Paper 204-30

# IRT-FIT: SAS® Macros for Fitting Item Response Theory (IRT) Models

Sung-Hyuck Lee, University of Oklahoma, Norman, OK

Robert Terry, University of Oklahoma, Norman, OK

## ABSTRACT

Psychometrics has recently seen the development of complex measurement models to better represent test and item data. Item Response Theory (IRT), in particular, comprises a set of non-linear latent variable models that appear to have several conceptual and empirical properties that make them more valuable in practice than classical test theory methods. However, IRT-based models typically require the availability of costly and computationally-intensive software for estimating parameters and assessing model fit. In this paper, we present a set of SAS Macros called IRT-FIT, which use SAS /IML® and SAS/GRAPH® to estimate, fit, and graph two- and three-parameter IRT models to binary test data. The macros currently developed use Bock and Aitkin's (1981) Marginal Maximum Likelihood (MML) estimation algorithm for fitting models and estimating parameters as the basis for the computations. Additionally, we have extended the MML routines by implementing Bayesian Estimation concepts as suggested in Mislevy (1986). All computational routines are written in SAS/IML, and output data sets are produced containing the parameter estimates along with their associated standard errors and overall model fit statistics. Optionally, SAS/GRAPH plots are available of the estimated Item Characteristic Curves (ICC's), the item and test information curves, as well as the standard error curve for estimated latent trait scores. Finally, if the test data come from a rating experiment and a cut-point along the latent variable can be determined, ROC curves using IRT-based estimates of Signal-Detection-Theory concepts are plotted to visually represent rater performance.

## INTRODUCTION

Psychometrics has recently seen the development of complex measurement models to better represent test and item data. For example, the recent development and subsequent focus on the concept of construct representation (Embretson and Reise, 2000) suggests that models of items, not tests, are more appropriate for many applications. Moreover, models that can serve as the basis for process analysis show potential for understanding the nature of the person-item interaction within the context of a particular stimulus presentation. As theories of measurement develop, such models are the future of test construction and development.

Item Response Theory (IRT) is a mathematical and statistical model of item responses in a population of individuals (Birnbaum, 1968). The basic model can take on several forms, depending upon the chosen parameterization and the functional form of the Item Characteristic Curve (ICC). In typical applications, logistic or normal ogive ICC's are chosen to represent the mathematical form of the ICC. Finally, another aspect of the basic IRT model is the limitation of unidimensionality; the basic IRT model assumes a single latent trait is the only systematic source responsible for variations in observed item response patterns.

Most IRT models fall within three basic types of model parameterization. The simplest IRT model contains only a single parameter and is usually identified as the Rasch model (1960). For the Rasch model, items are believed to vary only in their relative difficulty. Once the relative difficulty of each item is estimated, person-estimates of test scores can be obtained by subsequently treating the item difficulty as fixed.

The Rasch IRT model is a popular method thanks to several desirable statistical properties. The most important of these properties is that sufficient statistics exist to estimate the person-score independent of the item parameters, hence computation of parameters is simplified as compared to more complex parameterizations. Estimation of the Rasch IRT model parameters can be accomplished with many general-purpose programs to model categorical data, such as SAS PROC GENMOD or SAS PROC LOGISTIC.

The Rasch IRT model, however, will not adequately represent empirical data in the typical situation. In these instances, either two- or three-parameter models are necessary. The two-parameter model (Birnbaum, 1968) adds a varying slope to the Rasch IRT model, allowing item responses and person abilities to interact. In the two-parameter model, each ICC is represented by a different slope; it follows that each item varies in average reliability. Using the two-parameter model as a baseline, the three-parameter model further adds a “pseudo-guessing” parameter to each ICC, with the intent of accounting for observed performance of those persons with very low levels of the latent trait.

It has been mathematically demonstrated that both the two- and three-parameter models do not have sufficient statistics for estimating item parameters that are independent of person-scores (Birnbaum, 1968). In essence, no statistic exists that can be used to estimate the item parameters in an optimum way that does not depend upon the conditioning variable itself - in this case the unknown latent trait.

For two- and three-parameter IRT models, no easy solution for estimating parameters will exist. As a non-linear latent variable model, special algorithms must be developed to accurately estimate the model parameters. However, most available software is typically costly, difficult to use, and inconvenient because of their stand-alone nature. In this paper, we present a set of SAS macros called IRT-FIT, written in PROC IML, which implement state-of-the-art parameter estimation routines for both the 2- and 3-parameter models. These macros provide a cheap, yet computationally accurate means of estimating item and ability parameters within the context of an existing statistical software package.

Finally, although IRT models exist for non-binary data, the current macros are restricted to analyzing data with binary responses only. Additional work is under way to develop SAS programs for polytomous unidimensional IRT models, as well as multidimensional IRT models.

Before getting to the specific of the macros themselves, we turn to a brief introduction of the mathematics of IRT models.

## MATHEMATICS OF IRT MODELS

We use Birnbaum’s (1968) notation and the logistic function to mathematically represent the model. For any item $X_{ij}$ denoting the response for the jth person on the ith item, we can write:

$$(1)\quad P(X_{ij} \mid \theta_j) = p_{ij}^{X_{ij}} q_{ij}^{1-X_{ij}}$$

where $x_{ij} = 0,1$ depending upon the correctness of the response, $\theta$ is the value of the latent trait for the jth person, $p_{ij}$ is the(unknown) probability of getting the ith item correct for person j and with $q_{ij} = 1 - p_{ij}$.

Using the logistic function, we then write the three-parameter logistic (IRT-3PL) as:

$$(2)\quad p_{ij} = P(X_{ij} = 1 \mid \theta_j) = c_i + (1 - c_i)\frac{\exp[D\alpha_i(\theta_j - \beta_i)]}{1 + \exp[D\alpha_i(\theta_j - \beta_i)]}$$

In equation (2), the α, β, and c parameters represent slope, difficulty, and pseudo-guessing parameters, respectively. When c=0, the model reduces to the two-parameter IRT model (IRT-2PL). When $\alpha_j = \alpha$ for all j , the model becomes the Rasch-IRT model. Finally, if D=1.702, the logistic ICC is scaled in the same metric as the normal ogive ICC (Birnbaum, 1968).

Given (1) and (2), the likelihood of a particular response pattern for the ith person is given by:

$$(3)\quad P(X_{i1},\ldots,X_{ij} \mid \theta_i, \underline{\alpha}, \underline{\beta}, \underline{c}) = \prod_j P_{ij}^{X_{ij}} Q_{ij}^{1-X_{ij}} \quad \text{or more compactly as}$$

$$P(\underline{X}_i \mid \theta_i, \underline{\alpha}, \underline{\beta}, \underline{c}).$$

Finally, given (3), the Data Likelihood (over all θ) is then given as:

$$(4)\quad L(\underline{X}) = \prod_i P(\underline{X} \mid \theta_i, \underline{\alpha}, \underline{\beta}, \underline{c})$$

Computations are radically simplified if one applies a log transformation to (4), achieving log-likelihoods equations for the data.

Equation (4) clarifies the likelihoods' dependence upon the unknown latent trait θ. Bock and Aitkin (1981) solved the problem by expressing the response pattern probabilities as expectations regarding a population distribution for θ. Thus, the observed data are regarded as a random sample from a population of persons with latent trait distribution g(θ). Mathematically:

$$(5)\quad L(\underline{X}_j) = \int_\theta \prod_i P_{ij}^{X_{ij}} Q_{ij}^{1-X_{ij}} g(\theta)\partial\theta$$

## IRT-FIT AND MARGINAL MAXIMUM LIKELIHOOD (MML) ESTIMATION

The basis for estimating the item parameters in equation (5) is the Bock and Aitkin (1981) Marginal Maximum Likelihood (MML) solution. Because the person-ability parameters can be considered nuisance parameters for estimating item parameters, the Bock and Aitkin solution effectively marginalizes the likelihood equations, integrating away the person-ability distribution as in equation (5). Using Gauss-Hermite quadrature to perform the numerical integration, the resulting marginalized equations are then subjected to the EM algorithm for generating the expected values to be maximized. Because not every response is - or even can be - observed, Bock and Aitken (1981) suggested the EM-algorithm as an appropriate mechanism for handling the missing item response patterns that are sure to exist. Once obtained via the EM algorithm, the item parameters are treated as fixed and the person-scores are estimated using normal complete-data ML methods.

## IRT-FIT AND BAYESIAN (BAYES) ESTIMATION

In some situations, it becomes difficult to get convergence in estimation for the item and person-score parameters using MML solutions. To ameliorate this potential problem, we have also implemented a Bayesian-augmentation routine for use with very small or ill-behaved data sets. We often know enough about the distribution of certain model parameters to use this prior information to augment the sparse empirical data. Provided the parameters of this Bayesian prior are chosen wisely, the additional information can often fix convergence issues should they arise.

We consider here the estimation of person-scores θ. Item parameter estimation follows similar lines. Briefly, we can write the Bayes equations for IRT models thusly:

$$(6)\quad P(\theta \mid \underline{X}) \propto L(\underline{X} \mid \theta) P(\theta)$$

In equation (6), P(θ) is the (assumed) prior distribution of the latent trait, L(X| θ) is the data likelihood, and the revised likelihood P(θ | X) is referred to as the posterior density of θ. The mode of this distribution is the most probable value for θ, given the data likelihood and the prior distribution. Estimation of the mode is this manner is called Maximum A Posteriori (MAP) estimation (Mislevy, 1986). One could also compute the mean of the posterior distribution - Expected A Posteriori (EAP) estimation - as an alternative. Both MAP and EAP estimates of the person-scores are implemented in the IRT-FIT macro as an option.

We have also implemented Bayesian augmentation methods for estimating item parameters. In both the 2- and 3-parameter models, prior distributions for the difficulty, slope, and pseudo-guessing parameters are available. Our experience suggests it is usually necessary to specify priors for the 3PL model in order to obtain convergence in the item parameter estimation phase.

## IMPLEMENTING THE IRT-FIT MACROS.

We have developed two SAS macros for this paper. IRT_FIT_2P is the first macro and is used for estimating parameters in the unidimensional two-parameter IRT model. IRT_FIT_3P is the second macro and is used only for estimating parameters in the unidimensional IRT model that contains pseudo-guessing parameters. Code development and implementation was based on the work of Baker (1992).

The macros are invoked in the usual way - %IRT_FIT_2P(options) and %IRT_FIT_3P(options).

The following options are available for both the 2P and 3P macros:

ITEM_TOT = i – user-supplied number (REQUIRED)
SAMPLE_TOT = j – user-supplied number (REQUIRED)

For these two options, the response option i indicates the number of items in the problem and the response option j indicates the number of persons (examinees, stimuli) is the sample data set.

The program has two options for how the estimates are computed. In the first option, the program simultaneously estimates all item parameters at the same time. This option is very fast but it requires the creation of a super-matrix for computing purposes, in which most of the matrix elements are zero. For small problems, this solution is both elegant and fast.

The second option computes the parameter estimates for each item one at a time; the assumption of local independence in all IRT models makes this a feasible solution. Matrix storage requirements are reduced at the expense of computational speed.

The program automatically chooses the method of computation based on the size of examinee by item matrix. The program default preference is for speed whenever feasible. .

EST_METHOD = MMLE (default)
BAYES

The default is to use the MMLE solution. IF BAYES is chosen, the user can choose to specify the location and standard deviation of the prior distributions or use the defaults provided by the program. Using the BAYES option is often useful in handling items for which the estimated parameters are problematic (e.g. slope less than 0).

METRIC = LOGISTIC
NORMAL (default)

The NORMAL solution is achieved by applying the scaling factor D=1.702 to the LOGISTIC solution, not by solving the normal ogive representation of the ICC. Prior research suggests that using the scaling factor reproduces the normal ogive ICC's within .02 of their value along the entire curve.

SCORE = MAP (Posterior modal estimate)
EAP (Posterior mean estimate)
MLE (Default)

The default is to use data-based scoring (MLE) only. If an examinee has missed all items, an approximate solution is obtained by changing the examinees response to the easiest item from 0 (wrong) to ½. (partially correct). Likewise, if an examinee has responded correctly to every item, an approximate solution is obtained by changing the response to the most difficult item from 1 (correct) to ½. (partially correct).

If MAP or EAP is chosen, a NORMAL prior for THETA is assumed. Finally, if item parameters are estimated to be negative or too close to zero, these items are not used during the THETA calibration step.

M_LOGSLOPE = 0 (default)
m - user-supplied number

If METHOD=BAYES is chosen, the default is to use a log-normal distribution on item slopes, with a default mean of 0 (which translates to a mean of 1.13 in the normal metric).

The user, however, can specify any prior estimate of location desired. This option is especially useful for simulation work.

STD_LOGSLOPE = .5 (default)
s – user-supplied number

If METHOD=BAYES is chosen, the default is to use a log-normal distribution on item slopes, with a default standard deviation of .5 (which translates to STD = .6 in the normal metric). The user, however, can specify any prior estimate of scale desired. This option is also especially useful for simulation work.

THETA = (m,s)
m – location parameter (mean) - default (m=0)
s – scale parameter (standard deviation) – default (s=1)

The origin of the THETA scale is indeterminate in the IRT model; this is a common characteristic of any latent variable model. There are two simple solutions to this problem. One can fix the item slopes to fix the scale of measurement, or one can fix the location and scale of THETA. This option gives the user the capability of rescaling the THETA scale to constants (m,s) rather than using the default mean=0, SD = 1. This can be very useful in psychological or educational settings where metrics are well-established (e.g. IQ scores, clinical T-scores).

NQP = 10 (default)
q – user-supplied number

NQP is the number of quadrature points used to numerically integrate the ability distribution in the marginalization step of the Bock and Aitken (1981) algorithm. Using 10 quadrature points is typically sufficient for most problems. Using more than 10 quadrature points increases the computational cost in speed with little gain in accuracy.

EM_CRIT = .001 (default)  
t – User-supplied number

This is the convergence criterion necessary for the EM cycles to terminate. The value t implies that no item parameter has changed by more than t from the previous iteration.

EM_CYCLES = 50 (default)  
t – User-supplied number

EM algorithms are notoriously slow to converge. The default of 50 usually is sufficient for well-behaved data. Increasing the number of EM cycles is often successful is some situations.

NR_ITEM_CRIT = .001 (default)  
t – User-supplied number

Following the EM step, Newton-Raphson cycles are implemented in the final step. Again, the value t implies that no item parameter has changed by more than t from the previous iteration.

NR_ITEM_CYCLES = 2 (default)  
t – User-supplied number

In most cases, the Newton-Raphson cycle converges in a single cycle.

NR_ABILITY_CRIT = .001 (default)  
t – User-supplied number

Once item parameter estimates are obtained, a Newton-Raphson iterative algorithm is applied to obtain THETA estimates treating item parameter estimates as fixed. The value t implies that no THETA parameter has changed by more than t from the previous iteration.

NR_ABILITY_CYCLES = 30 (default)  
t – User-supplied number

Once item parameter estimates are obtained, a Newton-Raphson iterative algorithm is applied to obtain THETA estimates treating item parameter estimates as fixed. Thirty cycles is usually sufficient for convergence to be met in most situations.

PRINT_OPT = 0 (None)  
1 (Item descriptive statistics)  
2 (Item parameters + item descriptive statistics)  
3 (Person-scores)  
4 (ALL of the above – default)

The user can control the level of printed output desired. The program automatically creates three SAS data sets (ITEM_DES, ITEM_PAR,THETA_PAR) for the user after the macro has finished executing. These data sets may be printed as well using the PRINT_OPT option for control of printed output.

PRINT_ITER = NO (default)
YES

The user can control the printing of the iteration history and the parameter changes as they occur. This option is rarely needed or used. The iteration history is primarily useful for diagnosing convergence problems and for comparing the speed of convergence to other algorithms.

PLOT_OPT = 0 None (default)
1 ICC's only
2 Item Information Curves Only
3 ICC's plus Item Information Curves
4 Test Information Curve Only
5 Test Characteristic Curve Only
6 ALL of the above

Several plots are available using the GPLOT procedure in SAS. The ICC plots show the curve relating THETA to the estimated probability of success, P(θ). There is one plot for each item.

Item Information Curves (IIC) show the amount of information each item provides for estimating THETA along the THETA scale. In general, IIC's are steeper when the item slope is high, and are centered at the estimated difficulty level, accounting for guessing. Items with mostly flat IIC's are considered unreliable.

The Test Information Curve (TIC) is the aggregate of all the IIC's. This curve can be used to show the sensitivity of the entire test for local regions of THETA. If the TIC is low and fairly flat for some local region of THETA, it suggests that THETA (person-scores) in this local region are unreliable, even though the test as a whole may be quite reliable.

The Test Characteristic Curve (TCC) is a plot relating THETA to the True-Score Metric (i.e. 0 to N – the number of items in the test). The TCC can be used to make True-score metric approximations (e.g. in integer values) based on IRT methodology. The TCC is also useful when attempting to construct two equivalent forms of a test, a process known as equating.

ROC = 0 None (default)
1 Plot of typical (average) item ROC curve
2 Plot of each items ROC curve against the typical curve.
3 Plot of Test ROC Curve
4 Plots 1 and 3 only
5 ALL plots produced

CUTPOINT = c – user-specified number (Required if ROC not equal 0).

For mastery tests, it is often the case that a standard exists. This standard may be known independently of the test characteristics, perhaps based on the application of a standard-setting procedure using subject-matter experts. It may also be inferred from knowledge of the population, such that one wishes to select only the top 15%, say, of the examinees for further consideration in a selection context.

If such a cut-point can be determined, signal-detection theory can be used to evaluate the sensitivity of the test (items) relative to the false-alarm rate in making binary decisions of mastery versus non-mastery classifications. Options 1 and 2 produce ROC Curves at the item level. Option 3 produces a single ROC curve for the entire test, based on the Test Characteristic Curve (TCC).

.

## OUTPUT OF THE PROGRAM

Output consists of three data sets. Data set one, ITEM_DES, contains the usual item statistics found in classical test theory. Data set two, ITEM_PAR, contains item parameter estimates, standard errors of estimates, and goodness-of-fit information for each item. Data set three, THETA_PAR, consists of person-scores along with their associated standard errors. These are always produced and may be used as the user sees fit. Finally, the various plots listed above are produced if requested. A complete description of the program output is detailed in the Technical Manual *IRT-FIT: Fitting IRT models in SAS®* (Terry, Lee, and Milburn, 2005).

## SPECIAL CONSIDERATIONS

**Guessing Models**. It is not uncommon for 3PL IRT models to fail to achieve convergence. We have implemented both MML and Bayesian estimation, although it is rare that the MML solution converges. The program automatically chooses the priors based on theoretical and empirical considerations, which currently is not under the control of the user.

**Missing Data**. It is likely that most applications of testing will result in missing item responses. Assuming that the missing item responses are Missing At Random (MAR), the program easily accommodates missing data via the EM algorithm and the use of a selection matrix. Nothing need be specified by the user; the program assumes the standard SAS convention for missing data (e.g. a period). Missing data formats are especially convenient for IRT analyses such as equating and differential item functioning (Embretson and Reise, 2000)

**Calibration to existing software**. We have used several data sets and compared the output from the IRT-FIT macros to that produced by existing software (e.g. BILOG 3; Zimowski, Muraki, Mislevy and Bock, 2000). The results suggest little discrepancy between parameter estimates, with a mean discrepancy (d) of approximately d=.02.

**Rater calibration**. Rater calibration models can be viewed as a special case of the IRT model, with raters serving as items and the rated stimulus set playing the role of examinees (Terry, 2000). The program produces special statistics appropriate for the analysis of such data., along with any desired plots of rater sensitivity.

## CONCLUSION

The IRT-FIT macros developed in this paper can be used for multiple purposes. Since they are written in SAS PROC IML, they provide a very low cost and convenient alternative to existing software. For example, the IRT-FIT macros are useful for teaching IRT-based measurement to students, who can see both the exact nature of the code and obtain results for very little cost. Additionally, the IRT-FIT macros are valuable for creating simulation and Monte Carlo studies, using SAS Data Management capabilities for further collection and analysis of the simulated results. These IRT-FIT macros provide a needed step next step by bringing fairly low-cost programs for fitting IRT models to the everyday psychometric user.

Of course, ongoing development is part of the process. We are currently in the process of developing new macros for non-binary data. As we implement more options, we update the macros regularly along with the technical manual. The complete set of SAS IRT-FIT macros and accompanying technical manual can be obtained from the second author via email request.

## REFERENCES

Baker, F. (1992), *Item Response Theory: Parameter Estimation Techniques*, New York: Marcel Dekker, Inc..

Birnbaum , A. (1968), “Some Latent Trait Models and Their Use in Inferring an Examinee’s Ability”, In F.M Lord and M.R. Novick (eds.), *Statistical Theories of Mental Test Scores*, (pp. 397-472), Reading, MA: Addison-Wesley.

Bock, R.D. and Aitken, M. (1981), “Marginal Maximum Likelihood Estimation of Item Parameters. An Application of an EM Algorithm”, *Psychometrika*, 46, 443-459.

Embretson S. E. and Reise, S. P. (2000), *Item Response Theory for Psychologists*, Mahwah, New Jersey: Lawrence Erlbaum Associates.

Rasch, G. (1960), *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: University of Chicago Press.

Terry, R. (2000), “Recent Advances in Measurement Theory and the Use of Sociometric Techniques”, in A.H.N. Cillessen and W.M. Bukowski (eds.), *New Directions for Child and Adolescent Development: Recent Advances in the Measurement of Acceptance and Rejection in the Peer System*, 88, (pp. 27-53),San Francisco: Jossey-Bass..

Terry, R,, Lee, S.H., and Milburn, N (2005). *IRT-FIT: Fitting IRT models in SAS®* , Technical Manual for Users. .

Zimowski, M.F., Muraki, E., Mislevy, R.J., and Bock, R.D. (2000). *BILOG-3: IRT Analysis and Test Maintenance for Binary Items,* Chicago: Scientific Software, Inc.
.

## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Robert Terry
Company: University of Oklahoma
Address: 808 DAHT
City state ZIP: Norman OK 73019
Work Phone: (405)-325-4593
Fax: (405)-325-4737
Email: r terry@ou.edu