

References of Non-Commercial Software for IRT Analyses¹

Nina Deng
University of Massachusetts Amherst

Please send any comments, updates, or corrections to (ndeng@educ.umass.edu) . Thank you very much for the support and I hope you find this brief report helpful!

A Computer Program for Simulation Evaluation of IRT Ability Estimators

Author: David Thissen

Source: <http://eric.ed.gov/>

Capabilities:

A computer program for simulation evaluation of item response theory (IRT) ability estimators.

Applicable Models:

Not mentioned.

Features:

- Contains a program which graphs the robust simulation results.
- Simulated in a unidimensional test.
- Published in Nov. 1984 by ETS and restricted to microfiche of Education Resources Information Center (ERIC).

ADTEST

Authors: Javier Revuelta, Vicente Ponsoda, & Julio Olea.

Source: *Applied Psychological Measurement*, Vol. 17 No. 1, March 1993, p. 28.

Capabilities:

A program implements the computerized adaptive testing (CAT) algorithm based on three parameter logistic model.

Applicable Models:

The three-parameter logistic model.

Features:

- Trait estimates are estimated by maximum likelihood , using the Newton-Raphson method.
- Data of each selected item, difference between stimulatees' estimates and true trait levels, and standard error of trait level estimation are provided.
- Coded in Turbo Pascal 6.0 and work on PC-compatible computers.

ANALYSIS

Author: ReLabs Research Laboratories Ltd.

Source: http://www.relabs.org/pb/wp_37493a8c/wp_37493a8c.html

Capabilities:

¹ **Center for Educational Assessment Research Report No. 699.** Amherst, MA: University of Massachusetts, Center for Educational Assessment. The author is grateful to Professor Ronald Hambleton for his assistance with the project.

A free program manages data and calibrates item and person parameters based on Rasch models.

Applicable Models:

Simple Logistic Rasch Model, Partial Credit Model, and Rating Scale Model.

Features:

- User-friendly and Windows interfaced.
- Allows users to import and export data to and from SPSS, Excel, Access, etc. Has many data management features.
- Provides classical item descriptive statistics.
- Compares the means of performance of different groups.
- Produces graphs including ICCs and frequency charts.

BB-CLASS v 1.1

Author: Robert L. Brennan

Source: The University of Iowa,

http://www.education.uiowa.edu/casma/computer_programs.htm

Capabilities:

An ANSI C computer program that uses the beta-binomial model (and its extensions) for classification consistency and accuracy based on Hanson and Brennan (1990) and Livingston and Lewis (1995) procedures.

BIGSTEPS

Source: <http://www.winsteps.com/bigsteps.htm>

Capabilities:

A free DOS-based Rasch measurement program. It has most of the functionality of WINSTEPS, but lacks a Windows interface and recent enhancements. Its capacity is 3,000 items and 20,000 persons (cases).

BIRT

Author: Frank Baker

Source: <http://edres.org/irt/baker/software.htm>

Capabilities:

A software package that accompanies the *Basics of Item Response Theory* book while learning and reviewing item response theory.

Features:

- Originally written in AppleBasic and later converted to Visual Basic 5.0. A bit old interface.
- Runs under WINDOWS 95 +.
- Requires the Visual Basic 5.0 run-time package, Msvbvm50.dll, and MSFLXGRD.OCX.

CIPE

Author: Michael J. Kolen

Source: Center for Advanced Studies in Measurement and Assessment, University of Iowa.

http://www.education.uiowa.edu/casma/computer_programs.htm#equating

Capabilities:

Common item program for equating performs mean, linear, and equipercentile equating in the common item nonequivalent groups design.

Features:

- Implements the equating methods including Tucker mean (TMEAN), Levine mean for internal common items (LMEAN), Braun/Holland mean (BMEAN), Tucker linear (TLIN), Levine linear for internal common items (LLIN), Braun/Holland linear (BLIN), unsmoothed frequency estimation equipercentile (UNSMOOTHED), and smoothed frequency estimation equipercentile, with up to 8 different degrees of cubic spline smoothing.
- calculates standard errors of equating for the Tucker linear, Levine linear, and unsmoothed equipercentile methods.

ConstructMap (formerly GradeMap)

Source: http://bearcenter.berkeley.edu/GradeMap/index.php?page_id=1

Capabilities:

A software package combines a multidimensional IRT engine for estimating item and person parameters with tools for managing cross-sectional and longitudinal student response data and interpreting findings from such data.

Applicable Models:

Multidimensional IRT models.

Features:

- Graphical maps and reports are designed for use in settings in which progress on multiple measures can be examined and analyzed.
- Users can select expected a posteriori (EAP), maximum likelihood, or plausible value estimates of multivariate proficiency estimates.
- Accepts dichotomous, rating scale, or partial credit items in between-item (each response is an indicator of a single dimension) or within-item (a response may be an indicator of multiple dimensions) multidimensional models.
- Produces Wright maps that align person estimates with item estimates on a logit scale, and item characteristic and cumulative probability curves.
- Differential item functioning and item bias can be explored by partitioning the response data on user-defined grouping criteria.
- Traditional item-analysis statistics and modeling fit statistics are produced.
- Graphical and menu-driven.

DFITD4

Source: <http://work.psych.uiuc.edu/irt/>

Capabilities:

A program implements DFIT (Differential Functioning of Items and Tests) developed by Raju, van der Linden, & Fleer (1995), which detects DTF/DIF by comparing test characteristic curves (TCCs).

Features:

- Identifies DIF items and computes DTF, and DTF is found, determines which DIF items, if any, should be removed to establish measurement equivalence.
- Linking coefficients, item parameters, and latent trait scores are required to compute the TCCs.

DIFCUT

Authors: James Alice O. Nanda, T. C. Oshima, & Phill Gagne

Source: *Applied Psychological Measurement*, Vol. 30 No. 2, March 2006, p. 150–151.

Capabilities:

A program conducts significance tests for differential functioning of items and tests (DFIT) for dichotomously scored test data using item response theory.

Applicable Models:

The models from BILOG-MG3

Features:

- Uses method of item parameter replication (IPR) to determine the cutoff scores.
- Calculates actual DIF and DTF and identifies the level of significance for each item and the test as a whole.
- Can be modified to accommodate any other DIF indices that make use of IRT-based item parameter estimates.
- Written in SAS/IML and runs on SAS-PC.

DIMENSION

Authors: John Hattie & Krzysztof Krakowski

Source: *Applied Psychological Measurement*, 1993, 17(3), 252.

Capabilities:

A program generates item response data according to several unidimensional and multidimensional item response models.

Applicable Models:

The compensatory and noncompensatory models

Features:

- Assumes that examinee trait levels are normally distributed.
- The models, the number of items/variables(max of 60), dimensions(max of 5), number of examinees (max of 1000) can be selected.

DRAWICC

Author: Christine DeMars

Source: *Applied Psychological Measurement*, Vol. 24 No. 3, September 2000, p. 224

Capabilities:

A program reads item parameter files created by PARSCALE or BILOG, and graphs the item response functions, and the item information functions for all items.

Applicable Models:

Models fit BILOG or PARSCALE.

Features:

- Runs on any Windows-based computer with SAS/GRAPH installed

EO-FIT

Authors: Pere J. Ferrando & Urbano Lorenzo-Seva

Source: *Educational and Psychological Measurement*, 2001, 61 (5), 895-902.

Capabilities:

A program checks the model-data fit of unidimensional logistic item response models for binary and ordered polytomous responses based on comparison of observed and expected test score distributions.

Applicable Models:

The one-, two-, and three-parameter logistic models, Samejima's graded response model (GRM) and Masters' partial credit model (PCM).

Features:

- Makes extensive use of graphical displays.
- An additional χ^2 -type statistic is reported.
- Allows cross-validation procedures to be used.
- Allows the fit of different models to be compared.
- Developed in Visual C++ *Applied Psychological Measurement* and executed under the Microsoft Windows 95/98/NT operative system.

EQUATE

Author: Frank B. Baker, Ali Al-Karni, & Ibrahim M. Al-Dosary.

Source: *Applied Psychological Measurement*, 1991, 15 (1),78.

Capabilities:

A program implements test characteristics curve equating procedure due to Stocking & Lord (1983).

Applicable Models:

Not mentioned.

Features:

- Uses item parameter estimates of the common anchor items to compute the scale (A) and intercept (K) coefficients of the linear transformation of the ability metric.
- Both transformed item and ability estimates are stored in files named by the user in a standard format.
- Written in Professional FORTRAN for MS-DOS computers.

EQUATE 2.0

Author: Frank B. Baker.

Source: *Applied Psychological Measurement*, Vol.17 No. 1, March 1993, 20.

Capabilities:

A program implements the test characteristics curve method of test equating for dichotomously, graded and nominally scored items.

Applicable Models:

Models for dichotomous response items, and graded or nominal response items.

Features:

- Extends the capabilities of EQUATE (Baker, 1991) by including graded and nominal scored items.
- Accepts input files produced by MULTILOG or produced in a user-supplied format.
- Written in FORTRAN for DOS computers.
- Runs interactively with the user supplying the necessary specifications.

eRm: extended Rasch models

Authors: Patrick Mair & Reinhold Hatzinger.

Source: <http://cran.r-project.org/src/contrib/Descriptions/eRm.html>

Applicable Models:

Rasch models (RM), linear logistic test models (LLTM), rating scale model (RSM), linear rating scale models (LRSM), partial credit models (PCM), and linear partial credit models (LPCM).

FACET 3.22

Source: <http://www.winsteps.com/facdos.htm>

Capabilities:

A free DOS-based Rasch measurement program. It has most of the functionality of the current version of Facets, but lacks a Windows interface and recent enhancements. Does not run under Windows XP Professional x64 Edition. Its capacity is about 20,000 persons (elements).

FIRESTAR

Author: Seung W. Choi

Source: Northwestern University, Feinberg School of Medicine, Center on Outcomes, Research and Education.

<http://depot.northwestern.edu/~swc807/>

Capabilities:

A computer program for simulating computerized adaptive testing (CAT) with polytomous items. Designed to run on Windows-based computers with R installed.

Applicable Models:

Samejima's graded response model (GRM), Muraki's generalized partial credit model (GPCM), Master's partial credit model (PCM), and Andrich's rating scale model.

Features:

- Provides various item selection techniques, stopping criteria, interim and final theta estimators, and output files.
- Provides choice of exposure control, prior distribution, first item selection, and standard error calculation methods.
- R code can be generated by the software.

FREEIRT Project Programs

Source: <http://freeirt.org/index.php?file=database/edittheme.php&new=yes>

Capabilities:

A website consists of programs applied in wide areas of Rasch models and other measurement applications.

Applicable Models:

Rasch models, and other IRT models.

Format_PCI.sas & Format_ICC.sas

Author: Chong Ho Yu

Source: *Applied Psychological Measurement*, Vol. 30 No. 3, May 2006, 247–248.

Capabilities:

Format_PCI.sas is a macro SAS program formatting the input file for Winsteps, and Format_ICC.sas is for adding graphical presentation of the Winsteps item parameter output.

Applicable Models:

The models from Winsteps

Features:

- A document titled “sf.html” is included with the package to help beginners interpret the step function yielded from partial-credit items.
- The graphical presentations include TIF, IIF, TCC, and ICC.
- A file “report.html” is created as exam-level report including item parameter information.
- The results are Web ready and can be shared among colleagues through the Internet or IntraNet.
- Requires SAS Version 9.1.3

GGUM2000

Authors: James S. Roberts

Source: *Applied Psychological Measurement*, Vol.25, No. 1, March 2001, 38.

Capabilities:

A program estimates item parameters in the GGUM using marginal maximum likelihood. It derives person estimates using an expected a posteriori approach.

Applicable Models:

The generalized graded unfolding model (GGUM), and seven other constrained versions of the model.

Features:

- Allows for 100 items, with up to 10 response categories per item, and up to 2,000 respondents
- Output includes parameter estimates, associated standard errors, and various indices of model, item, and person fit
- Runs under MS-DOS or in an MS-DOS shell under Windows 95/98
- With use’s guide

GGUM 2004

Authors: James S. Roberts, Haw-ren Fang, Weiwei Cui, & Yingji Wang

Source: *Applied Psychological Measurement*, Vol. 30 No. 1, January 2006, 64–65.

Capabilities:

A program estimates parameters for a family of unidimensional unfolding item response theory (IRT) models.

Applicable Models:

The generalized graded unfolding model (GGUM), and seven other models derived.

Features:

- Includes and extends the capabilities of GGUM2000.
- Allows the number of response categories to vary across items.
- Allows for missing item responses under the assumption that those responses are missing at random.
- Calculates new item fit statistics and information criteria relating to model fit.
- Can run under the Windows 98SE, Windows 2000 Professional, and Windows XP Professional operating systems.

GGUMLINK

Author: ROBERTS James S. & Chun-wei Huang.

Source: *Behavior research methods, instruments & computers*, Vol. 35 No. 4, 2003, 525–536.

<http://www.education.umd.edu/EDMS/tutorials/index.html>

Capabilities:

A computer program links parameter estimates of the generalized graded unfolding model from item response theory.

Features:

- Reexpresses parameter estimates from two separate GGUM calibrations in a common metric.
- Secures a common metric by using one of five methods that have been generalized to the GGUM.

GR-GRAPH

Authors: David M. Gudanowski, Dawn L. Vreven, Lynda A. King, & Daniel W. King.

Source: *Applied Psychological Measurement*, Vol. 18 No. 3, September 1994, 292.

Capabilities:

A program generates values and produces graphs and tables for item response theory analysis.

Applicable Models:

Samejima's (1969) graded response model.

Features:

- The items must use a 5-point Likert-type rating format.
- Provides both graphical and tabular forms.
- Provides values for the operating response functions (ORFs), item information functions (IIFs), and test information functions.
- Written in Quattro Pro Templates and Macros and runs in DOS.

GRAPHDIF

Author: John H. Neel.

Source: *Applied Psychological Measurement*, Vol. 18 No. 3, September 1994, 299.

Capabilities:

A program identifies differential item functioning (DIF) by calculating the area between item response functions (IRFs), and using graphic displays.

Applicable Models:

Not mentioned

Features:

- Graphs are labeled with the areas (shaded) and item parameters, and are color-coded to item files.
- Items can be sorted by different values to assist DIF.
- The number of graphs displayed at one time ranges from 1 to 56.
- Written in C++ for DOS.

GUMJML

Authors: James S. Roberts, Kevin Lee & T.C. Oshima

Source: *Applied Psychological Measurement*, Vol. 22 No. 1, March 1998, 70

Capabilities:

A program estimates parameters of the graded unfolding model using a joint maximum likelihood approach.

Applicable Models:

The graded unfolding model.

Features:

- Operates up to 200 test items and 2,000 examinees.
- The number of response categories with each item ranges from 2 to 9.
- Offers several diagnostic indexes of b item and person fit.
- A DOS-based system written in Microsoft PowerStation FORTRAN.

HSGEN (Hierarchical Score Generator)

Author: Kyung T. Han

Source: <http://www.umass.edu/remf/software/hsgen/>

Capabilities: Generates examinees' scores mostly on IRT scale in a hierarchical structure (say, students' scores within each school and/or within each district).

Features:

- Specify a mean value for the highest level groups and standard deviations for each level.
- Specify the number of groups (or individuals at the lowest level) for each level.
- Handle up to 5 levels.

ICL- IRT Computer Language (Version 0.020301, March 2002)

Author: Bradley A. Hanson

Source: <http://www.b-a-h.com/software/irt/icl/>

Capabilities:

A computer program performs single- or multiple-group estimation for dichotomous items, and polytomous items.

Applicable Models:

The 1-, 2-, and 3-parameter logistic item response models, the partial credit model and generalized partial credit model.

Features:

- Compiled in versions for Windows 95/NT, Macintosh, and Linux.
- Written in C++ using the [Estimation Toolkit for Item Response Models](#) (ETIRM).
- A couple of bugs were found and have been fixed in the [CVS repository](#) at the SourceForge website

IPARM

Author: Richard M. Smith

Source: IPARM: *Item and person analysis with the Rasch model*. 1991. Chicago: Mesa Press.

Capabilities:

A program provides item analysis and person-fit statistics for Rasch model.

Features:

- Uses a variation of KIDMAP output (Wright, Mead, and Ludlow, 1980) to produce instructionally useful diagnostic information in score reporting.
- Performs item and person analysis at a higher level of detail than calibration programs.
- Creates detailed maps of examinee performance for fixed length or computer adaptive tests.

IPLINK

Authors: Kevin Lee & T. C. Oshima

Source: *Applied Psychological Measurement*, Vol. 20 No. 3, September 1996, 230.

Capabilities:

A program estimates linking coefficients which place item parameter estimates from separate calibrations onto a common trait metric for test data that are multidimensional.

Applicable Models:

Multidimensional and unidimensional models. Default settings for those from BILOG and NOHARM are provided.

Features:

- Minimizes the differences between two sets of functions of item parameter estimates using one of the four methods described in Oshima, Davey, & Lee (1996).
- More than one pair of input data files can be entered. Multiple runs are allowed.
- Written in Turbo C++ and runs under Win 3.1 and Win 95.

IRT-CLASS v2.0

Author: Won-Chan Lee & Michael J. Kolen

Source: Center for Advanced Studies in Measurement and Assessment,
The University of Iowa.

<http://www.education.uiowa.edu/casma/DecisionConsistencyPrograms.htm>

Capabilities:

A FORTRAN computer program that computes classification consistency and accuracy indices for raw and scale scores.

Applicable Models:

The three-parameter logistic, normal ogive graded response (Samejima, 1997), logistic graded response (Samejima, 1997), generalized partial credit (Muraki, 1997), and nominal response (Bock, 1997) models.

Features:

Outputs the consistency index (ϕ), the kappa coefficient, accuracy index (γ), and false positive and false negative error rates.

IRTDIF

Authors: Seock-Ho Kim & Allan S. Cohen.

Source: *Applied Psychological Measurement*, Vol.16 No. 2, June 1992, 158.

Capabilities:

A program uses item response theory to provide measures of differential item functioning (DIF).

Applicable Models:

The one-, two-, three-parameter IRT models.

Features:

- Provides DIF measures including Lord's χ^2 statistics, the exact area measures (Raju, 1990), and the closed-interval area measures (Kim & Cohen, 1991).
- Written in IBM Professional FORTRAN.
- Execution of the program requires a numerical coprocessor.

IRTEQ

Author: Kyung T. Han

Source: <http://www.umass.edu/remf/software/irteq/>

Capabilities:

Implement various IRT based IRT scaling/equating methods with an anchor test design. The methods include Mean/Mean, Mean/Sigma, Robust Mean/Sigma, and TCC methods (Haebara; Stocking & Lord).

Applicable Models:

Logistic models for dichotomous responses (with 1, 2, or 3 parameters), Generalized Partial Credit Model (GPCM) (including Partial Credit Model (PCM), for polytomous responses, a mixture of IRT models. The number of response categories can vary.

Features:

- No limit for the number of items in each form or of the linking items (>10,000,000).
- Provide the option of various score distributions with TCC methods.

- Import item parameters and score data from WinGen (Han, 2007) and/or PARSCALE.
- Provide true scoring equating with test score conversion table provided.
- Test characteristic curves of each test form, and a, b, and c-parameter plots of the linking items are provided.

IRTFIT

Authors: Jakob B. Bjorner, Kevin J. Smith, Clement Stone, & Xiaowu Sun.

Source: QualityMetric Incorporated, School of Education, University of Pittsburgh
http://outcomes.cancer.gov/areas/measurement/irt_model_fit.html

Capabilities:

A SAS macro for item fit and local dependence tests under IRT models.

Applicable Models:

The 1-, 2-, and 3-parameter models; the graded response model, the generalized partial credit model, the generalized rating scale model, the partial credit model, the rating scale model, and the nominal categories model.

Features:

- Produces a variety of fit statistics including extensions of the S-X2 and the S-G2 tests for polytomous items and the X* and G* statistics as well as tests for local dependencies between pairs of items.
- Provides observed-expected fit plots.
- Reads the parameter files from various IRT softwares including MULTILOG, PARSCALE, BILOG, OPLM, and WINSTEPS.

IRTFIT-RESAMPLE

Authors: Clement A. Stone

Source: *Applied Psychological Measurement*, Vol. 28 No. 2, March 2004, 143–144.

Capabilities:

A program evaluates the fit of item response theory models based on posterior expectations when ability is estimated imprecisely.

Applicable Models:

Graded logistic response model; one-, two-, and three-parameter dichotomous logistic response models; generalized partial-credit model.

Features:

- Posterior probabilities are used to compute the goodness-of-fit statistics.
- A Monte Carlo resampling procedure is used for statistical test.
- Item sets representing mixed models can be evaluated.
- Output item-fit tables and graphical displays are produced.
- Written in SAS, accompanying manual.

IRTGEN

Authors: Tiffany A. Whittaker, Steven J. Fitzpatrick, Natasha J. Williams, & Barbara G. Dodd.

Source: *Applied Psychological Measurement*, Vol. 27 No. 4, July 2003, 299–300.

Capabilities:

A program to generate known trait scores (theta values) according to the distribution and item responses for examinees based on the IRT models.

Applicable Models:

The graded response, partial credit, generalized partial credit, rating scale, successive intervals, and three-parameter logistic models.

Features:

- User can specify either a normal or uniform distribution under which the known trait scores will be generated
- Able to generate responses to items with differing numbers of categories when one of the polytomous IRT models is used
- A SAS input data set needed including item parameters, name of IRT model, number of items and examinees
- Written in SAS macros
- Manual with SAS codes and examples available

IRTGRAPH

Authors: Ruth A. Childs & Karen Schlumpf

Source: *Applied Psychological Measurement*, Vol. 23 No. 3, September 1999, 262.

Capabilities:

A set of programs streamlines the graphs produced by MULTILOG and PARSCALE including item response functions, item and test information functions, and parameter comparison scatterplots.

Applicable Models:

(1) From MULTILOG, the one-, two-, and three-parameter logistic, and graded response models; (2) from PARSCALE, the two- and three-parameter logistic, graded response, and generalized partial credit models.

Features:

- Particularly useful for producing large numbers of uniformly formatted graphics
- Process files containing estimates for a single model
- Written in SAS macro language

IRTINFO

Authors: Steven J. Fitzpatrick, Seung W. Choi, Ssu-Kuang Chen, Liling Hou, and Barbara G. Dodd.

Source: *Applied Psychological Measurement*, Vol. 18 No. 4, December 1994, 390.

Capabilities:

A program computes item and test information for different models.

Applicable Models:

The graded response model, the partial credit model, the generalized partial credit model, the rating scale model, the successive intervals model, and the three-parameter logistic model.

Features:

- Can handle items with differing numbers of categories.
- Written in SAS Macro and suitable for use by SAS procedures.

IRTLRDIF v. 2

Source: <http://www.unc.edu/~dthissen/dl.html>

Capabilities:

A software computes likelihood ratio tests of DIF.

Applicable Models:

The 3P logistic and graded IRT models.

Irtoys

Author: Ivailo Partchev

Source: <http://cran.stat.ucla.edu/src/contrib/Descriptions/irtoys.html>

Capabilities:

Provide a simple common interface to the estimation of item parameters and plotting in IRT models for binary responses with three different programs (ICL, BILOG-MG, and ltm).

Applicable Models:

Models with dichotomous data.

IRT Painter

Author: Ning Han

Source: Conference paper presented at annual meeting of National Council on Measurement in Education, 2003, Chicago, Illinois.

Capabilities:

A program is designed to produce graphical displays associated with the common used item response theory (IRT) models.

Applicable Models:

The 1P-, 2P-, and 3P- Logistic models, the Graded Response Model, and the Generalized Partial Credit Model.

Features:

- Draws Item Characteristic Curves (ICC), Test Characteristic Curves (TCC), Item Information Functions (IIF), Test Information Functions (TIF), and Standard Error of Measurement (SEM).
- The charts appear on the screen and can be printed, or copied into an editable file such as Microsoft Word.
- The charts can be accessed through Excel interface and user can edit every element of the charts such as title, legend, axis labels, and font.
- Reads the outputs of *BILOG*, *BILOG-MG*, and *PARSCALE* directly or text files created by the user.
- Runs on Windows 98/ME/2000/XP/NT. Microsoft Office must be installed.

IRTScore

Source: <http://www.unc.edu/~dthissen/dl.html>

Capabilities:

A software computes summed-score to EAP(theta) translation tables, and the values and weights used in linear IRT response-pattern scoring, given parameters from Multilog output files or space- or tab-delimited files.

ITERLINK

Source: <http://work.psych.uiuc.edu/irt/>

Capabilities:

A program performs iterative linking and pairwise DIF detection, and puts the focal group parameters on the reference group metric using Lord's chi-square statistics.

Applicable Models:

The 2P- and 3P-logistic models.

ltm: Latent Trait Models under IRT

Author: Dimitris Rizopoulos.

Source: <http://wiki.r-project.org/rwiki/doku.php?id=packages:cran:ltm>

Capabilities:

A package analyzes multivariate dichotomous and polytomous data using latent trait models under the Item Response Theory approach.

Applicable Models:

Rasch Model, Two Parameter Logistic Model, Birnbaum's Three Parameter Model, Graded Response Model and Latent Trait Model up to two latent variables.

LDIP

Authors: Seock-Ho Kim, Allan S. Cohen, & Yuan-Horng Lin

Source: *Applied Psychological Measurement*, Vol. 30 No. 6, November 2006, 509-510

Capabilities:

A program provides indices of local dependence for polytomous items under item response theory.

Applicable Models:

The graded response and generalized partial credit model from either MULTILOG or PARSCALE.

Features:

- The indices of local dependence include the Pearson chi-square statistic χ^2 , the likelihood ratio chi-square statistic G2, Yen's index of local dependence Q3, and the Fisher-transformed correlation difference statistic Zd.
- Written in Fortran and compiled with Microsoft Fortran PowerStation.
- Can be executed in a command prompt under Microsoft Windows XP and run in an interactive mode.

LINKDIF

Author: N.G. Waller.

Source: *Applied Psychological Measurement*, Vol. 22 No.4, December 1998, 392.

Capabilities:

A program links IRT item parameters and calculates several measures of differential test (DTF) and item functioning (DIF).

Applicable Models:

Not mentioned.

Features:

- IRT a and b estimates are linked to a common metric by the test characteristic curve (TCC) method by Stocking & Lord (1983).
- DIF and DTF measures include Lord's X^2 , the signed and unsigned and their z values, compensatory and noncompensatory DIF indexes.
- Written in S-PLUS.

LOGLIN/KE

Authors: Alina von Davier, Henry Chen, Paul Holland, Dorothy Thayer, Ning Han, Ting Lu, & Michelle Najarian.

Source:

A free copy for non-commercial use can be obtained upon request from Educational Testing Service. A license agreement needs to be signed and the signed original needs to be sent via regular mail before the software could be distributed (contact email: dlembeck@ets.org).

Capabilities:

LOGLIN/KE consists of two stand-alone computer programs. LOGLIN is designed to fit loglinear models for pre-smoothing for univariate or bivariate score distributions. KE is a software package designed for performing Kernel Equating, see von Davier, A. A., Holland P. W. and Thayer, D. T. (2004) *The Kernel method of test equating*, Springer: New York.

Features:

- Implements Kernel Equating, adopting Gaussian Kernel Smoothing for continuization of the discrete score distributions.
- Applicable for a variety of equating designs including using common persons or common items. Chain equating and post-stratification equating methods are implemented for designs using common items.
- The standard errors of equating (SEE), the standard errors of the differences between equating functions (SEED), and the difference between Kernel Equating and equipercetile equating using linear interpolation are provided.
- User-friendly and Windows interfaced. Various plots are outputted in Excel files.

LTDOMAIN

Author: Yuan Hwang Li.

Source: *Applied Psychological Measurement*, Vol. 19 No. 1, March 1995, 50.

Capabilities:

A program generates a look-up table for the corresponding estimated one-parameter logistic model scale score and the unbiased domain score for each number-correct score.

Applicable Models:

The one-parameter logistic model.

Features:

- Written in PASCAL and runs under MS-DOS.

mdltn

Authors: Matthias von Davier, Xueli Xu.

Source: A License for non-commercial use can be obtained upon request from Educational Testing Service (email:dlembeck@ets.org).

Capabilities:

A program which estimates a variety of latent variable models using the multidimensional discrete latent trait models (mdltn).

Applicable Models:

Latent class models, IRT 1PL, 2PL, Generalized Partial Credit Model, multiple group and mixture distribution IRT models, and General Diagnostic Models.

Features:

- Using the graphical user interface.
- Suitable for nominal, dichotomous, polytomous and mixed format datasets.
- Handling omitted data, missing data at random, weighted data, and multi-group estimation.
- Allows for parameter constraints across populations and test forms in the concurrent calibrations.

MINIFAC

Source: <http://www.winsteps.com/minifac.htm>

Capabilities:

A reduced version of FACETS, has complete FACETS functionality, but is limited to 2,000 data points (responses). Free of charge or time-limit.

MINISTEP

Source: <http://www.winsteps.com/ministep.htm>

Capabilities:

A evaluation/student version of WINSTEPS, has complete WINSTEPS functionality, but is limited to 25 items and 75 persons (cases). Free of charge or time-limit.

MODFIT

Source: <http://work.psych.uiuc.edu/irt/>

Capabilities:

A computer program plots theoretical item response functions and examines the fit of dichotomous or polytomous IRT models to response data.

Features:

- Computes fit plots and chi-squares for item singles, doubles, and triples using the method described by Drasgow et al. (1995).

- Computes item/option response functions, information functions, test characteristic curves, test information functions, and conditional standard errors.
- Visual Basic for Applications program created using Microsoft Excel 2000.
- All files must be saved in the same folder to run the program.

MOKSCAL

Authors: Johannes Kingma & Terry Taerum

Source: *Applied Psychological Measurement*, Vol.12 No. 2, June 1988, 188.

Capabilities:

A program for the Mokken (1971) scale analysis based on a nonparametric item response model.

Applicable Models:

Nonparametric item response model (Mokken & Lewis, 1982).

Features:

- Tests the assumption of double monotony in the Mokken model.
- Tests the robustness of an established Mokken scale across different groups.
- Computes three coefficients of scalability and decides whether they meet the criterion of monotone of homogeneity.
- Computes four different coefficients of reliability and the biserial correlations.
- Two versions are available, an SPSS-X user PROC file, and a stand-alone program, both written in FORTRAN 77.

MULT-CLASS v 3.0

Author: Won-Chan Lee

Source: The University of Iowa,

http://www.education.uiowa.edu/casma/computer_programs.htm

Capabilities:

A FORTRAN computer program that employs the multinomial and compound multinomial error models (Lee, 2007) for computing classification consistency and accuracy indices.

Features:

- An extension of BB-CLASS (Brennan, 2004) to multivariate situations.
- The multinomial model is used for a test with a single item set, and the compound multinomial model is used for a test consisting of mixtures of different item sets.
- Can be used for mixture of polytomous and dichotomous data.

MULTIRA V 1.65 (in German)

Authors: C. H. Carstensen & J. Rost

Source: www.multira.de

Capabilities:

A computer program implements an algorithm for the multidimensional item

component Rasch model and other Rasch models.

Applicable Models:

The Multidimensional Item-Component Rasch-Model (MultiRa)

Features:

- Get person and item parameter estimates according to the MultiRa-Model.
- Joint maximum Likelihood and conditional Maximum Likelihood Estimation.
- Modelfit measures including Martin Loef Test on Item Homogeneity. Residual Mean Square Item Fit, Bootstrap GoF.

NOHARM

Authors: Colin Fraser & R. P. McDonald

Source: <http://people.niagaracollege.ca/cfraser/download/>

Capabilities:

A Windows program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory based on theory developed by R. P. McDonald.

OpenStat

Author: William G. Miller

Source: <http://www.statpages.org/miller/openstat/>

Capabilities:

A general program for the analyses in statistics and measurement.

PARAM-3PL/IPL

Author: Lawrence M. Rudner

Source: <http://edres.org/irt/param/>

Capabilities:

A program written in PowerBasic calibrating items and individuals using Newton-Raphson maximum likelihood estimation.

Applicable Model:

The 3 parameter logistic item response theory model; Rasch model.

Features:

- Calibrate either ability given item responses and item parameters, or item parameters given ability and item responses, or a, b parameters given c parameter, ability and responses (in 3PL), or ability, item parameters given item responses.
- No limit of the number of examinees; a maximum number of items of 9,999.
- Click-and-point. User-friendly.

PIE & POLYEQUATE

Author: B. A. Hanson & Zeng, L. (PIE)

Kolen, M. J. & Brennan, R. L. (POLYEQUATE)

Source: <http://www.education.uiowa.edu/casma/EquatingLinkingPrograms.htm>

Capabilities:

Computer programs conduct IRT true and observed score equating analyses described in Kolen and Brennan (2004).

Applicable Models:

PIE conducts IRT true and observed scoring equating for dichotomously scored tests.

POLYEQUATE conducts IRT true and observed scoring equating for dichotomously and polytomously scored tests.

PLotIRT

Authors: Cheryl D. Hill & Michelle M. Langer

Source: *Applied Psychological Measurement*, Vol. 31 No.5, September 2007, 456.

<http://www.unc.edu/~dthissen/dl.html>

Capabilities:

A set of R functions facilitates plotting a variety of curves associated with item response theory.

Applicable Models:

The two-parameter logistic, three-parameter logistic, graded, nominal, and multiple-choice models.

Features:

- Handles curves including item and test characteristic curves, item and test information curves, expected item score functions, and posterior density distributions.
- The user can specify colors, labels, the number of plots in a window, the number of items within a plot, font sizes, and a variety of other options.
- Basic R functions can be used to read in an assortment of item parameter files, or item parameters can be directly typed into R arrays.
- Runs on any Windows or Mac-based computer with R installed.

PML

Author: Karl Bang Christensen

Source:

<http://freeirt.org/index.php?file=database/detailpgm.php&cond=idprogram=25>

Capabilities:

A SAS macro that can be used to test the assumption of unidimensionality in a polytomous Rasch model.

POLYBIF

Authors: Robert D. Gibbons & Donald Hedeker.

Source: University of Illinois at Chicago, Center for Health Statistics,

<http://www.uic.edu/labs/biostat/bifactor.html>

Capabilities:

Carries out the full-information item bifactor analysis for ordinal and dichotomous data.

Features:

- Number of categories and thresholds can vary over the items.

- Item can load on the primary dimension only (essential unidimensional).
- Limitations: 15 factors, 10 quad points, 100 items, 5000 patterns, and 10 categories.

POLYCSEM

Author: Michael J. Kolen.

Source: <http://www.education.uiowa.edu/casma/DownloadOtherPrograms.htm>

Capabilities:

A Fortran 77 computer programs used to estimate conditional standard errors of measurement and reliability of raw and scale scores for tests under an IRT model.

Applicable Models:

The three-parameter logistic, Samejima's normal ogive graded response (Samejima, 1997), Samejima's logistic graded response (Samejima, 1997), Bock's nominal model (Bock, 1997), and Muraki's generalized partial credit model (Muraki, 1997).

POLYST

Author: Seonghoon Kim & Michael J. Kolen

Source: The University of Iowa,

http://www.education.uiowa.edu/casma/computer_programs.htm#equating

Capabilities:

An ANSI C computer program for implementing item response theory (IRT) scale transformation methods to place item parameter estimates from a scale of a group of examinees to a scale determined by a base or reference group of examinees.

Applicable Models:

1) three-parameter logistic (3PL) model, 2) graded response (GR) model, 3) generalized partial credit (GPC) model, 4) nominal response (NR) model, 5) multiple-choice (MC) model.

Features:

- Implements four scale transformation methods including the mean/sigma (Marco, 1977), mean/mean (Loyd & Hoover, 1980), Haebara (Haebara, 1980) and Stocking-Lord (Stocking & Lord, 1983) methods.
- Estimates the slope and intercept of a linear transformation relating the two IRT scales resulting from separate calibrations.
- No limitation in the number of common items and the number of response categories for an item. All items do not need to have the same number of response categories as well.

PRASCH

Author: John M. Grego

Source: *Applied Psychological Measurement*, Vol. 17 No. 3, September 1993, 238.

Capabilities:

A program models latent class polytomous response Rasch models using Conditional Maximum Likelihood Estimates (CMLEs).

Applicable Models:

The latent class polytomous response Rasch models.

Features:

- Provides parameter estimates, goodness-of-fit statistics, posterior mean scores, and conditional probabilities of item response levels for a given latent class.
- Finds CMLEs and tests a moment condition for polytomous response Rasch model.
- Computes the likelihood gradient function for a latent class model.
- Compiled on a DEC VAX8300 using the VS Fortran compiler.

PRED

Author: Ning Han

Source: University of Massachusetts Amherst, Center of Educational Assessment.

Capabilities:

Predicts the score distribution by Monte-Carlo simulation of item responses when the examinee ability estimates and the item parameter estimates are known.

Applicable Models:

The 1p, 2p, and 3p logistic dichotomous and polytomous models.

Features:

- The item parameter files are the outputs of BILOG, BILOG-MG, and PARSCALE.
- The user selects whether an existed ability parameter is used or normal distribution is assumed.
- The simulation times can be specified.

RASCH/ECIZ

Authors: Randall B. Nelson & Steven P. Chatman

Source: *Applied Psychological Measurement*, Vol.9 No. 3, September 1985, 325.

Capabilities:

A SAS PROC MATRIX program computes item and person parameters for the Rasch model, and provides several person-fit statistics.

Applicable Models:

Rasch model.

Features:

- Uses the PROX method and the iterative maximum-likelihood routine (Wright & Stone, 1979) to produce estimates of the model's parameters.
- Applies Anderson's (1973) correction for the biased estimates.
- Reports the standard errors of parameter estimates and mean square fit statistics.
- Five measures of person fit are reported including mean square fit, ECI2, ECI4, ECIZ2, and ECIZ4.

Rasch Scaling Program (RSP)

Authors: Cees Glas & Jules Ellis

Source of Review: *Applied Psychological Measurement*, Vol. 23 No. 1, March 1999, 90-94.

Capabilities:

A software package applies the one-parameter logistic (Rasch) model to dichotomously scored item response data. Applications of the model might include (1) estimating statistics for items (2) estimating latent trait scores (3) identifying item bias (4) designing tests (5) identifying aberrant persons

Applicable Model:

Rasch Model

Features:

- Includes conditional and marginal maximum likelihood (MML) estimation
- Include an array of fit statistics for persons, items, and score distributions
- Handles incomplete data designs
- “Caution indices” related to person fit
- Contains a manual and Rasch model introduction
- Run under MS-DOS or from within Win3.1 or Win95

RASCHSIM

Authors: G. Edward Miller

Source: *Applied Psychological Measurement*, Vol. 26 No. 3, September 2002, 355.

Capabilities:

A algorithm creates simulated data sets of observations and item variables which simulate realizations, and computes Rasch difficulty and ability parameter estimates and their standard errors..

Applicable Model:

Rasch model

Features:

- Uses the unconditional maximum likelihood estimation algorithm
- The first 43 lines of the algorithm code are comment and instruction statements
- Used in SAS version 6.0 or above

RASCHTEST

Author: Jean-Benoit Hardouin.

Source: <http://ideas.repec.org/c/boc/bocode/s439001.html>

Capabilities:

A program estimates the parameters of a Rasch model.

Features:

- The estimation method can be chosen between conditional maximum likelihood (CML), marginal maximum likelihood (MML) and generalized estimating equations (GEE).
- Offers a set of tests, to valuate the fit of the data to the Rasch model, or detect non homogeneous items.
- Several graphical representations can be easily obtained: comparison of the observed and theoretical Item Characteristic Curves (ICC), Map

difficulty parameters/Scores, results of the split tests, and information function.

RBF.sas

Author: Edward W. Wolfe

Source: *Applied Psychological Measurement*, Vol. 32 No. 7, October 2008, 585-586.

Capabilities:

A SAS Macro for estimating critical values for Rasch model fit statistics.

Applicable Model:

Dichotomous, rating scale, and partial credit Rasch model

Features:

- Utilizes a bootstrap procedure to estimate critical values for item and person fit statistics produced by WINSTEPS.
- The 2.5th, 5th, 95th and 97.5th percentile values are computed for fit statistics including the unweighted and weighted mean square, and the standardized unweighted and weighted mean square, item-total score correlations, item slope estimates, and item lower asymptote estimates.
- The user can specify the number of simulated data sets.

RESFIT

Author: Yue Zhao

Source: University of Massachusetts Amherst, Center of Educational Assessment.

Capabilities:

An R program that generates item fit plots and produces residuals.

Features:

- Input data are read from PARSCALE.
- Various graphical presentations.

RESGEN Item Response Generator. 1990 Version 1.01

Author: Eiji Muraki

Source: Research report at Educational Testing Service. Report No. ETS-RR-92-7.
http://eric.ed.gov/ERICDocs/data/ericdocs2/content_storage_01/0000000b/80/27/38/a8.pdf

Capabilities:

A computer program designed to generate simulated latent trait distributions and then dichotomous or polytomous item responses based on item response models.

Applicable Models:

(1) one-, two-, or three-parameter model; (2) logistic or normal ogive model; (3) unidimensional or multidimensional model; (4) dichotomous or polytomous model; and (5) graded response or partial credit model.

Features:

- The latent trait distributions can be univariate or multivariate normal, log-normal, uniform, or gamma.

- Simulates realistic testing situations by employing multiple matrix sampling designs, including multiple blocks, multiple subtests (booklets), multiple groups, multiple latent trait dimensions, and multiple sampling units.
- Free of charge if used within ETS.

ResidPlots-2

Authors: Tie Liang, Kyung T. Han, & Ronald K. Hambleton

Source: University of Massachusetts Amherst, Center for Educational Assessment.

<http://www.umass.edu/remf/software/residplots/>

Applied Psychological Measurement, Vol. 33 No. 5, July 2009, 411-412.

Capabilities:

A computer program produces the residual plots for IRT models.

Applicable Models:

The 1p, 2p, and 3p dichotomous models, and the polytomous models including graded response model, generalized partial credit model, and nominal response model.

Features:

- Input file read from PARSCALE, BILOG-MG, or Multilog syntax file, and three sets of fit plot are produced including item-level, test-level, and test score distribution plots.
- Users can specify the confidence bands and number of intervals.
- Allows users to decide what type of error bars to display and to specify the number of standard errors represented by the error bars.

SAS macro-program "%Detect"

Author: Jean-Benoit Hardouin

Source: <http://anaqol.org/index.php?file=editonepgm.php&idpgm=5&fr=>

Capability:

“%Detect” allows computing, for a partition of the items, the DETECT, Iss and R indexes defined by Zhang, et al. (1999). These indexes allow evaluating the psychometric qualities of a partition of dichotomous items. The greatest partition of items maximizes the DETECT index. A good partition had Iss and R indexes close of 1.

SAS macro-program “%MSP”

Author: Jean-Benoit Hardouin

Source: <http://anaqol.org/index.php?file=editonepgm.php&idpgm=2&fr=>

Capability:

The Mokken Scale Procedure (MSP) allows selecting items which verify the three fundamental assumptions of the Item Response Theory (IRT): Unidimensionality, monotonicity, local independence. The items are selected by using the Loevinger H coefficients (Loevinger, 1948). This procedure is proposed by Hemker, Sijtsma and Molenaar (1995).

Splus/R Program—Mokken Version 0.1

Author: L. Andries Van Der Ark

Source:

<http://freeirt.org/index.php?file=database/detailpgm.php&cond=idprogram=37>

Capabilities:

A Splus/R package contains principal functions for Mokken scale analysis.

SCD (statistical analysis of categorical data) & DIGRAM

Author: Svend Kreiner

Source: www.biostat.ku.dk/~skm/skm/

Capabilities:

Supports analyses by graphical models, loglinear Rasch models, multi-dimensional Markov chain models, analysis of marginal and conditional homogeneity of repeated ordinal measurements, of collapsibility of categories in multidimensional contingency tables by stepwise multiple comparison analysis, non-parametric analysis of ordinal data by loglinear models.

ST, POLYST & STUIRT

Source: <http://www.education.uiowa.edu/casma/IRTPrograms.htm>

Capabilities:

Computer programs conduct various IRT scale transformations.

Applicable Models:

ST conducts item response theory (IRT) scale transformations for dichotomously scored tests.

POLYST conducts IRT scale transformations for dichotomously and polytomously scored tests.

STUIRT conduct IRT scale transformations for mixed-format tests.

Features:

Windows PC console and graphical user interface (GUI) versions and Macintosh OS9 console and OS10 GUI versions are available for at least some of the programs.

SIMCAT 1.0

Authors: Gilles Rai che & Jean-Guy Blais

Source: *Applied Psychological Measurement*, Vol. 30 No. 1, January 2006, 60–61.

Capabilities:

A program simulates adaptive testing sessions under different adaptive expected a posteriori (EAP) proficiency-level estimation methods.

Applicable Models:

The one-parameter Rasch logistic model

Features:

- Also computes empirical and estimated skewness and kurtosis coefficients, such as the standard error.
- Allows comparing empirical and estimated properties of the estimated proficiency-level sampling distribution under different variations of the EAP estimation method.

- Coded in SAS 6.08—and ulterior versions.

STDIF

Author: Frederic Robin

Source: University of Massachusetts Amherst, Center for Educational Assessment.

Capabilities:

A DOS-based program to compute DIF indices of conditional p-value differences between two groups of interest (the reference group and the focal group) with two different indices of SDIF (signed DIF) and UDIF (unsigned DIF).

TESTGRAF

Author: J.O. Ramsay.

Source: <ftp://ego.psych.mcgill.ca/pub/ramsay/testgraf/>

Capabilities:

A software for MS-DOS machines provides a graphical analysis of multiple-choice test items and/or rated responses using Ramsay's "kernel smoothing" approach to item response theory.

WINIRT

Authors: Hua Fang & George A. Johanson.

Source: <http://oak.cats.ohiou.edu/~hf101702/winirt.htm>

Capabilities:

A Windows-Based Item Response Theory Data Generator with an Equating and Differential Item Functioning Simulation Guide.

WinGen2

Author: Kyung T. Han

Source: *Applied Psychological Measurement*, 2007, 31 (5), 457- 459.

<http://www.umass.edu/remf/software/wingen/>

Capabilities:

Windows software generates IRT parameters and item responses.

Applicable Models:

(1) dichotomous IRT models with one, two, and three parameters, (2) polytomous IRT models including the partial credit model, generalized partial credit model, graded response model, rating scale model, and nominal response model, (3) non-parametric models, and (4) multidimensional models.

Features:

- Allows a mixture of more than one IRT models in a set of items.
- Generates parameter values from various distributions for realistic data including normal, uniform, beta, or lognormal distribution.
- Provides an intuitive and user-friendly interface and plots item characteristic curves (ICC), test characteristic curves (TCC), item information curves (IFC), and test information curves (TIC).
- Syntax files can be used. Thousands of syntax files can be run in a cue.

- Replication data can be simulated up to 1,000,000 sets, and syntax files are generated for other IRT programs.
- Different examinee data can be generated for each replication.
- Analyzes DIF by computing RMSD, MAD, and BIAS.
- Developed on *Microsoft .NET frameworks 2.0*. Runs on either 32bit Windows series (Windows XP) or 64bit Windows series (Windows vista).

WPERFIT

Authors: Pere J. Ferrando & Urbano Lorenzo

Source: *Educational and Psychological Measurement*, 2000, 60 (3), 479-487

Capabilities:

A program computes different person-fit measures under different parametric item response models for binary items.

Applicable Models:

Rasch model and the two- and three-parameter logistic models.

Features:

- The person-fit indices include the standardized log-likelihood index lz , the standardized extended caution index $ECI4z$, and the chi-square statistic of Trabin and Weiss.
- Plots observed and expected person response curves (PRC).
- No clear limit on the maximum number of subjects and items that can be analyzed.
- Developed in Visual C++ and can be executed under the Microsoft Windows 95/NT operating system.

July, 2009