PROC IRT: A SAS Procedure for Item Response Theory

Applied Psychological Measurement
2017, Vol. 41(4) 311–320
© The Author(s) 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621616685062
journals.sagepub.com/home/apm

Ki Matlock Cole¹ and Insu Paek²

Abstract

This article reviews the procedure for item response theory (PROC IRT) procedure in SAS/STAT 14.1 to conduct item response theory (IRT) analyses of dichotomous and polytomous datasets that are unidimensional or multidimensional. The review provides an overview of available features, including models, estimation procedures, interfacing, input, and output files. A small-scale simulation study evaluates the IRT model parameter recovery of the PROC IRT procedure. The use of the IRT procedure in Statistical Analysis Software (SAS) may be useful for researchers who frequently utilize SAS for analyses, research, and teaching.

Keywords

item response theory, multidimensional item response theory, SAS

Item response theory (IRT) is an important technique to calibrate and evaluate items and to score examinees from item responses on tests, surveys, and similar instruments. Its most notable application is within the fields of education and psychology, but it is becoming increasingly popular in health and marketing. Statistical Analysis Software (SAS) is a widely used tool for data management, research, and analysis that is able to perform an extensive amount of statistical analyses utilized in a variety of industries. With the increased popularity of IRT across industries, the introduction of the IRT procedure in SAS could meet a need of those who must otherwise use separate software for calibrating test items, linking, and scoring. Furthermore, as many higher institutions have a SAS license, the students will be able to learn IRT on familiar software. The purpose of this article is to provide a review of the procedure for item response theory (PROC IRT) procedure including a small-scale basic evaluation for the model parameter recovery of some popular IRT models.

Features

The PROC IRT procedure in SAS/STAT® was released in December 2013. Table 1 presents the models supported, calibration methods, output, and limitations within PROC IRT.

¹Oklahoma State University, Stillwater, OK, USA

²Florida State University, Tallahassee, FL, USA

Corresponding Author:

Ki Matlock Cole, Research, Evaluation, Measurement, and Statistics, College of Education, Oklahoma State University, 313 Willard Hall, Stillwater, OK 74078, USA.

Email: ki.matlock@okstate.edu

Table 1. Table of Features for PROC IRT.

Models supported	
Dichotomous data	Rasch, IPL, 2PL, ^a 3PL, 4PL
Polytomous data	GR (Samejima, 1969), ^a GPC (Muraki, 1992)
Dimensionality	UD and MD
Rotation method	Orthogonal: Biquartimax, Equamax, Parsimax, Quartimax, Varimax ^a Oblique: Biquartimin, Covarimin, Biquartimax, Equamax, Parsimax, Quartimax, Varimax, Quartimin
Calibration	
Link function	Probit, Logit ^a
Item calibration	MML
Maximization method ^b	Quasi-Newton, ^a Newton–Raphson with ridging (Bock & Lieberman, 1970), EM (Bock & Aitkin, 1981)
Quadratic approximation method	Adaptive, ^a Nonadaptive Gauss–Hermite (see Rabe-Hesketh, Skrondal, & Pickles, 2002)
Latent score calibration	ML, EAP, MAP ^a
Output	
Model fit	Log likelihood, Akaike information criterion and Bayesian information criterion, the likelihood ratio chi-square G^2 statistic, and Pearson's chi-square
Item fit ^c	Likelihood ratio, Pearson's chi-square statistics
Plots	Item and test information curves, item characteristic curves
Classical item statistics ^d	Mean or p value, point-biserial item-total correlations including (unadjusted) and excluding (adjusted) the specific item, item means within i subgroups ($i = 4$ by default provides means within quartiles)
Program limitations	
Missing data	None
Sample size ^e	None
Number of items ^e	None
Mixture models	None
Polytomous score categories	11 ^f
Item prior specification	No capability

Note. IPL = one-parameter logistic; 2PL = two-parameter logistic; 3PL = three-parameter logistic; 4PL = four-parameter logistic; GR = graded response; GPC = generalized partial credit; UD = unidimensional; MD = multidimensional; MML = marginal maximum likelihood; EM = expectation–maximization; ML = maximum likelihood; EAP = expected a posteriori; MAP = maximum a posteriori; SAS = Statistical Analysis Software.

^aDenotes default.

^bThe maximum number of function calls and iterations can be specified.

^cOnly available for unidimensional models.

^dProvided for binary data with one factor using the option “itemstat.” This feature was not available in the SAS/STAT® 13.1 release but is available in the 14.1 version (SAS Institute, Inc., 2015).

^eJ. Tao (personal communication, March 8, 2016).

^fThe SAS/STAT 14.1 User's Guide indicates the maximum number of levels is 11, but the response models and VARIABLE (VAR) statement documentation indicate at most 19 levels are allowed; the authors confirm that the maximum number of levels is 11.

Interface and Documentation

The syntax of the PROC IRT follows a familiar language to other procedures in SAS. The supporting documentation and SAS/STAT® User's Guide (SAS Institute, Inc., 2015) are easily accessible at support.sas.com, with details of syntax options, model estimations, and examples. Users are encouraged to use the SAS/STAT® 14.1 User's Guide for the IRT procedure (SAS

Institute, Inc., 2015) because many features, including classical item statistics and the generalized partial credit (GPC) model, were added in the update.

Input and Output Files

Item response data can be supplied where each row of observations corresponds to an individual's set of responses to items or where each row corresponds to a unique item response string with a second variable used to indicate the frequency of each response pattern. Output is provided using HTML documentation. Datasets of estimated item parameters and standard errors, item and model fit statistics, item information, and the estimated latent trait scores can be saved to output files using the Output Delivery System in SAS.

Availability

Software licensing is available annually for individuals, business, academic institutions, and government. PROC IRT is available within the SAS/STAT module and is, therefore, an add-on to the Base SAS package. SAS/STAT® also performs more advanced statistical procedures, in addition to IRT analyses. Users can run the procedure “PROC SETINIT NOALIAS” to confirm the installation of SAS/STAT® on their system.

Performance

A small-scale simulation study was used to provide some indication of the item parameter recovery of the PROC IRT procedure for some popular IRT models.

Parameter Specifications

Two unidimensional dichotomous response models (the two-parameter logistic [2PL] and three-parameter logistic [3PL] models), two unidimensional polytomous response logistic models (the graded response [GR] and GPC models), and a simple structure dichotomous response four-dimensional logistic model (multidimensional 2PL [M2PL] model) were independently evaluated using model-generated score response data. A 40-item test was simulated for dichotomously scored data, whereas a 20-item test was simulated for the four-category polytomous scored data. Sample sizes (N s) were 1,500 for 2PL, 3,000 for 3PL, 1,500 for GR, 1,500 for GPC, and 1,500 for M2PL with the dimension correlation of .7.

The model parameterization for the 2PL and 3PL models follows the conventional IRT parameterization $a(\theta - b)$, where a , b , and θ are the slope (discrimination), difficulty, and person ability parameters, respectively. The M2PL model has the form $\mathbf{a}'\theta + d$, that is, a compensatory multidimensional item response theory (MIRT) model form, where $\mathbf{a}'$ and θ are vectors of the slope and person ability parameters, respectively, across dimensions, and d is the item intercept parameter. The slope parameters for data simulation were from a log-normal distribution with mean equal to zero and a standard deviation (SD) of 0.2. The b parameters for the 2PL and 3PL models and the intercept parameter for the M2PL models were from the standard normal distribution. The pseudo-guessing parameter for the 3PL model was from a uniform distribution with a minimum and maximum of 0.05 and 0.30, respectively. The specifications are similar to those from other simulation studies (see, for example, Sinharay, 2015). Note that the same a and b parameters were used in the 2PL and 3PL model data generation.

The step parameters (or category or boundary response curve intersection parameters) for the GPC and GR models were generated from a uniform distribution from -2.0 to -0.1 with

0.1 intervals; the second and third step parameters were separated from the first 1 and 2 units, respectively (DeMars, 2004). In other words, the first item's step parameters were -2 , -1 , and 0 ; the second item's step parameters were -1.9 , -0.9 , 0.1 ; and the last item's step parameters were -0.1 , 0.9 , and 1.9 .

The unit normal distribution, $\theta \sim N(0, 1)$, was used to generate person ability parameters for the unidimensional dichotomous and polytomous conditions. For the multidimensional model, abilities were drawn from a multivariate, two-dimensional, normal distribution with all dimension means equal to zero, dimension variances equal to one, and a correlation of $.7$ between dimensions. Item response data were generated using WinGen Version 3.1.13.447 (Han, 2007). Each condition had 100 replications.

Software and Computer Specifications

The default estimation settings and convergence criterion of PROC IRT were used for all datasets. These included the logit link function using the marginal maximum likelihood (MML) with the quasi-Newton algorithm, a maximum of 200 iterations, and adaptive Gauss-Hermite quadrature procedure. No item priors were specified, as this option is not available in the current PROC IRT procedure (J. Tao, personal communication, March 23, 2016). The multidimensional model in SAS has the Varimax rotation as the default; thus, the PROC IRT was specified with no rotation for the simple structure M2PL model.

All simulations were run on a computer with Intel® Core™ i5-4310U CPU at 2.00 GHz, 2601 MHz, two cores, and four logical processors with 8 GB of physical RAM. The operating system was Microsoft Windows 7 Enterprise. SAS 9.4 was used with the SAS/STAT® 14.1 component.

Appendix B provides the syntax of the PROC IRT procedure for the Rasch, 2PL, 3PL, M2PL, GR, and GPC models used in the simulation. The authors of this study also provide an evaluation of the IRT procedure that was established by Gierl and Ackerman (1996) in Appendix A. The criteria were established as a way to identify useful programs that implement IRT analyses according to item and latent parameter estimations, flexibility and ease of use, output, and support.

Results

Evaluation criteria included run time (i.e., the real time, not CPU), percentage of nonconvergence datasets, and measures of item parameter root mean square error (RMSE) and bias.

The 2PL, M2PL, GR, and GPC models showed a 100% convergence rate with the settings used in this simulation in less than 1 min for all sets. The 3PL model had a 91% convergence rate, and took 66.34 s, on average for each set, to run. This poor convergence is a problematic issue for the PROC IRT procedure for the estimation of the 3PL model. It would be beneficial if the item priors could be specified or constrained, but this specification has not yet been made available.

Table 2 provides the overall summaries of RMSE and bias. Box plots of the RMSE and bias for item parameters from all models are provided in Figures 1 through 6. Item parameters for the 2PL model ($N = 1,500$) were estimated with an RMSE less than 0.090 on average across all items for all slope (a) and difficulty (b) parameters. Average biases for the slope and difficulty parameters from the 2PL model were 0.050 and 0.034, respectively.

The average RMSE of the slope parameter across all items from the 3PL model ($N = 3,000$) was 0.169, and the average bias was 0.024. The difficulty parameter was estimated with larger RMSE (0.304 on average across all items) and bias (-0.042 on average across all items), which

Table 2. Average RMSE and Bias of Item Parameters Across All Items.

Model	Parameter	RMSE	Bias
2PL	a	0.088	0.050
	b	0.089	0.034
3PL ^a	a	0.169	0.024
	b	0.304	−0.042
	c	0.094	−0.005
M2PL	a_1	0.121	0.061
	a_2	0.124	0.061
	a_3	0.112	0.042
	a_4	0.116	0.049
	d	0.071	−0.009
GR	a	0.077	0.048
	b_1	0.108	0.073
	b_2	0.073	0.030
	b_3	0.079	−0.001
GPC	a	0.083	0.044
	b_1	0.102	0.067
	b_2	0.069	0.033
	b_3	0.077	−0.001

Note. RMSE = root mean square error; 2PL = two-parameter logistic; 3PL = three-parameter logistic; M2PL = multidimensional two-parameter logistic; GR = graded response; GPC = generalized partial credit.

^aNine replications of the 3PL model did not reach convergence.

Figure 1. Boxplot of RMSEs of 40 items for each parameter from the 2PL and 3PL models.

Note. RMSEs = root mean square errors; 2PL = two-parameter logistic; 3PL = three-parameter logistic.

was in the opposite direction indicating an underestimate of the true value than from the 2PL model. Larger RMSE values of the slope and difficulty parameter estimates in the 3PL model than in the 2PL model are also clearly indicated in Figure 1. The pseudo-guessing parameter showed an RMSE of 0.094 and a bias of −0.005.

The M2PL dataset had a four-factor simple structure design. The average RMSE of the estimated item discrimination across subsets of 10 items with nonzero true discrimination was 0.121, 0.124, 0.112, and 0.116 across the four dimensions, and the bias was 0.061, 0.061, 0.042,

Figure 2. Boxplot of the bias of estimations for 40 items for each parameter from the 2PL and 3PL models.
Note. 2PL = two-parameter logistic; 3PL = three-parameter logistic.

Figure 3. Boxplot of RMSEs of estimated discrimination across 10-item subsets with nonzero discrimination and RMSE of estimated easiness across all 40 items from the M2PL model.
Note. RMSEs = root mean square errors; M2PL = multidimensional two-parameter logistic.

Figure 4. Boxplot of the bias of estimated discrimination across 10-item subsets with nonzero discrimination and bias of estimated easiness across all 40 items from the M2PL model.
Note. M2PL = multidimensional two-parameter logistic.

Figure 5. Boxplot of RMSEs of 20 items for each parameter from the GRM and GPCM models.
Note. RMSEs = root mean square errors; GRM = graded response model; GPCM = generalized partial credit model.

Figure 6. Boxplot of the bias of estimations for 20 items for each parameter from the GRM and GPCM models.
Note. GRM = graded response model; GPCM = generalized partial credit model.

and 0.049. RMSE and bias of the estimated easiness (or intercept) parameter were 0.071 and -0.009 , respectively.

Polytomous model item parameters were estimated with RMSE less than 0.11 on average across all items for slope and threshold parameters from the GR and GPC models. The slope parameters had an average bias of 0.044 and 0.048 for the GR and GPC models, respectively. Average biases of the first threshold were largest (0.067 and 0.073 for the GR and GPC, respectively), and smaller for the second and third thresholds (0.033 and 0.030 for the second and -0.001 and -0.001 for the third from the GR and GPC, respectively).

Conclusion

Because of the widespread availability and use of SAS at academic institutions, the SAS IRT procedure could draw attention from researchers and possibly practitioners as well. The current

article provides a review of the procedure and a small-scale evaluation mainly for research-oriented use of the SAS IRT procedure. PROC IRT appears to be an efficient procedure in SAS for conducting some popular IRT model applications with large sample sizes within the scope of the simulation conducted in this article. Researchers, including graduate students, may benefit from the PROC IRT procedure for investigating empirical studies of applying IRT analysis to a single dataset or for composing a simulation study to evaluate the results under different conditions, data specifications, model estimation procedures, and scoring procedures. It is also beneficial for teaching purposes at institutions that have an academic license to the SAS software.

For small sample sizes and test lengths under different conditions from the current limited evaluation conditions, further investigations should be conducted to provide more solid information for researchers. For the use of the SAS PROC IRT procedure in operational settings, the need for more systematic assessment and comparison studies of the performance of the procedure regarding IRT equating and calibration is even greater. Some of the future improvements for the SAS IRT procedure includes the capability to incorporate item parameter constraints and priors into the calibration, which is especially useful for stabilizing the estimation of more complex models such as the 3PL and the four-parameter logistic (4PL) models.

Appendix A

Evaluation of procedure for item response theory (PROC IRT).

This evaluation criterion was established by Gierl and Ackerman (1996) to identify a useful program for practitioners and researchers using IRT.

Item- and trait-level parameter estimation

The use of IRT requires procedures for estimating the parameters of the items and of the examinees' trait levels on a test.

Does the program provide item parameter estimates?	Yes
Does the program provide trait-level parameter estimates?	Yes
Can different methods be used to estimate the item and trait parameters?	Yes

Flexibility and ease of use

The computer program must be flexible, providing users with options for different testing situations, yet easy to operate, given that most users will have limited knowledge about IRT.

Are the data easily formatted and prepared for the analysis?	Yes
Is the graphical user interface well-structured? For example, Are the icons clearly marked and are the function buttons carefully described?	NA
Does the program contain an editor for the data, input, and output files?	Yes
Does the program contain online help?	Yes
Is the program fast and quickly executed?	Yes
Can the program be used with large samples and with large tests containing many items?	Yes

Comprehensive output

The output must be comprehensive—providing the user with information at both the item and test levels.

Does the output include a classical item analysis?	Available with options
Does the output include an IRT analysis in which summaries of the model fit, the standard errors, and the information functions are available at the item and test levels?	Available with options
Does the program graph the item response functions and the test response function?	Available with options
Does the program graph the item information function and the test information function?	Available with options
Is the output well-formatted and easy to read?	Yes

(continued)

Appendix A (continued)

Can the input and output files be easily exported to other applications?	Yes
<i>Manual and technical support</i>	
The operation and execution of the program must be clearly presented in the manual, and the user should receive free technical product support.	
Does the program come with a comprehensive manual in which data formatting, program formatting, user options, estimation procedures, and error codes are carefully described?	Yes
Does the program come with datasets and does the manual contain detailed sample runs in which typical examples encountered in test development are examined and explained?	No datasets Examples provided
Are program features documented and explained in the manual (e.g., linking functions, scaling options, comparisons across parameter estimation options)?	Yes

Note. IRT = item response theory.

Appendix B**Syntax**

Rasch model:

```
proc irt data=data itemfit out=fscore2PL itemstat;
  var i1-i40;
  model i1-i40/resfunc=rasch;

run;
```

Two-parameter* logistic model:

```
proc irt data=data itemfit out=fscore2PL itemstat;
  var i1-i40;
run;
```

*The two-parameter logistic model is the default model for binary items. Therefore, the specification “resfunc=twop” is unnecessary.

Three-parameter logistic model:

```
proc irt data=data itemfit out=fscore3PL;
  var i1-i40;
  model i1-i40/resfunc=threep;

run;
```

Four-dimensional, simple structure, two-parameter logistic model:

```
proc irt data=data nfactor=4 out=fscoreM2PL rotate=none;
  var i1-i40;
  factor Factor1->i1-i10,
         Factor2->i11-i20,
         Factor3->i21-i30,
         Factor4->i31-i40;

run;
```

Graded response model:

```
proc irt data=data itemfit out=fscoreGR;
  var i1-i20;
  model i1-i20/resfunc=gr;
```

```
run;
Generalized partial credit model:

proc irt data=data itemfit out=fscoreGPC;
  var i1-i20;
  model i1-i20/resfunc=GPC;
run;
```

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443-459.
- Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, 35, 179-197.
- DeMars, C. E. (2004, April). *A comparison of the recovery of parameters using the nominal response and generalized partial credit models*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.
- Gierl, M. J., & Ackerman, T. (1996). XCALIBRE™ marginal maximum-likelihood estimation program, Windows™ version 1.10. *Applied Psychological Measurement*, 20, 303-307.
- Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement*, 31, 457-459.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176.
- Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *Stata Journal*, 2, 1-21.
- Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometrika Monograph Supplement No. 17). Retrieved from <https://www.psychometricsociety.org/sites/default/files/pdf/MN17.pdf>
- SAS Institute, Inc. (2015). *SAS/STAT® 14.1 User's Guide*. Cary, NC: Author.
- Sinharay, S. (2015). The asymptotic distribution of ability estimates: Beyond dichotomous items and unidimensional IRT models. *Journal of Educational and Behavioral Statistics*, 40, 511-528.