selected-response items that are typically scored dichotomously, items that are scored for partial credit (e.g. short answer, essays, or more complex performance exercises), or a mixture of different items types—a trend that is increasingly common (Hambleton et al, 2000; Lane, 2005).

Regardless of the test format, Item Response Theory (IRT) plays a prominent role in most major testing programs. When employing IRT methods, assuming the fit between data and the selected IRT model(s) is adequate, the resulting person and item parameters have many desirable qualities[1]. However, estimating IRT item and person parameters requires the use of specialized software. The earliest IRT estimation programs produced for wide scale use were capable of estimating only relatively simple IRT models for dichotomously scored items. With the need to estimate complex models (e.g. polytomous IRT models, multidimensional IRT models), the desire for advanced IRT based analyses (e.g. DIF, equating), and the formulation of superior estimation procedures, new software packages were developed that encompassed these capabilities. In order to make informed decisions about which software package to use, researchers have conducted extensive simulation and real data studies that allow for the

[1] These features include, but are not limited to, the invariant properties of item and person parameters and the properties of the information function. See Lord (1980), Hambleton & Swaminathan (1985) or Hambleton, Swaminathan, & Rogers (1991) for detailed description of these and other features or IRT models.

comparison of the item and person parameter estimates from competing estimation software packages.

For the most part, comparisons between IRT software packages are limited to commercial package's performance in the estimation of parameters from dichotomous models (Mislevy & Stocking, 1989; Yen, 1987; and many others). There is substantially less research comparing programs that estimate parameters for polytomous IRT models (Reise & Yu, 1990; Ankenmann, & Stone, 1992; Childs & Chen, 1999). Demars (2002) compared the performance of PARSCALE (Muraki & Bock, 2002) and MULTILOG (Thissen, 1991) in recovering item parameters for the graded response (Samejima, 1969) and generalized partial credit models (Muraki, 1992). PARSCALE and MULTILOG performed equally well in recovering item and person parameters under a number of simulated conditions. Few studies have explored how various commercial software packages perform in the recovery of item and person parameters in test that contain both polytomous and dichotomous items (i.e. mixed format examinations).

In recent years, several high quality open-source freeware IRT software options have become available. Open-source and/or freeware programs offer a number of practical advantages over their commercial available counterparts. These programs allow users to examine and manipulate the underlying program code, creating virtually limitless opportunities in what can be accomplished. Commercial packages constrain users to limited models and options selected for inclusion by the developer. The ability to change and share changes can foster a powerful community where users can help each other find, modify, or even create the extensions they require to make the best use of the program[2]. For example, Mead, Morris, and Blitz (2007) describe how Hanson's (2000a) freeware program, IRT Control Language (ICL),

[2] The R statistical programming community is an excellent example of this type of community.

provides more flexibility and options in estimation methods than the commercial IRT program BILOG. However, as Mead and his colleagues warn, this added flexibility in ICL is often only practical for expert users who have both advanced programming skills and advanced theoretical knowledge of the workings of IRT models and estimation methods. Perhaps the most evident and most important advantage that programs like ICL have over commercially available programs is that freeware programs have no cost attached. With the price of a single license for a commercial software package costing hundreds of dollars, a freeware alternative allows access to IRT estimation programs to many practitioners, researchers, and students when cost would be an otherwise prohibitive factor.

Open source programs provide many benefits, but they are certainly not infallible. Unfettered and unregulated access to the programming language allows any user to create extensions to programs like ICL; thus, there is little to no quality control on these customizations. If such extensions are flawed, and applied in research or operational settings, decisions and conclusions made using the resulting parameters could be called into question. The fact that a company does not develop and support programs like ICL, results in no formalized technical support. Users with questions are constrained to asking for support though informal and unofficial channels, where answers may be delayed, confusing, or simply incorrect. This detail also limits the documentation available for ICL. While an in-depth manual of ICL is available, it is not edited and updated in the manner a commercial manual would be.

Ironically, the fact that commercially available IRT software programs are purchased for considerable sums allows for their major benefits. Scientific Software International (SSI), the company that markets some of some major IRT software packages (e.g. BILOG-MG, PARSCALE, MULTILOG, TESTFACT) has the ability to hire professional programmers and developers to create and troubleshoot the program on a daily basis. These packages continually release new and updated versions, often with improved features (though users are of course limited to only the options included in the package). Before any new release, quality control processes are employed to identify errors. Technical support is provided in any number of convenient formats (e.g. email, fax, or telephone), and assistance is essentially immediate. These benefits have made packages like PARSCALE widely accepted in the academic and business community.

While several high quality open-source IRT estimation programs are available (e.g. ICL, ETRIM, TESTGRAF), very little work is available to show how these programs function when compared to commercially available alternatives. In one rare example, Mead, Morris and Blitz (2007) conducted a comparison between BILOG's (Zimowski et al, 2003) and ICL in terms of dichotomous item parameter estimation and features provided by the program. The authors illustrate many of the benefits of commercial software discussed above when discussing the features, such as BILOG's practical interface, technical support and extensive manual. BILOG also provides more features and output for those not skilled in the programming language on which ICL is based. In terms of item parameter recovery, the authors compared item characteristic curves generated by ICL to BILOG across a 50-item test with set item parameters, manipulating sample sizes (100, 1,000 and 50,000). Results showed that ICL and BILOG perform comparably in estimating true item parameters. While this study shows great promise in the use of ICL, there is insufficient research on other types IRT models to accurately assess this claim. Little information is known about the estimation of polytomous models and even less so about mixed format tests which incorporate polytomous and dichotomous questions.

Using simulated data, the current study investigates how well ICL and PARSCALE recover the true item and person parameters, as well as how the two programs estimated parameters compare to each other. Comparisons will be made over tests that include just dichotomous items, just polytomous items, or both type of items. The study will also examine performance across various sample sizes, test lengths, and in the case of the mixed format examinations, the percent of polytomous items included.

PARSCALE (Muraki & Bock, 2001) is a commercially available IRT software package that provides numerous models for estimating item and person parameters, as well as other operationally convenient options[3]. PARSCALE has the ability to estimate both dichotomous and polytomous data. Dichotomous response data can be modeled through the one, two, or three parameter logistic models. PARSCALE also offers the generalized partial credit model, partial credit model (Masters, 1982), and graded response models for the treatment of polytomous items. Item parameters are estimated using marginal maximum likelihood estimation (MML) via the Expectation-Maximization (E-M) algorithm. Person ability parameters are estimated with either, straight maximum likelihood methods or using expected a posteriori (EAP) estimations. More details are available from the user manual available from SSI

ICL is an open source and freeware program for the estimation of IRT parameters. ICL's basic package consists of a single complied executable file capable of fitting the one, two, and three parameter models for dichotomous data and/or the generalized partial credit model and partial credit model for the modeling of polytomous data using MML-EM methods. Users can submit commands (as documented by the user manual) line by line to the program or submit a

[3] These include the ability to complete DIF studies, using data that results from multiple raters, and methods for combining subtest scores into a single score. The SSI manual for PARSCALE contains detailed descriptions of these options

prepared file containing a set of command language. The program has a number of flexible features such as, the ability to simulate data from any combination of the models, convenient data manipulation capabilities, and refined control over the E and M steps of the EM algorithm. The executable, source code, user manual, example data, and command language are all freely available online on the website dedicated to the software's author, Bradley Hanson (http://www.b-a-h.com/software/irt/icl/).

## Methods

### Data Simulation

Data were generated using a variety of conditions that simulate some typical testing situations. Total sample sizes, test length, percent of total items that are polytomous, and model choice were manipulated in the study. Conditions where polytomous items accounted for 100%, 50%, 25%, and 0% of the total number of items were considered in this study. The last two categories (0% polytomous and 25% polytomous items) were selected as tests that include all, or a large majority of, dichotomously scored items are common in high-stakes testing. The other two levels were included to demonstrate the capabilities of the software programs in more extreme, though less common, conditions; where half or all of the items on a test are polytomously scored. Reise and Yu (1990 suggested a minimum sample size of 500 when estimating polytomous models and hence, this value was chosen as the smallest sample size condition for this study. A second level of the sample size condition, 1500, was selected as this represents a sample size that should be more than adequate for stable parameter estimation. Four test different lengths were considered in the study as well. In the 0%, 25%, and 50% polytomous items conditions, test lengths of 40 and 80 items were considered. These test lengths represent typical test lengths that might be observed in practice and, for the most part, allow for realistic numbers of dichotomous and polytomous items to be included in each simulated test. As 40 and 80 polytomous items would not be practical or common in practice, two smaller test lengths of 10 and 20 items were used in the 100% polytomous item condition. These conditions were fully crossed with model choice for data generation.

To generate the data for this simulation study, the IRT two parameter logistic (2PL) model and the Generalized Partial Credit Model (GPCM; Muraki, 1992) were used for dichotomous and polytomous responses respectively. Constraining all of the *a* parameters to be equal across a set of items reduces the 2PL model to the one parameter logistic model (1PL) and the GPCM to the partial credit model (PCM). Both the one and two parameter models will be considered in this study. In mixed format test conditions, data will be generated either the one parameter models (1PL/PCM) or the two parameter models (2PL/GPCM). In all cases, the scaling constant, D, was set to 1.7.

True item parameters for the data simulation were generated using one of these two sets of models (i.e., the 1PL/PCM models or the 2PL/GPCM for dichotomous/polytomous items). The true *a* parameters are constrained to equal 1 for the 1PL and/or PCM model. True *a* parameters for the 2PL model were generated form a lognormal distribution with a mean of 0.0 and standard deviation of 0.1, yielding an expected mean *a* value of 1 and 99% of values falling between 0.7 and 1.3. Polytomous items tend to have lower discrimination in practice, and hence, for the GPCM conditions the *a* parameters were sampled form a lognormal distribution with a mean of -0.5 and a standard deviation of 0.1. This yielded a set of parameters with an expected mean of 0.6 with 99% of value falling between 0.4 and 0.8.

The category parameters, *d*, were generated using the same method for PCM and GPCM items. All items were simulated to have five score categories (scores ranging from 0-4). In many ways, this choice was arbitrary as the number of score categories for a polytomous item on a mixed format test would be highly variable. This choice did allow a convenient parallel to Demars (2006) study comparing MULTUILOG and PARSCALE as this study also used five score categories for each polytomous item. Additionally, In keeping with the constraint mentioned in the discussion of the ICRF of the GPCM above, $d_0$ was set to zero. The next two categories parameters were set by generating two numbers from a uniform distribution with a minimum value of -0.15 and maximum value of - 0.5. The first random number was added to $d_0$ to create $d_1$ and $d_2$ was created by adding both random numbers to $d_{0.}$ The last two categories parameters were created by generating two numbers from a uniform distribution with a minimum value of 0.15 and maximum value of 0.5. The first random number was added to $d_0$ to create $d_3$ and $d_4$ was created by adding both random numbers to $d_{0.}$ This method of item parameter generation allowed for a small number of items to have categories located at the extreme ends of the ability scale, as might be expected in real testing situations.

True abilities for simulated test-takers were generated from a standard normal distribution. The item and person parameters were used in conjunction with the IRF/IRCF functions presented in the equations above to generate 50 replications of data for each condition in the simulation using a script written in the R statistical package. The true model was then fit to each generated data set with PARSCALE and ICL. Options were set in both programs in such a way that differences observed could not be attributed to easy to set user options[4]. The package default values were used for any additional options (e.g. parameter starting values). EAP scoring

[4] A maximum of 100 E-M cycles was specified, the convergence criteria set at 0.001, and option we set in each program in such a way that the posterior distribution was estimated after each M step and scale to have a mean of 0 and standard deviation of 1 ( options "FREE=(0,1)" and "POSTERIOR" in PARSCALE and "-estim_dist**"** and "-scale_points" in ICL)

is available for both ICL and PARSCALE, and was used to produce the final person parameter estimates. Table 1 contains a summary of the all the simulation conditions included in the study.

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 is mathematically defined as:

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

where is the $\lambda$ is the true value of a item or person parameter, $\hat{\lambda}$ is the estimated value of that parameter using either ICL or PARSCALE, and *n* is the total instances of that type of parameter within a replication (i.e. sample size for ability, number of items for discrimination and difficulty, and 4 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, and is given by:

$$RMSE_{\lambda} = \sqrt{\frac{\sum_{j=1}^{n} \hat{\lambda}_j - \lambda_j \;{}^{2}}{n}}$$

where terms in the equation are defined as they are with bias. As with bias, RMSE is calculated for each parameter type in a replication, and an average for each condition is found within each condition.

Equally as important as the recovery of true parameters values in this study, is a comparison of estimation between the two IRT programs. It is unlikely that any IRT program will recover truth perfectly, and in real data situations, truth is never known. However, if two programs produce highly similar parameter estimates, the researcher or practitioner can feel confident in choosing either program as a sufficient estimator of true parameters. Similar to RMSE, the root mean squared difference (RMSD) is the square root of the average squared difference between parameter estimates of competing programs. It is defined by:

$$RMSD_{\lambda} = \sqrt{\frac{\sum_{j=1}^{n} \hat{\lambda}_{j,ICL} - \hat{\lambda}_{j,PSC} \quad {}^{2}}{n}}$$

where $\hat{\lambda}_{j,ICL}$ and $\hat{\lambda}_{j,PSC}$ are the estimates of any parameter estimated by ICL and PARSCALE respectively. The remaining terms are defined as they are in calculation of bias and RMSE. As with the previous two statistics, RMSD is calculated for each parameter type within each replication in the study, and an average for each condition is found for each condition.

## Results

The results of the simulation show that ICL performs comparably to PARSCALE under all conditions tested. Both programs estimate true item and person parameters with similar precision as RMSEs are virtually identical. RMSDs between ICL and PARSCALE are approximately zero for nearly all conditions, with the maximum value of .028. Low RMSDs suggest that when estimation errors are made the direction of these errors are consistent across programs. Both programs estimate truth in a manner consistent with the literature (Demars, 2002; Mead, Morris & Blitz 2007); parameter estimations become more stable given ideal conditions (i.e., adequate sample size for number of items given). Additionally, bias was near zero for all conditions, indicating that both programs produced parameter estimates that show no evidence of systematic error of estimation.

Under the dichotomous only model ICL and PARSCALE perform similarly across all conditions. As shown by table 2, RMSD's differ by less than a hundredth of a decimal place in all conditions but one. Polytomous model comparisons show similar results (see Tables 3 and 4). In both programs, the partial credit model estimated truth more accurately than the generalized partial credit model; however, there is less variability in program estimates and lower RMSDs in the GPCM.

When estimating the 25% polytomous mixed format condition, the two programs again show equivalent results. Dichotomous parameters were estimated accurately and congruently; RMSDs and RMSEs were small for each condition (see Tables 5 and 7). As with the polytomous only condition, the PCM results in lower RMSEs when compared to GPCM; however RMSDs under the GPCM do not exceed .007. Both programs struggle in estimating category parameters (b, b-d) in polytomous models especially under the GPCM. All of the trends hold true for the 50% polytomous item condition as well (see tables 6 and 8), though the increase in the number unique item parameter to be estimated produces slightly larger values of RMSE (though RMSD remain very similar to the 25% condition).

## Conclusions

Overall, both programs recover true item and person parameters in a manner consistent with other studies in the literature. As with these studies, precision and accuracy improve as samples size and test format conditions (e.g. test lengths that are not excessively long or too short) are sufficient for IRT estimation. More importantly for this study, the two programs produced nearly identical estimates across all conditions. This should provide some sense of assurance that decisions about which software package to use should be made on considerations other than accuracy in estimation (e.g. cost, ease of use, institutional availability, need for customization).

Of course, simulation studies often demonstrate performance under ideal situations. In this case, the true IRT model was known and fit can be assumed nearly perfectly. Future studies should use these IRT programs to fit these models to real data. No comparison to truth would be available; however, the estimates under each program could be compared. Absent from this study was the very popular three parameter logistic model. This model was excluded only to limit the scope of this first initial study. Future studies comparing ICL to other programs (including programs like PARSCALE, MULTILOG, and BILOG-MG) must include this model as well given its prevalence in standardized testing. Conditions should also be explored that expanding current simulation variables to include more levels (i.e., sample sizes, test lengths, and percent of polytomous items). In addition, as mixed format tests may contain polytomous items that have variable numbers of score categories, the number of score categories could also be manipulated in further studies. This will not only help create other realistic testing situations, but will also test the ability of each program to perform in situations where there are more or less item parameters overall.

## References

Andrich, D. (1978). Application of a psychometric rating scale model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2,* 581-594.

Ankenmann, R.D., & Stone, C.A. (1992, April). *A Monte Carlo study of marginal maximum likelihood parameter estimates for the graded model.* Paper presented at the annual meeting of the National Council for Measurement in Education, San Francisco, CA. (ERIC Document Reproduction Service No. ED34189)

Childs, R.A., & Chen, W.-H. (1999). Software Note: Obtaining comparable item parameters estimates in MULTILOG and PARSCALE for two polytomous IRT models. *Applied Psychological Measurement, 21,* 89-90.

Choi, S. W., Cook, K. F., and Dodd, B. G. (1997). Parameter recovery for the partial credit model using MULTILOG. *Journal of Outcome Meausrment, 1*, 114-142.

Demars, C. E. (2002). *Recovery of Graded Response and Partial Credit Parameters in MULTILOG and PARSCALE.* Paper presented at the annual meeting of the National Council for Measurement in Education, Chicago, IL. (ERIC Document Reproduction Service No. ED476138)

Hambleton, R. K. (2000). Advances in Performance Assessment Methodolgy. *Applied Psychological Measurement, 24*(4), 291-293.

Hambleton, R., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Boston, MA: Kluwer-Nijhoff Publishing.

Hambleton, R.., & Swaminathan, H. & Rogers, J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications.

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

Hanson, B. A. (2002b). *IRT Command Language*. Computer software manual. [Available athttp://www.b-a-h.com/software/irt/icl/icl_manual.pdf]

Lane, S. (2005). *Status and future directions for performance assessments in education*. Paper presented at the Annual Meeting of the American Educational Research Association.

Lord, F. (1980). *Applications of Item Repsonse Theory to Practical Testing Problems*. Hillsdale, NJ: Lawerence Erlbaum AssociatesMislevy,R, & Stocking, M. (1989). A consumer guide to LOGIST and BILOG. *Applied Psychological Measurement, 13,* 57-75.

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

Mead, A.D., Morris, S.B. & Blitz, D.L. (2007). *Open-source IRT: A comparison of BILOG-MG and ICL features and item parameter recovery.* Unpublished manuscript

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG *Applied Psychological Measurement, 13,* 57-75.

Muraki, E. (1992). Ageneralized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.

Muraki, E. & Bock, D. (2002) PARSCLE 4.1 Computer program.  Chicago: Scientific Software International, Inc.

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

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

Thissen, D. (1991). *MULTILOG user's guide: Multiple categorical item analysis and test scoring using item response theory (Version 6.0)* [Softw, are manual]. Chicago: Scientific Software.

Yen, W. M. (1987).  A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika, 52*, 275-291.

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. [Computer software.]

Table 1: Simulation Conditions

| Model | % Polytomous Items | Sample Size | Test Length | Replications |
|---|---|---|---|---|
| 1PL and/or PCM | 0% | 1500 | 40 | 50 |
| | | | 80 | 50 |
| | | 500 | 40 | 50 |
| | | | 80 | 50 |
| | 25% | 1500 | 40 | 50 |
| | | | 80 | 50 |
| | | 500 | 40 | 50 |
| | | | 80 | 50 |
| | 50% | 1500 | 40 | 50 |
| | | | 80 | 50 |
| | | 500 | 40 | 50 |
| | | | 80 | 50 |
| | 100% | 1500 | 10 | 50 |
| | | | 20 | 50 |
| | | 500 | 10 | 50 |
| | | | 20 | 50 |
| 2PL and/or GPCM | 0% | 1500 | 40 | 50 |
| | | | 80 | 50 |
| | | 500 | 40 | 50 |
| | | | 80 | 50 |
| | 25% | 1500 | 40 | 50 |
| | | | 80 | 50 |
| | | 500 | 40 | 50 |
| | | | 80 | 50 |
| | 50% | 1500 | 40 | 50 |
| | | | 80 | 50 |
| | | 500 | 40 | 50 |
| | | | 80 | 50 |
| | 100% | 1500 | 10 | 50 |
| | | | 20 | 50 |
| | | 500 | 10 | 50 |
| | | | 20 | 50 |

Table 2: Average Bias, RMSE, & RMSD for Conditions with 100% Dichotomous Items

| Model | Sample | Items | Parameter | ICL | | PARSCALE | | |
|---|---|---|---|---|---|---|---|---|
| | | | | *Bias* | *RMSE* | *Bias* | *RMSE* | *RMSD* |
| 1PL | 500 | 40 | a | -- | -- | -- | -- | -- |
| | | | b | 0.006 | 0.067 | 0.006 | 0.067 | 0.001 |
| | | | Theta | -0.002 | 0.249 | -0.001 | 0.249 | 0.002 |
| | | 80 | a | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | | | b | -0.003 | 0.065 | 0.000 | 0.065 | 0.003 |
| | | | Theta | -0.002 | 0.191 | 0.001 | 0.191 | 0.004 |
| | 1500 | 40 | a | -- | -- | -- | -- | -- |
| | | | b | 0.036 | 0.056 | 0.038 | 0.057 | 0.002 |
| | | | Theta | 0.038 | 0.264 | 0.040 | 0.264 | 0.003 |
| | | 80 | a | -- | -- | -- | -- | -- |
| | | | b | 0.031 | 0.051 | 0.033 | 0.052 | 0.001 |
| | | | Theta | 0.038 | 0.196 | 0.039 | 0.196 | 0.003 |
| 2PL | 500 | 40 | a | -0.006 | 0.111 | 0.005 | 0.115 | 0.013 |
| | | | b | 0.008 | 0.073 | 0.008 | 0.072 | 0.005 |
| | | | Theta | -0.001 | 0.251 | 0.000 | 0.251 | 0.008 |
| | | 80 | a | -0.008 | 0.103 | 0.000 | 0.105 | 0.009 |
| | | | b | -0.002 | 0.074 | -0.002 | 0.074 | 0.003 |
| | | | Theta | 0.001 | 0.189 | 0.000 | 0.190 | 0.007 |
| | 1500 | 40 | a | 0.020 | 0.074 | 0.025 | 0.076 | 0.005 |
| | | | b | 0.037 | 0.058 | 0.037 | 0.058 | 0.002 |
| | | | Theta | 0.039 | 0.265 | 0.040 | 0.266 | 0.004 |
| | | 80 | a | 0.024 | 0.065 | 0.027 | 0.066 | 0.003 |
| | | | b | 0.034 | 0.055 | 0.034 | 0.056 | 0.001 |
| | | | Theta | 0.039 | 0.197 | 0.039 | 0.197 | 0.004 |

Table 3: Average Bias, RMSE, & RMSD for Conditions with 100% PCM Items

| Sample | Items | Parameter | ICL | | PARSCALE | | |
|---|---|---|---|---|---|---|---|
| | | | *Bias* | *RMSE* | *Bias* | *RMSE* | *RMSD* |
| 500 | 10 | a | -- | -- | -- | -- | -- |
| | | b | 0.013 | 0.067 | 0.013 | 0.069 | 0.012 |
| | | d | -- | 0.160 | -- | 0.160 | 0.014 |
| | | b-d | 0.013 | 0.174 | 0.013 | 0.175 | 0.019 |
| | | Theta | -0.002 | 0.337 | -0.001 | 0.338 | 0.011 |
| | 20 | a | -- | -- | -- | -- | -- |
| | | b | 0.011 | 0.054 | 0.011 | 0.053 | 0.012 |
| | | d | -- | 0.157 | -- | 0.157 | 0.013 |
| | | b-d | 0.011 | 0.167 | 0.011 | 0.166 | 0.018 |
| | | Theta | -0.001 | 0.241 | -0.001 | 0.241 | 0.012 |
| 1500 | 10 | a | -- | -- | -- | -- | -- |
| | | b | 0.047 | 0.058 | 0.049 | 0.059 | 0.018 |
| | | d | -- | 0.098 | -- | 0.098 | 0.022 |
| | | b-d | 0.047 | 0.114 | 0.049 | 0.114 | 0.028 |
| | | Theta | 0.037 | 0.338 | 0.038 | 0.339 | 0.019 |
| | 20 | a | -- | -- | -- | -- | -- |
| | | b | 0.042 | 0.056 | 0.043 | 0.056 | 0.008 |
| | | d | -- | 0.093 | -- | 0.093 | 0.009 |
| | | b-d | 0.042 | 0.109 | 0.043 | 0.109 | 0.012 |
| | | Theta | 0.039 | 0.249 | 0.039 | 0.249 | 0.006 |

Table 4: Average Bias, RMSE, & RMSD for Conditions with 100% GPCM Items

| Sample | Items | Parameter | ICL | | PARSCALE | | |
|---|---|---|---|---|---|---|---|
| | | | *Bias* | *RMSE* | *Bias* | *RMSE* | *RMSD* |
| 500 | 10 | a | 0.020 | 0.091 | 0.015 | 0.091 | 0.006 |
| | | b | -0.002 | 0.111 | -0.003 | 0.113 | 0.007 |
| | | d | -- | 0.253 | -- | 0.255 | 0.004 |
| | | b-d | -0.002 | 0.279 | -0.003 | 0.282 | 0.008 |
| | | Theta | -0.001 | 0.437 | -0.001 | 0.437 | 0.004 |
| | 20 | a | -0.006 | 0.071 | -0.008 | 0.072 | 0.003 |
| | | b | 0.021 | 0.090 | 0.021 | 0.091 | 0.004 |
| | | d | -- | 0.247 | -- | 0.248 | 0.002 |
| | | b-d | 0.021 | 0.264 | 0.021 | 0.266 | 0.004 |
| | | Theta | 0.000 | 0.321 | -0.001 | 0.321 | 0.002 |
| 1500 | 10 | a | 0.021 | 0.036 | 0.019 | 0.035 | 0.003 |
| | | b | 0.053 | 0.075 | 0.053 | 0.074 | 0.003 |
| | | d | -- | 0.141 | -- | 0.141 | 0.002 |
| | | b-d | 0.053 | 0.160 | 0.053 | 0.161 | 0.003 |
| | | Theta | 0.037 | 0.439 | 0.037 | 0.439 | 0.002 |
| | 20 | a | 0.010 | 0.037 | 0.007 | 0.036 | 0.003 |
| | | b | 0.047 | 0.080 | 0.046 | 0.079 | 0.003 |
| | | d | -- | 0.134 | -- | 0.135 | 0.001 |
| | | b-d | 0.047 | 0.157 | 0.046 | 0.157 | 0.004 |
| | | Theta | 0.039 | 0.331 | 0.039 | 0.330 | 0.003 |

Table 5: Average Bias, RMSE & RMSD with 25% PCM Items and 75% 1PL Model

| Sample | Items | Parameters | | ICL | | PARSCALE | | |
|---|---|---|---|---|---|---|---|---|
| | | | | *Bias* | *RMSE* | *Bias* | *RMSE* | *RMSD* |
| 500 | 40 | Dichotomous Items | a | -- | -- | -- | -- | -- |
| | | | b | 0.021 | 0.105 | 0.021 | 0.105 | 0.001 |
| | | Polytomous Items | a | -- | -- | -- | -- | -- |
| | | | b | 0.009 | 0.052 | 0.009 | 0.052 | 0.008 |
| | | | b-d | 0.009 | 0.180 | 0.009 | 0.181 | 0.011 |
| | | | d | -- | 0.172 | -- | 0.173 | 0.008 |
| | | All Items | theta | -0.002 | 0.258 | -0.002 | 0.258 | 0.003 |
| | 80 | Dichotomous Items | a | -- | -- | -- | -- | -- |
| | | | b | 0.001 | 0.094 | 0.001 | 0.094 | 0.002 |
| | | Polytomous Items | a | -- | -- | -- | -- | -- |
| | | | b | 0.004 | 0.076 | 0.004 | 0.076 | 0.010 |
| | | | b-d | 0.004 | 0.188 | 0.004 | 0.189 | 0.014 |
| | | | d | -- | 0.172 | -- | 0.173 | 0.010 |
| | | All Items | theta | -0.003 | 0.187 | -0.003 | 0.187 | 0.003 |
| 1500 | 40 | Dichotomous Items | a | -- | -- | -- | -- | -- |
| | | | b | 0.033 | 0.069 | 0.032 | 0.068 | 0.002 |
| | | Polytomous Items | a | -- | -- | -- | -- | -- |
| | | | b | 0.044 | 0.058 | 0.045 | 0.056 | 0.018 |
| | | | b-d | 0.044 | 0.111 | 0.045 | 0.109 | 0.025 |
| | | | d | -- | 0.094 | -- | 0.092 | 0.018 |
| | | All Items | theta | 0.038 | 0.272 | 0.038 | 0.272 | 0.006 |
| | 80 | Dichotomous Items | a | -- | -- | -- | -- | -- |
| | | | b | 0.034 | 0.067 | 0.033 | 0.067 | 0.002 |
| | | Polytomous Items | a | -- | -- | -- | -- | -- |
| | | | b | 0.038 | 0.061 | 0.039 | 0.060 | 0.014 |
| | | | b-d | 0.038 | 0.120 | 0.039 | 0.118 | 0.019 |
| | | | d | -- | 0.103 | -- | 0.102 | 0.013 |
| | | All Items | theta | 0.038 | 0.191 | 0.038 | 0.190 | 0.003 |

Table 6: Average Bias, RMSE & RMSD with 50% PCM Items and 50% 1PL Model

| Sample | Items | Parameters | | ICL | | PARSCALE | | |
|---|---|---|---|---|---|---|---|---|
| | | | | *Bias* | *RMSE* | *Bias* | *RMSE* | *RMSD* |
| 500 | 40 | Dichotomous Items | a | -- | -- | -- | -- | -- |
| | | | b | 0.041 | 0.097 | 0.041 | 0.098 | 0.006 |
| | | Polytomous Items | a | -- | -- | -- | -- | -- |
| | | | b | 0.005 | 0.070 | 0.004 | 0.069 | 0.011 |
| | | | b-d | 0.005 | 0.189 | 0.004 | 0.188 | 0.015 |
| | | | d | -- | 0.175 | -- | 0.175 | 0.011 |
| | | All Items | theta | -0.002 | 0.216 | -0.003 | 0.216 | 0.006 |
| | 80 | Dichotomous Items | a | -- | -- | -- | -- | -- |
| | | | b | -0.006 | 0.094 | -0.006 | 0.095 | 0.006 |
| | | Polytomous Items | a | -- | -- | -- | -- | -- |
| | | | b | 0.002 | 0.073 | 0.002 | 0.073 | 0.011 |
| | | | b-d | 0.002 | 0.189 | 0.002 | 0.189 | 0.016 |
| | | | d | -- | 0.174 | -- | 0.174 | 0.011 |
| | | All Items | theta | -0.003 | 0.160 | -0.003 | 0.160 | 0.006 |
| 1500 | 40 | Dichotomous Items | a | -- | -- | -- | -- | -- |
| | | | b | 0.041 | 0.067 | 0.040 | 0.066 | 0.005 |
| | | Polytomous Items | a | -- | -- | -- | -- | -- |
| | | | b | 0.037 | 0.054 | 0.037 | 0.053 | 0.013 |
| | | | b-d | 0.037 | 0.109 | 0.037 | 0.108 | 0.018 |
| | | | d | -- | 0.094 | -- | 0.094 | 0.012 |
| | | All Items | theta | 0.038 | 0.229 | 0.038 | 0.229 | 0.005 |
| | 80 | Dichotomous Items | a | -- | -- | -- | -- | -- |
| | | | b | 0.033 | 0.064 | 0.033 | 0.063 | 0.005 |
| | | Polytomous Items | a | -- | -- | -- | -- | -- |
| | | | b | 0.037 | 0.060 | 0.037 | 0.059 | 0.015 |
| | | | b-d | 0.037 | 0.122 | 0.037 | 0.120 | 0.019 |
| | | | d | -- | 0.106 | -- | 0.105 | 0.013 |
| | | All Items | theta | 0.038 | 0.164 | 0.038 | 0.164 | 0.005 |

Table 7: Average Bias, RMSE & RMSD with 25% GPCM Items and 75% 2PL Model

| Sample | Items | Parameters | | ICL | | PARSCALE | | |
|---|---|---|---|---|---|---|---|---|
| | | | | *Bias* | *RMSE* | *Bias* | *RMSE* | *RMSD* |
| 500 | 40 | Dichotomous Items | a | -0.023 | 0.108 | -0.018 | 0.110 | 0.006 |
| | | | b | 0.008 | 0.074 | 0.008 | 0.074 | 0.002 |
| | | Polytomous Items | a | 0.015 | 0.073 | 0.015 | 0.074 | 0.002 |
| | | | b | 0.018 | 0.088 | 0.018 | 0.090 | 0.003 |
| | | | b-d | 0.018 | 0.264 | 0.018 | 0.266 | 0.004 |
| | | | d | -- | 0.248 | -- | 0.249 | 0.002 |
| | | All Items | theta | -0.002 | 0.241 | -0.002 | 0.241 | 0.002 |
| | 80 | Dichotomous Items | a | -0.009 | 0.099 | -0.004 | 0.101 | 0.007 |
| | | | b | -0.001 | 0.071 | -0.001 | 0.071 | 0.002 |
| | | Polytomous Items | a | 0.006 | 0.077 | 0.005 | 0.077 | 0.002 |
| | | | b | 0.008 | 0.131 | 0.008 | 0.132 | 0.003 |
| | | | b-d | 0.008 | 0.309 | 0.008 | 0.311 | 0.004 |
| | | | d | -- | 0.278 | -- | 0.280 | 0.003 |
| | | All Items | theta | -0.001 | 0.182 | -0.001 | 0.182 | 0.004 |
| 1500 | 40 | Dichotomous Items | a | 0.021 | 0.075 | 0.022 | 0.076 | 0.002 |
| | | | b | 0.033 | 0.056 | 0.033 | 0.056 | 0.001 |
| | | Polytomous Items | a | 0.021 | 0.042 | 0.021 | 0.042 | 0.001 |
| | | | b | 0.046 | 0.075 | 0.046 | 0.075 | 0.001 |
| | | | b-d | 0.046 | 0.148 | 0.046 | 0.148 | 0.001 |
| | | | d | -- | 0.126 | -- | 0.126 | 0.001 |
| | | All Items | theta | 0.039 | 0.258 | 0.039 | 0.258 | 0.001 |
| | 80 | Dichotomous Items | a | 0.024 | 0.063 | 0.025 | 0.064 | 0.002 |
| | | | b | 0.033 | 0.053 | 0.034 | 0.053 | 0.001 |
| | | Polytomous Items | a | 0.014 | 0.047 | 0.014 | 0.047 | 0.001 |
| | | | b | 0.039 | 0.082 | 0.039 | 0.082 | 0.001 |
| | | | b-d | 0.039 | 0.169 | 0.039 | 0.169 | 0.001 |
| | | | d | -- | 0.147 | -- | 0.147 | 0.001 |
| | | All Items | theta | 0.039 | 0.184 | 0.039 | 0.185 | 0.003 |

Table 8: Average Bias, RMSE & RMSD with 50% GPCM Items and 50% 2PL Model

| Sample | Items | Parameters | | ICL | | PARSCALE | | |
|---|---|---|---|---|---|---|---|---|
| | | | | *Bias* | *RMSE* | *Bias* | *RMSE* | *RMSD* |
| 500 | 40 | Dichotomous Items | a | -0.012 | 0.111 | -0.008 | 0.113 | 0.006 |
| | | | b | 0.017 | 0.070 | 0.017 | 0.070 | 0.001 |
| | | Polytomous Items | a | -0.003 | 0.071 | -0.004 | 0.071 | 0.002 |
| | | | b | 0.009 | 0.110 | 0.009 | 0.112 | 0.003 |
| | | | b-d | 0.009 | 0.279 | 0.009 | 0.281 | 0.004 |
| | | | d | -- | 0.256 | -- | 0.257 | 0.002 |
| | | All Items | theta | -0.002 | 0.237 | -0.002 | 0.237 | 0.002 |
| | 80 | Dichotomous Items | a | -0.004 | 0.099 | -0.002 | 0.101 | 0.005 |
| | | | b | -0.007 | 0.072 | -0.007 | 0.072 | 0.001 |
| | | Polytomous Items | a | 0.000 | 0.067 | -0.001 | 0.068 | 0.003 |
| | | | b | 0.007 | 0.116 | 0.006 | 0.117 | 0.003 |
| | | | b-d | 0.007 | 0.285 | 0.006 | 0.287 | 0.004 |
| | | | d | -- | 0.260 | -- | 0.262 | 0.002 |
| | | All Items | theta | -0.002 | 0.179 | -0.002 | 0.179 | 0.002 |
| 1500 | 40 | Dichotomous Items | a | 0.036 | 0.084 | 0.037 | 0.085 | 0.002 |
| | | | b | 0.038 | 0.057 | 0.038 | 0.057 | 0.000 |
| | | Polytomous Items | a | 0.003 | 0.038 | 0.002 | 0.038 | 0.001 |
| | | | b | 0.035 | 0.070 | 0.034 | 0.070 | 0.001 |
| | | | b-d | 0.035 | 0.154 | 0.034 | 0.154 | 0.001 |
| | | | d | -- | 0.137 | -- | 0.137 | 0.001 |
| | | All Items | theta | 0.038 | 0.251 | 0.038 | 0.251 | 0.001 |
| | 80 | Dichotomous Items | a | 0.020 | 0.061 | 0.021 | 0.061 | 0.002 |
| | | | b | 0.034 | 0.051 | 0.034 | 0.051 | 0.000 |
| | | Polytomous Items | a | 0.020 | 0.047 | 0.019 | 0.047 | 0.001 |
| | | | b | 0.035 | 0.081 | 0.035 | 0.081 | 0.001 |
| | | | b-d | 0.035 | 0.166 | 0.035 | 0.166 | 0.001 |
| | | | d | -- | 0.144 | -- | 0.144 | 0.001 |
| | | All Items | theta | 0.038 | 0.181 | 0.038 | 0.181 | 0.002 |