IRTPRO 3 FEATURES

IRTPRO is an entirely new application for item calibration and test scoring using IRT.

Item response theory (IRT) models for which item calibration and scoring are implemented in IRTPRO are based on unidimensional and multidimensional [confirmatory factor analysis (CFA) or exploratory factor analysis (EFA)] versions of the following widely used response functions:

- Two-parameter logistic (2PL) (Birnbaum, 1968) [with which equality constraints includes the one-parameter logistic (1PL) (Thissen, 1982)]
- Three-parameter logistic (3PL) (Birnbaum, 1968)
- Graded (Samejima, 1969; 1997)
- Generalized Partial Credit (Muraki, 1992, 1997)
- Nominal (Bock, 1972, 1997; Thissen, Cai, & Bock, 2010)

These item response models may be mixed in any combination within a test or scale, and any (optional) user-specified equality constraints among parameters, or fixed values for parameters, may be specified.

IRTPRO implements the method of Maximum Likelihood (ML) for item parameter estimation (item calibration), or it computes Maximum *a posteriori* (MAP) estimates if (optional) prior distributions are specified for the item parameters. That being said, alternative computational methods may be used, each of which provides best performance for some combinations of dimensionality and model structure:

- Bock-Aitkin (BAEM) (Bock & Aitkin, 1981)
- Bifactor EM (Gibbons & Hedeker, 1992; Gibbons *et al*., 2007; Cai, Yang & Hansen (2011)
- Generalized Dimension Reduction EM (Cai, 2010-a)
- Adaptive Quadrature (ADQEM) (Schilling & Bock, 2005)
- Metropolis-Hastings Robbins-Monro (MHRM) (Cai, 2010-b, 2010-c)

- Markov Chain Monte Carlo (MCMC) Patz-Junker's (1999-a, 1999-b)

The computation of IRT scale scores in IRTPRO may be done using any of the following methods:

- Maximum a posteriori (MAP) for response patterns
- Expected a posteriori (EAP) for response patterns (Bock & Mislevy, 1982)
- Expected a posteriori (EAP) for summed scores (Thissen & Orlando, 2001; Thissen, Nelson, Rosa, & McLeod, 2001)

Data structures in IRTPRO may categorize the item respondents into groups, and the population latent variable means and variance-covariance matrices may be estimated for multiple groups (Mislevy, 1984, 1985). [Most often, if there is only one group, the population latent variable mean(s) and variance(s) are fixed (usually at 0 and 1) to specify the scale; for multiple groups, one group is usually denoted the "reference group" with standardized latent values.]

To detect differential item functioning (DIF), IRTPRO uses Wald tests, modeled after a proposal by Lord (1977), but with accurate item parameter error variance-covariance matrices computed using the Supplemented EM (SEM) algorithm (Cai, 2008).

Depending on the number of items, response categories, and respondents, IRTPRO reports several varieties of goodness of fit and diagnostic statistics after item calibration. The values of –2 log likelihood, Akaike Information Criterion (AIC) (Akaike, 1974) and the Bayesian Information Criterion (BIC) (Schwarz, 1978) are always reported. If the sample size sufficiently exceeds the number of cells in the complete cross-classification of the respondents based on item response patterns, the overall likelihood ratio test against the general multinomial alternative is reported. For some models, the M_2 statistic (Maydeu-Olivares & Joe, 2005, 2006; Cai, Maydeu-Olivares, Coffman, & Thissen, 2006) is also computed. Diagnostic statistics include generalizations for polytomous responses of the local dependence (LD) statistic described by Chen & Thissen (1997) and the $SS\text{-}X^2$ item-fit statistic suggested by Orlando & Thissen (2000, 2003).

ORGANIZATION OF THE USERS GUIDE

The user's guide has been written to introduce item response theory (IRT) models to researchers new to this field. It also serves as a guide to researchers who are already familiar with the existing IRT programs distributed by Scientific Software International and are upgrading to a program that has an easy to use graphical users interface (GUI) and can handle multidimensional models. In this guide the focus is on the "how to" part of IRT.

Chapter 2 provides a short description of the GUI, since the examples in the remaining chapters further illustrate the features of the user's interface.

IRTPRO uses its own data format, displayed in spreadsheet form. Data may be imported from

a long list of statistical software packages and spreadsheet programs. Chapter 3 deals with data import and manipulation and Chapter 4 deals with the calculation of traditional summed-score statistics.

Chapters 5 to 7 deals with the estimation (calibration) of IRT models. Chapter 5 is concerned with the fitting of unidimensional models and Chapter 6 deals with multiple groups and differential item functioning (DIF). In Chapter 7, we describe how IRTPRO handles exploratory and confirmatory factor analysis models. This chapter also contains examples illustrating the fit of bifactor and one and two-tier testlet response theory models.

Unlike classical test theory, IRT does not in general base the estimate of the respondent's ability (or other attribute) on the number-correct (NC) or summed score. To distinguish IRT scores from their classical counterparts, we refer to them as "scale" scores. The computation of IRT scale scores in IRTPRO may be done using one of the three methods discussed in Chapter 8.

One way of evaluating the impact of the violation of model assumptions, as well as studying factors such as the impact of choice of models, examinee sample sizes, the shape of ability distributions, and test length, and many other factors, is via simulation studies, also referred to as Monte Carlo studies. The purpose of the IRTPRO simulation module, the topic of Chapter 9, is to simulate examinee item response data given true model parameter values (both items and subjects). The simulation module creates data files in a form that can be directly run in the IRTPRO software as well as being saved for future use.

Graphics are often a useful data-exploring technique through which the researcher may familiarize her- or him with the data. IRTPRO offers both model-based and data-based graphs. The Model-based graphs discussed in Chapter 10 cover item- and test- characteristic curves; information and total information curves and are available for unidimensional IRT models only. This chapter also contains four types of graphical displays that serve as diagnostic tools for the MCMC method of estimation.

The MCMC graphical procedure discussed in Chapter 11 produces four types of plots that can aid further in convergence checks as well as identifying problems associated with a the specification of an IRT model.

In the case of the data-based graphs presented in Chapter 12, IRTPRO distinguishes between univariate and bivariate graphs. Univariate graphs are particularly useful to obtain an overview of the characteristics of a variable. However, they do not necessarily offer the tools needed to explore the relationship between a pair of variables.

For most unidimensional and bifactor IRT models parameter estimation can be done effectively selecting the Bock-Aitkin EM algorithm (the default estimation method). In the case of multidimensional models, the method of estimation depends to a large extent on the

number of dimensions of the model to be fitted. A general rule is that two-dimensional models can be handled effectively using Bock-Aitkin or adaptive quadrature. For three- to four-dimensional models, the estimation methods of choice are adaptive quadrature and MH-RM. Higher dimensional models are handled most effectively using MH-RM and MCMC. Chapter 13 provides a short description of the options available for each of these estimation methods.

Each analysis created by the GUI produces a syntax file, essentially being a record of a user's selections from the sequence of dialogs. If a syntax file is opened, IRTPRO automatically fills the relevant GUI dialogs that can be viewed and modified. These aspects are dealt with in Chapter 14.

MONTE CARLO-MARKOV CHAIN (MCMC) ESTIMATION

An interest in fully Bayesian MCMC methods for estimating complex psychometric models has significantly increased in recent years. These sampling based methods are more flexible and can provide a more complete picture of the posterior distributions of all parameters in the model than maximum marginal likelihood estimation methods. They can be applied in situations (*e.g.*, small sample size) where the likelihood methods tend to break down. The samples produced by the MCMC procedure can also be used creatively for conducting model fit diagnosis (*e.g.*, posterior predictive checking), model selection, and model-based prediction.

The MCMC algorithm implemented in IRTPRO is based on the Patz-Junker's (1999-a, 1999-b) blocked Metropolis algorithm. The methodology developed in IRTPRO to impose parameter constraints and to implement multiple-group features enables the user to fit specialized IRT models using MCMC.

MCMC GRAPHICS

The MCMC graphical procedure produces the following four types of plots that can aid further in convergence checks:

- Autocorrelations
- Trace Plots
- Running Means
- Posterior densities

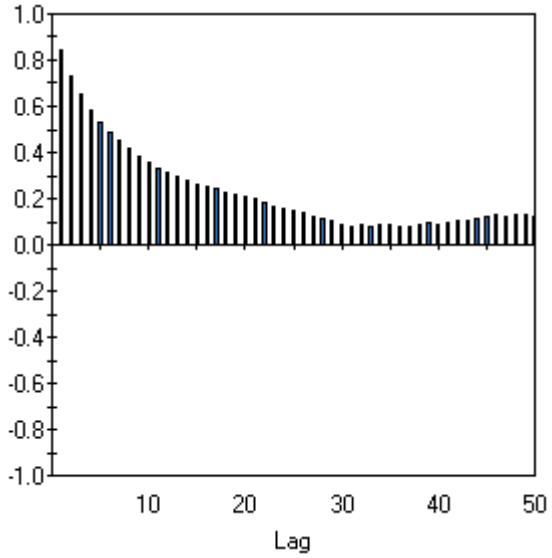
Autocorrelations

The sample autocorrelation coefficient measures the similarity between MCMC draws as a function of the time separation between them. It should be expected that the *h*-th lag autocorrelation is smaller with increase in *h* (for example, the 2nd and 30th draws should

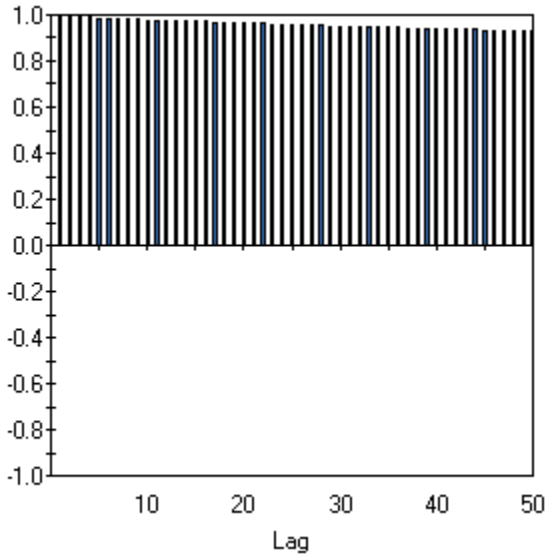
be less correlated than the 2nd and 4th draws). If the autocorrelation is still relatively high for higher values of *k*, it indicates a high degree of correlation between draws and therefore slow mixing.

The display below is an example of an autocorrelation plot that indicates good mixing (left pane) and one that indicates poor mixing (right pane).

Good Mixing



Bad Mixing



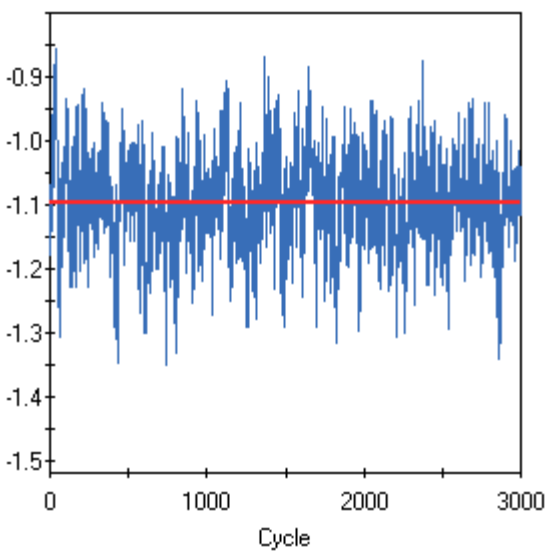
Trace Plots

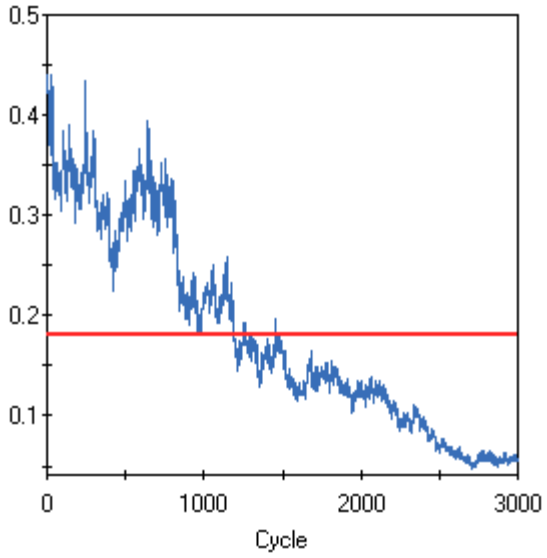
A trace plot shows the values that the relevant parameter took during the runtime of the chain. The mean (parameter estimate) of all the MCMC draws is represented by a horizontal red line.

The display below is an example of a trace plot that indicates good mixing (left pane) and one that indicates bad mixing (right pane).

Good Mixing

Bad mixing

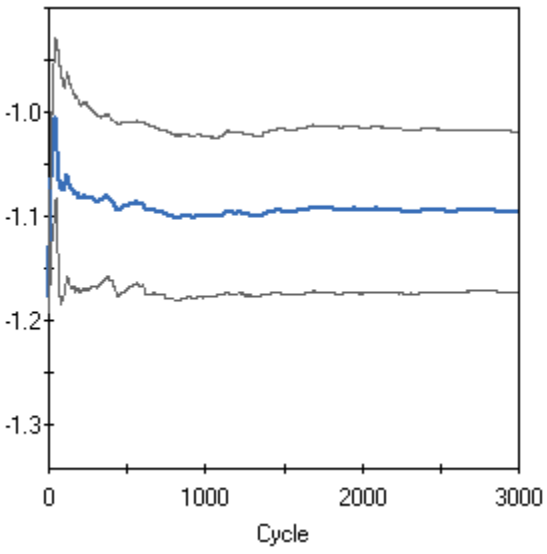




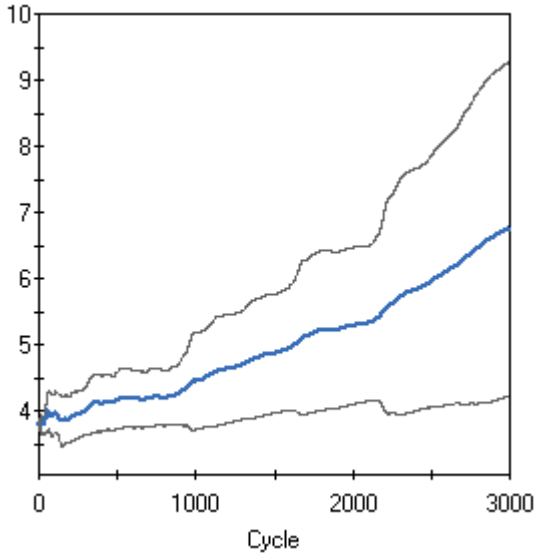
Running Means

Running means plots are used to check how well the MCMC chains are mixing. The **Running Means** plot updates the means and standard deviations for each new cycle. In other words, once (for example) the mean is known for the first few observations, it is updated using a special algorithm by just adding the value of the next observation, and so on. The plots below show the means (blue line) + or – one standard deviation.

Good Mixing



Poor Mixing

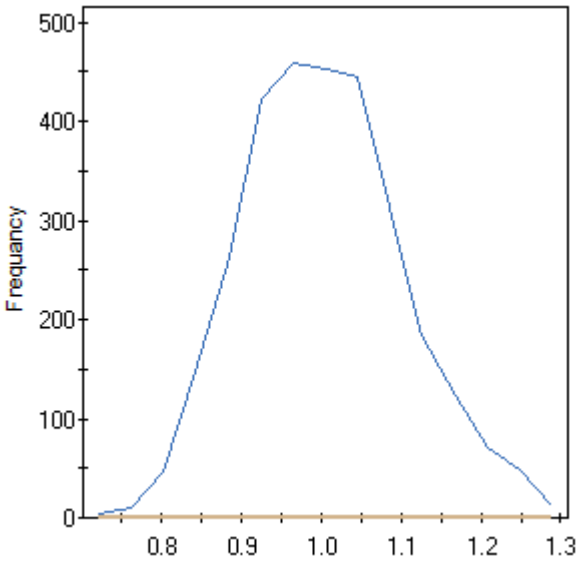


The display above is an example of a running means plot that indicates good mixing (left pane) and one that indicates bad mixing (right pane).

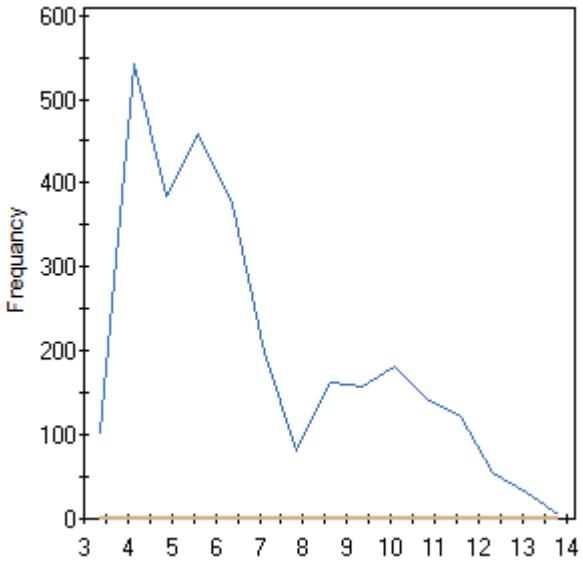
Posterior Densities

A posterior density plot is the histogram of the values in the trace-plot, *i.e.* the distribution of the values of the relevant parameter in the chain. These plots are usually called marginal density plots. The display below is an example of a trace plot that indicates good mixing (left pane) and one that indicates bad mixing (right pane).

Good Mixing



Bad Mixing



IRT SIMULATION

IRT models often rely on, or are based on three assumptions: (1) the mathematical form of the item characteristic functions, (2) the dimensionality of the item space and (3) the distribution of the abilities. Violations of these assumptions can produce serious negative consequences to the measurement process. One way of evaluating the impact of violating these assumptions, as well as studying factors such as the impact of choice of models, examinee sample sizes, the shape of ability distributions, and test length, and many other factors, is via

simulation studies, also referred to as Monte Carlo studies.

Monte Carlo studies have become an important tool in the development of theory and methods. For example, if the properties of an estimator or fit statistic are very difficult to work out analytically, a Monte Carlo study may be conducted to estimate those properties. Monte Carlo studies often provide a significant amount of the available knowledge of the properties of statistical techniques, especially under various alternative models. A large proportion of the articles in the statistical literature include Monte Carlo studies. For example, in recent issues of the *Journal of the American Statistical Association* almost half of the articles report on Monte Carlo studies that supported the research.

Simulation brings to the surface inconsistencies and inefficiencies, for example in situations where the sample size is small; the number of dimensions is large; and/or a mixture of different model types (for example the 2-parameter logistic, graded and nominal) are fitted to the items. A further example is deviations from model assumptions. When the same simulation procedure with a specified starting random seed is repeated several times, the generated data can be used to assess accuracy of the parameter estimates and the robustness of standard errors and fit statistics. In practice, it is often difficult and costly to repeat the exact circumstances under which data was collected. In contrast, with simulation software one can test the same system repeatedly in a time-efficient manner with different inputs and under different scenarios.

The purpose of the IRTPRO simulation module is to simulate examinee item response data given true model parameter values (both items and subjects). The simulation module creates data files in a form that can be directly run in the IRTPRO software as well as being saved for future use.

IRTPRO has the capability of generating data from all implemented IRT models, including multidimensional, bifactor, and multilevel models. The command syntax for generating data has the same basic structure as the model-fitting syntax.