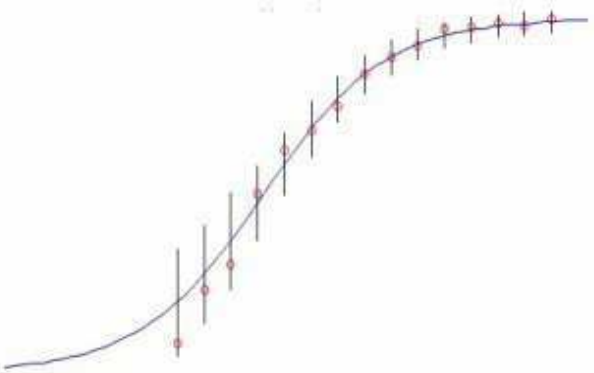
ResidPlots-2
User's guide



Tie Liang
Kyung T. Han
Ronald K. Hambleton

University of Massachusetts Amherst

Partial support for development of this software at the University of Massachusetts Amherst came from the College Board and we are grateful to them for their assistance.

User's Guide for ***ResidPlots-2***:

Computer Software for IRT Graphical Residual Analyses (Version 2.1)

Tie Liang, Kyung T. Han, Ronald K. Hambleton[1]

University of Massachusetts Amherst

July 1, 2009

[1] The authors have been working for more than one year on the software and felt that it was time to release version 2.1. We take no responsibility for any errors that may remain in the software. The software is freely distributed and users proceed at their own risk. We plan to keep the software available at http://www.umass.edu/remp/software/residplots/ and as errors are identified, or changes or extensions are made, we will update the software. The second author is now at the Graduate Management Admission Council.

The manual can be cited as follows: Liang, T., Han, K.T., & Hambleton, R. K. (2008). User's guide for ResidPlots-2: *Computer software for IRT graphical residual analyses, Version 2.0* (Center for Educational Assessment Research Report No. 688). Amherst, MA: University of Massachusetts, Center for Educational Assessment.

Acknowledgments

The three authors are grateful to the College Board for partially funding the development of *ResidPlots-2* through contracts in 2007 and 2008 to the University of Massachusetts (Ronald Hambleton and Stephen Sireci, PIs) to support the College Board Advanced Placement Psychometric Research, to Yue Zhao (see Zhao, 2008) for her initial studies on graphical residual analyses that provided the basis for some of the current methods that were implemented in the software, and to Peter Baldwin, who was helpful, in some of the early thinking about what the IRT model fit software might include. Though we are grateful for the assistance received, we recognize that any misjudgments or errors that remain in the manual or in the software are the responsibility of the authors of the manual.

TABLE OF CONTENTS

PART I

Introduction

Because item response theory (IRT) provides many advantages over classical test theory, the use of IRT models in constructing tests, equating scores, assessing item bias, and/or estimating proficiency scores, continues to grow (van der Linden & Hambleton,1997). However, even when the technical demands of IRT are satisfied (e.g., model parameter estimation), the success of these IRT applications depends on satisfactory fit between the model and the observed data. The assessment of model fit remains a major hurdle to overcome for effective measurement (Hambleton & Han, 2005; Swaminathan, Hambleton, & Rogers, 2007).

A general strategy for evaluating model fit at the item level and the strategy adopted in preparing our software involves comparing observations with model-predicted expectations (Hambleton and Han, 2005; Hambleton, Swaminathan, & Rogers, 1991; Rogers and Hattie, 1987). Statistical significance tests can be used for this purpose; however, they are not without shortcomings. Such tests tend to be narrowly focused on a particular aspect of the relationship between the model and the data, often summarizing the evaluation into a single descriptive number or test result. Furthermore, the most widely used fit statistics are susceptible to Type I error rates (e.g., Yen's Q_1 (1981), McKinley and Mills' G^2 (1985), Bock's χ^2 (1972)). It's known that the inflated Type I error rates are primarily due to grouping examinees into intervals based on the $\ddot{\theta}$ values, which contain error (Orlando & Thissen, 2000). Furthermore, the degrees of freedom for the χ^2 distribution are unknown for some of the test statistics (Orlando & Thissen, 2000). Finally, all models are wrong, and given sufficient amounts of data, all statistical tests will detect model misfit. There is certainly more to establishing IRT model fit than conducting a simple hypothesis test.

We believe professional judgment and the study of practical consequences of model misfit, along with statistical analyses, should provide the basis for model selection decisions. In our software, practitioners have an opportunity to carefully study the patterns of model misfit for a number of models fit to their data, and even between data that actually fits their model (generated from their item and person parameter estimates) and the predictions from the model to provide a framework for interpreting residuals.

In contrast then to statistical analyses, our approach involves visually representing the discrepancy between the model-based expectations and the observed data. Such graphical displays, while less objective than a statistical test in some cases, have proven to be useful in evaluating the complex relationship between model and data (Hambleton, Swaminathan, & Rogers, 1991). Some commercial IRT estimation software includes raw residual plots in the standard output—e.g., BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003)—others, such as PARSCALE (Muraki, & Bock, 2003) and MULTILOG (Thissen, 2003) do not. Even among those software packages that provide such output, room for improvement exists. For example, users often want a bit more flexibility in setting up their residual plots than current software provides (such as the number of categories into which examinees are grouped).

The purpose of the present software, *ResidPlots-2,* is to provide a powerful tool for graphical residual analyses. The advantages of this software include several features.

(1) *ResidPlots-2* supports the most widely used IRT models including three dichotomous models (1PLM, 2PLM, 3PLM) and three polytomous models (GRM, GPCM, NRM). (And, more models will be added in the coming year or two.) There are no practical limits on the number of items or the examinee sample size.

(2) *ResidPlots-2* provides considerable flexibility with respect to the number and size of the intervals for which the residuals are computed. Different data have different features and therefore it is beneficial to give users choices about the intervals. For example, users are able to decide number of intervals, choose the interval size to provide equal width or equal frequency ("equal widths" conditions speeds up the analysis considerably), select the location of the data plot in each interval, eliminate intervals at the lower or higher end of the proficiency scale if desired, etc.

(3) Typically, it is helpful to provide error bars on raw residual plots. *ResidPlots-2* allows users to decide what type of error bars they wish to have displayed. Users can specify the number of standard errors represented by the error bars (e.g., 2 standard errors).

(4) *ResidPlots-2* provides three sets of plots, first, at the item level, raw residual plots and standardized residual plots with error bars, second, at the test level, *ResidPlots-2* can show standardized residual distributions (PDF and CDF) with corresponding tables, item fit plots and score fit plots from both empirical and simulated data, and finally, observed test score distributions and predictive score distributions (PDF and CDF) are produced too. The predictive test score distribution is based on simulation data generated from item and ability parameter estimates from the observed data.

(5) The user-friendly interface (see Figure 1) is convenient and straightforward. Users merely point to the syntax file used to run PARSCALE, BILOG-MG or MULTILOG and *ResidPlots-2* will provide the analysis. Additionally, users can access any plot by pointing and clicking the options (see Figure 2, in the bottom portion of the display).

ResidPlots-2 consists of two components: A component for computing residual statistics, and another component for communicating with users and for plotting the residual graphs. The first and second components were written in FORTRAN and Microsoft .NET Framework 3.5, respectively. Since the component for residual statistic computation is DOS compatible as well as MS Windows, users can utilize the program in DOS for a batch run. In addition (for international users), users should make sure the language setting in their computers is English.

Educational Importance

Evaluating model fit is very important in item response theory because many of the inferences that are made from the estimates assume model fit. Residual plots provide a useful visual inspection of model fit. Their utility has led some researchers to suggest that residuals are

"perhaps the most valuable goodness-of-fit data" (Hambleton et al, 1991). Because of its user-friendly interface, measurement specialists are likely to quickly understand and master *ResidPlots-2*. We believe the software is a valuable advance in convenience and flexibility for researchers and practitioners wishing to do graphical residual analysis.

Figure 1. User-friendly interface.

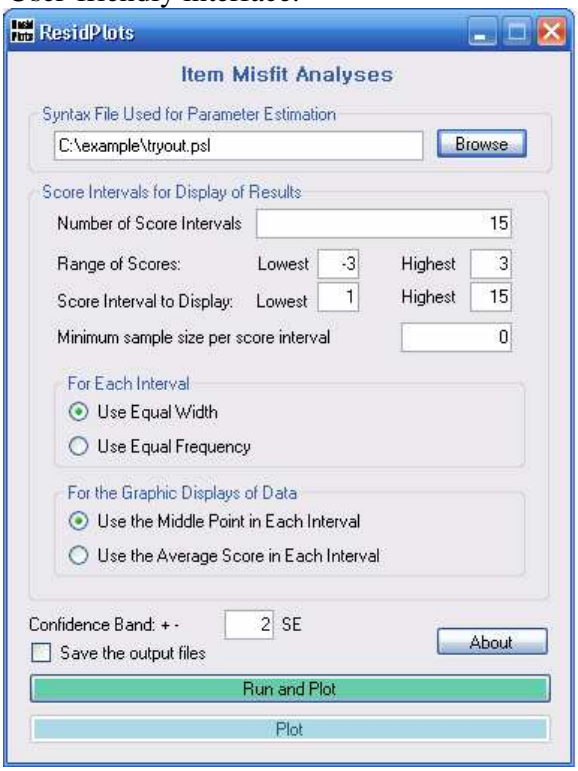
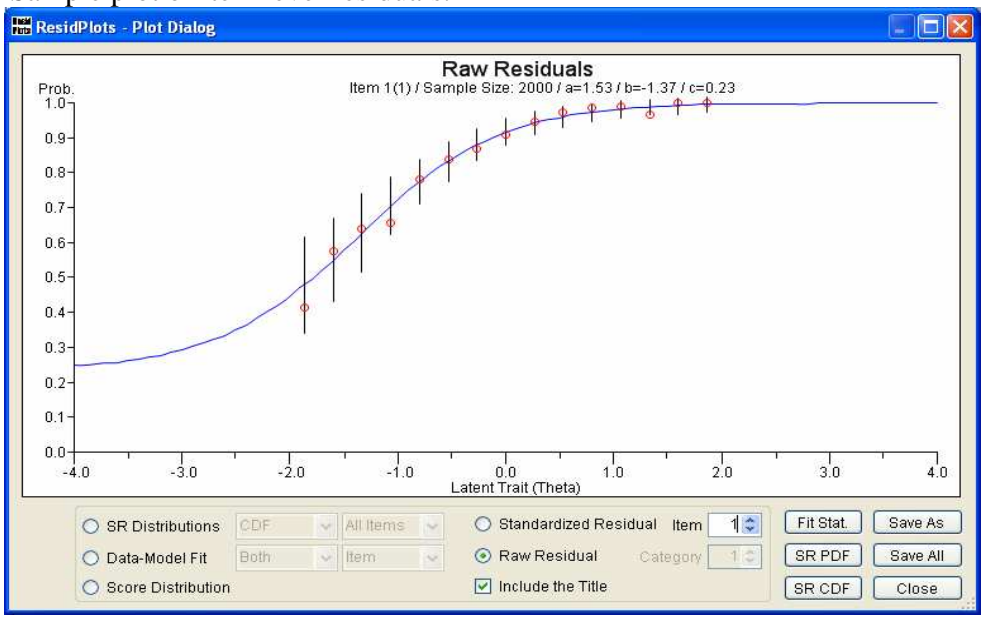


Figure2. Sample plot of item level residuals.

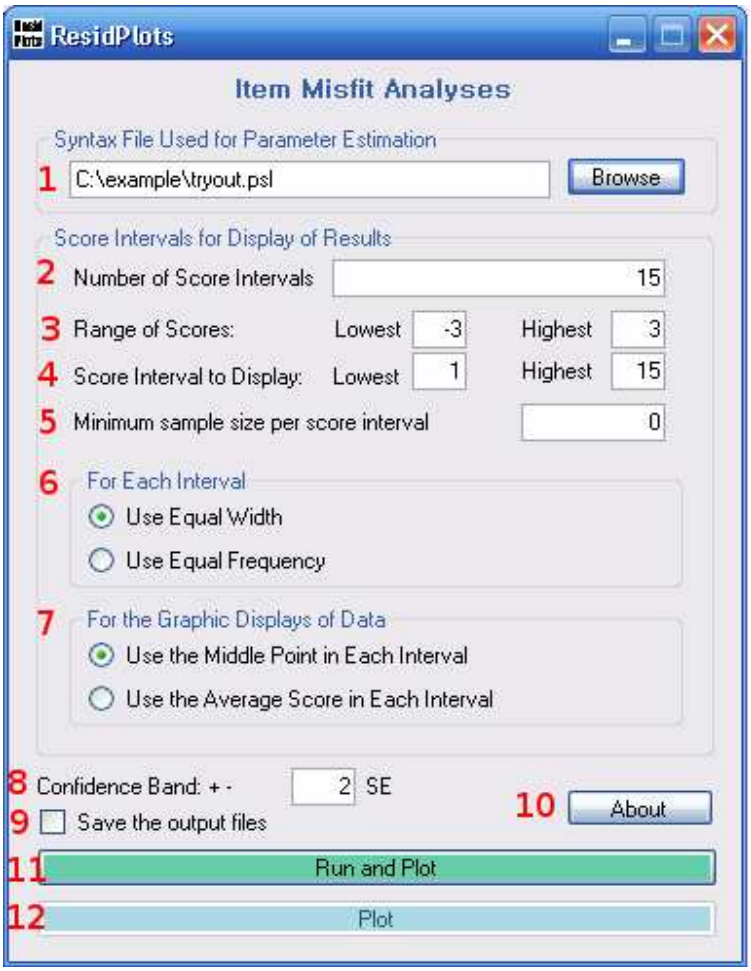


PART II

Using the Software

What follows is an explanation for the 11 variables that need to be specified to run the program.

Interface



1. The required information to run *ResidPlots-2* is provided below:

(a) Files

ResidPlots-2 requires users to enter the name and/or path of the syntax file which was already run in BILOG-MG, PARSCALE or MULTILOG successfully. For BILOG-MG and PARSCALE, there are basically five files needed to get *ResidPlots-2* to run: syntax file, data file together with key files for not reached and/or omitted data (if they were used in the original run), .PH2 file, .SCO file and .PAR file. For MULTILOG, four files are needed: syntax file used

for second run, data file, .SCO file and .PAR file. Users are required to put these files into one folder.

The name and/or path of the syntax file should NOT include any spaces.

(b) Commands and options in the syntax file

The following tables indicate the required information from the user's syntax file. All of the information is needed for the software to run.

BILOG-MG*

Commands	Options
>GLOBAL	DFNAME
>SAVE	PARM,SCORE
>INPUT	NTOTAL,NFNAME(if necessary),OFNAME(if necessary) ,KFNAME(if necessary)
	DATA FORMAT

* If users would like to specify the number of examinees in the syntax, please use "TAKE" not "SAMPLE".

PARSCALE*

Commands	Options
>FILE	DFNAME, NFNAME(if necessary),OFNAME(if necessary) ,KFNAME(if necessary)
>SAVE	FIT, PARM,SCORE
>INPUT	NTOTAL
	DATA FORMAT
>TEST	NBLOCK
>BLOCKS	NITEM, REPEAT(necessary for GRM or GPCM), NCAT,ORIGINAL,MODIFIED(if necessary), CSLOPE(necessary for 1PLM)
>CALIB	CSLOPE(necessary for 1PLM)

* Be sure that the option NCAT is in front of the ORIGINAL and/or MODIFIED in the BLOCK command.
* Users should be sure NOT to put more than one polytomous item in one BLOCK (except using the REPEAT command), because if they do, the .PAR file will NOT provide threshold parameters and *ResidPlots-2* will NOT run.
* If users would like to specify the number of examinees in the syntax, please use "TAKE" not "SAMPLE".

MULTILOG*

Commands	Options

>PROBLEM	DATA,NITEM,NG,NEXAMINEE
>TEST	MODEL, ALL(if necessary), ITEM (if necessary) ,NC (if necessary)
>START	PARAM
>SAVE	
	RESPONSE CODES AND DATA FORMAT

* The syntax file as input for ResidPlots-2 should be from second run in Multilog.
* When ITEM option is used, be sure to apply the reduced format (e.g., IT=(3(1)45))
* The a parameter of 3PLM shown on the plots and in the outputs in ResidPlots-2 absorbs 1.7.

Example 1 (BILOG-MG syntax)

```
Tryout test
A 40-item test fitted by 3PLM
>GLOBAL DFNAME='dich.dat',
        NPARM=3,
        LOGISTIC,
        SAVE;
>SAVE PARM='dich.PAR',
     SCO='dich.SCO';
>LENGTH NITEMS=(40);
>INPUT NTOTAL=40,
       TAKE=2000,
       NIDCHAR=10;
>ITEMS INAME=(i1(1)i40);
>TEST TNAME='tryout';
(10A1,40A1)
>CALIB NQPT=30,
       CYCLE=50,
       NEWTON=30,
       CRIT=0.0010,
       PLOT=1.0000,
       ACCEL=1.0000;
>SCORE ;
```

Example 2 (PARSCALE syntax)

```
A test with 40 multiple choice and 5 constructed response items
Fitted by the 3PM and the GRM
>FILE    DFNAME='mixgrm.dat',
         SAVE ;
>SAVE  SCORE='mixgrm.SCO',
        PARM='mixgrm.PAR'
              FIT='mixgrm.fit';
>INPUT  NID=4,NTOTAL=45,NTEST=1,TAKE=2000;
(4a1,6x,45A1)
>TEST1  TNAME=SOCSCI, ITEM=(1(1)45),NBLOCK=6;
```

```
>BLOCK1 NCAT=2,
        ORIGINAL=(0,1),
        MODIFIED=(1,2),
        GUESSING=(2,estimate),
        NITEMS=40;
>BLOCK2 NCAT=4,
        ORIGINAL=(0,1,2,3),
                MODIFIED=(1,2,3,4),
        NIT=1,
        REP=5;
>CALIB  GRA, LOG,
                SCALE=1.0,
                NQPTS=15, CYCLE=(50,1,1,1,1), NEWTON=2,
        CRIT=0.01;
>SCORE  ;
```

Example 3 (MULTILOG syntax)

A test with 40 multiple choice and 5 constructed response items
Fitted by the 3PM and the GRM

```
>PRO SC,
        IN,
        DA = 'D:\mlg\mixgrm.wgr',
        NI=45,
        NG = 1,
        NE=2000,
        NC=10;
>TEST ITEMS=(1(1)40),L3;
>TEST ITEMS=(41(1)45),nc=(4(0)5),GRA;
>START ALL, PARAM='mixgrm.par';
>SAVE;
>END ;
4
0123
11111111111111111111111111111111111111111111
22222222222222222222222222222222222222222222
00000000000000000000000000000000000033333
00000000000000000000000000000000000044444
(10a1,45a1)
```

2. Number of score intervals

This choice is for users to specify the number of score intervals. For example, 15 is often a good choice to establish the fit of an IRT model to the test data.

3. Range of scores

Users can enter any range of scores of interest. For example, the lowest ability score may be -3.0 and the highest score may be 3.0. Scores outside this specified range are not used in the calculations. A span of six standard deviations of scores would seem to be enough spread along the proficiency continuum for assessing the extent of model fit. Any wider range, and problems will surely arise in most applications with small frequency counts.

4. Score interval to display

Users have two numbers to enter—the lowest interval number, and the highest. For example, if there were 20 score intervals formed, the user could type in 1 and 20, or perhaps 4 to 17. This choice of range of intervals of interest will only apply to raw residual plots and standardized residual plots. Normally though the numbers from 1 to the number of intervals (N) will be used unless the user desires to see a smaller number of intervals displayed in the plots.

5. Minimum sample size per score interval

A large sample size in each score interval is always desirable for investigating model fit. But often this will not be possible. Users can use this option to suppress data points based on sample sizes less than some desired number. The default number is zero.

6. Select type of score interval

The choice of width of each interval gives users flexibility on how to look at residuals. Usually, if users are interested in the consistent display of residuals, they may need to choose equal width. This is the choice we often make. But the problem of "equal width" is that a small sample in an interval can result and so the plots can become unstable, and any statistical analyses can become problematic.

(a) Equal width

For this choice, *ResidPlots-2* divides the range of scores of interest by the number of intervals specified by the user.

(b) Equal frequency

For this choice, the number of examinees in each interval is decided by the total sample size divided by the number of intervals specified. Normally this choice will substantially slow the software and we recommend you avoid this option. We hope to improve this option in a later version of the software.

7. *Location of data points*

This option provides users flexibility on where to plot the data point within each interval. There are three choices: middle of the interval, or mean or median of scores in the interval. In practice, it is difficult to evaluate this difference in the placement of the data points, especially, if

there are lots of intervals being used to display the misfit information. **Note**: If the number of examinees within a score interval administered a particular item is zero (as might happen when a full item bank is being calibrated and the number of examinees taking some of the items might be small), no data point will show in the item level residual and standardized residual plots. Also, a table is available that displays the number of examinees in each interval administered each test item. This table will be useful when sample sizes administered test items vary considerably. Finally, SRs associated with expected probabilities below .05 or above .95 (of the maximum item score) are deleted from the SR distributions so as not to compromise the meaning of the actual and simulated distributions of SRs.

8. Size of error band

Users can set up the error band by entering the number of SE (e.g., 1, 2, or 3 with default=2).

9. Output file

If this option is unchecked, six tables will be saved in the same folder with the syntax file specified by users: "Name" is the same name as the syntax file name. They are "name"_RP_report.out, "name"_RP_sr_pdf.out, "name"_RP_sr_cdf.out, "name"_RP_fit.out, "name"_RP_ncount.out, "name"_RP_pfit.out. These tables are very useful for interpreting the misfit information. They are described in the "tables" section of PART II.

If this option is checked, in addition to the six tables, eight output files will also be saved in the folder. These eight outputs are the basis of the plots from *ResidPlots-2.* They are described below in detail:

Name_RP_id.out

It includes all item information. For example, item id, sample size, item parameters, chi square fit statistics (value, DF, p), G square fit statistic (value, DF, p). The first row in this file is reserved for reporting the scaling factor (1.7 or 1.0), the second row is to tell if it's a mixed test format or single test format (1-mixed, 2-dichotomous items only, 3-polytomous items only).

Name_RP_cb.out

It has the data for plotting raw residuals.
Row: first row is proficiency (latent trait) values. Starting from the second row, observed point, lower bound and upper bound for each item are reported.

Column: first three columns are for coding items. First column is used to distinguish x-axis (0) and actual values (1), second column is item number, third column is item category (2-two categories, otherwise, the category code will be following the order 0,1,2…). Starting from the fourth column, data information for each interval is presented.

Name_RP_sr.out

It has the data for plotting standardized residuals. The coding is the same as in name_RP_cb.out.

Name_RP_scodist.out

It has the data for plotting score distribution.
Row: First row is score points. Starting from the second row, PDF and CDF from both empirical data are reported.

Column: First column is data source (1-empirical data, 2-simulated data), second column is distribution type (1-PDF, 2-CDF). The first row is score points, so the two codes are always 0,0. Starting from the third column, they are score distribution frequency values for plotting.

Name_RP_srdist_pdf.out

It has the data for plotting the standardized residual frequency distribution.
Row: First row is proficiency (latent trait) values. Starting from second row, PDF and CDF from both empirical data and simulated data are reported.

Column: The first three columns are for coding. First column is data source (1-empirical data, 2-simulated data), second column is distribution type (1-PDF, 2-CDF), third column is test format (1-mixed, 2-dichotomous only, 3-polytomous only). The first row is proficiency (latent trait) values, so the three codes are always 0,0,0. Starting from the fourth column, there are 80 columns (relative frequency numbers) for plotting.

Name_RP_srdist_cdf.out

It has the data for plotting the standardized residual cumulative distribution based on PDF. The coding is the same as in name_RP_srdist_pdf.out.

Name_RP_chis.out

It has the data for plotting score fit plot for the whole test. The first column is score interval number, the second column is chi square number for each interval from the empirical data, the third column is the chi square number for each interval from simulated data. The chi square values are normalized based on the number of score categories of each item and the number of items in the test.

Name_RP_chii.out

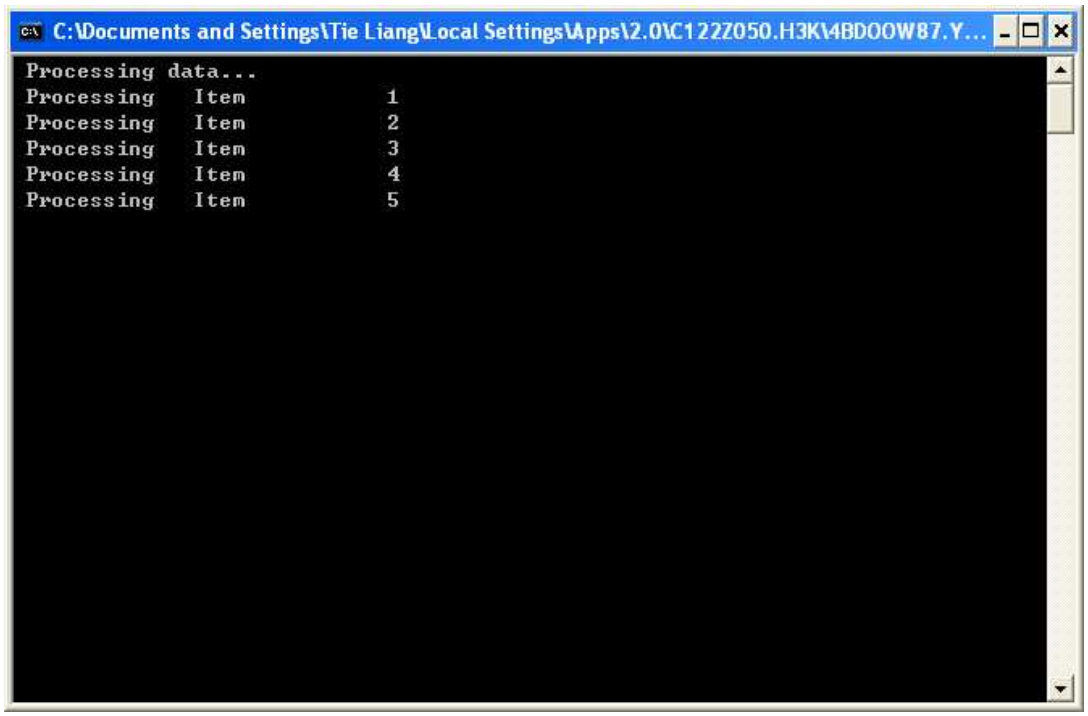
It has the data for plotting item fit plot for the whole test. The first column is item number, the second column is the chi square number for each item from the empirical data, the third column is the chi square number for each item from the simulated data.

10. About ResidPlots-2

Author and version information

11. Run and plot

Users can get all the plots by clicking this button. If the software is running correctly, users should be able to see the screen below. Sometimes, it will take some time to process data if there is data collapsing in the syntax file.



12. Plot

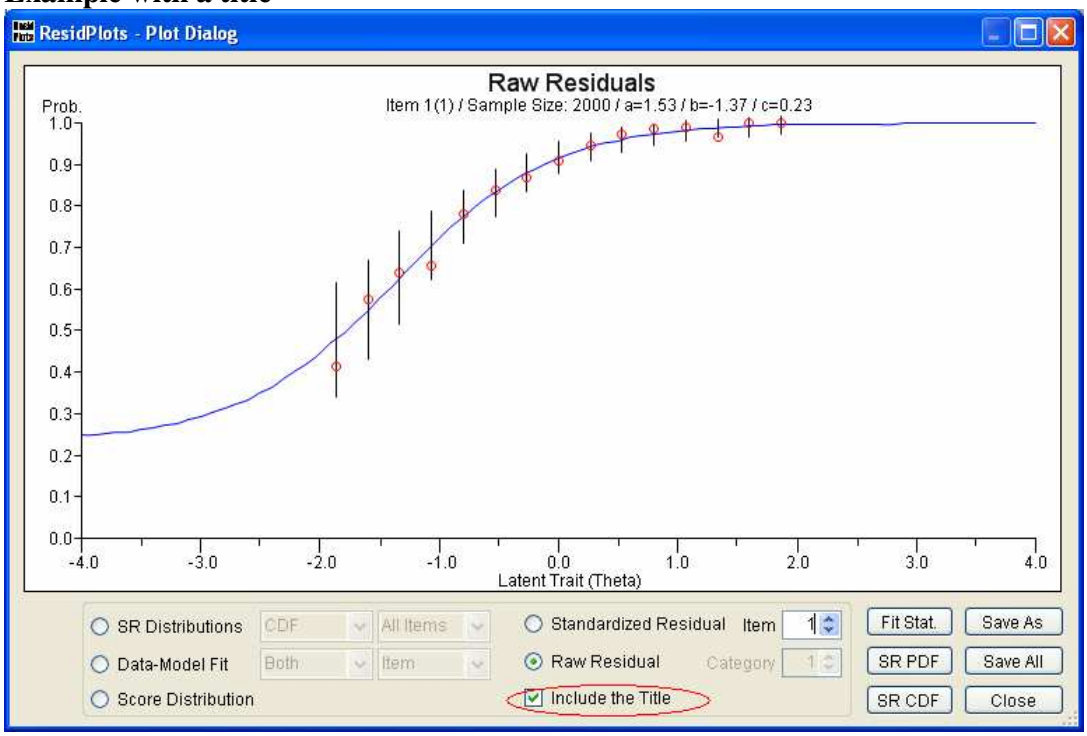
Get the plots for a second time without running ResidPlots-2 again.

Plots

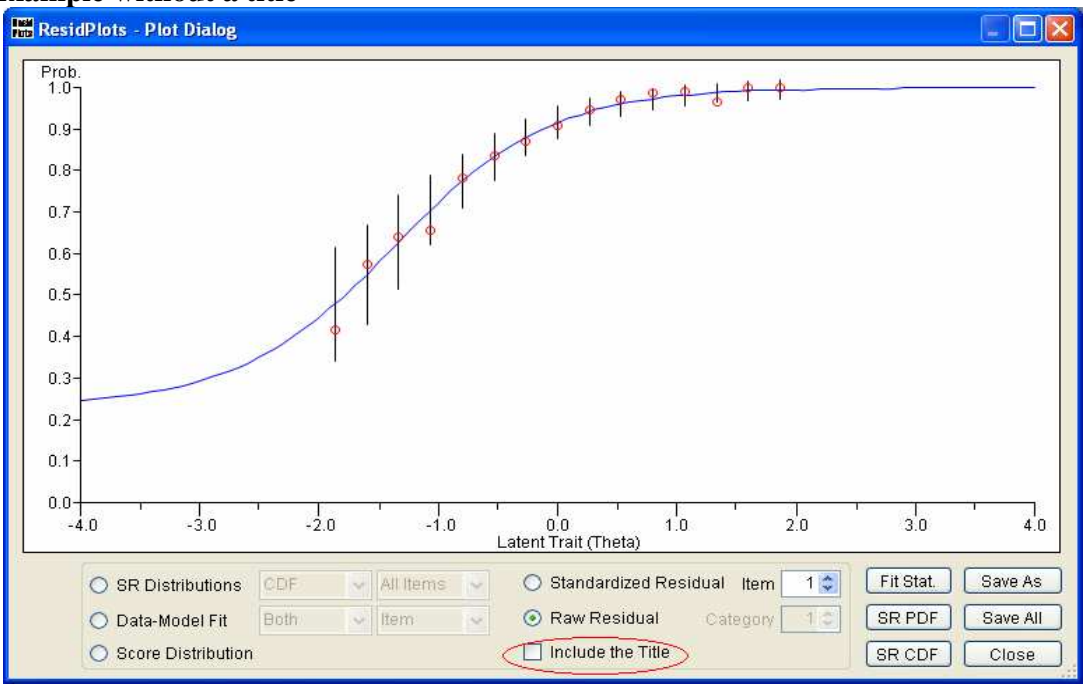
1. Plot title

Users can decide whether to include the graph title or not.

Example with a title



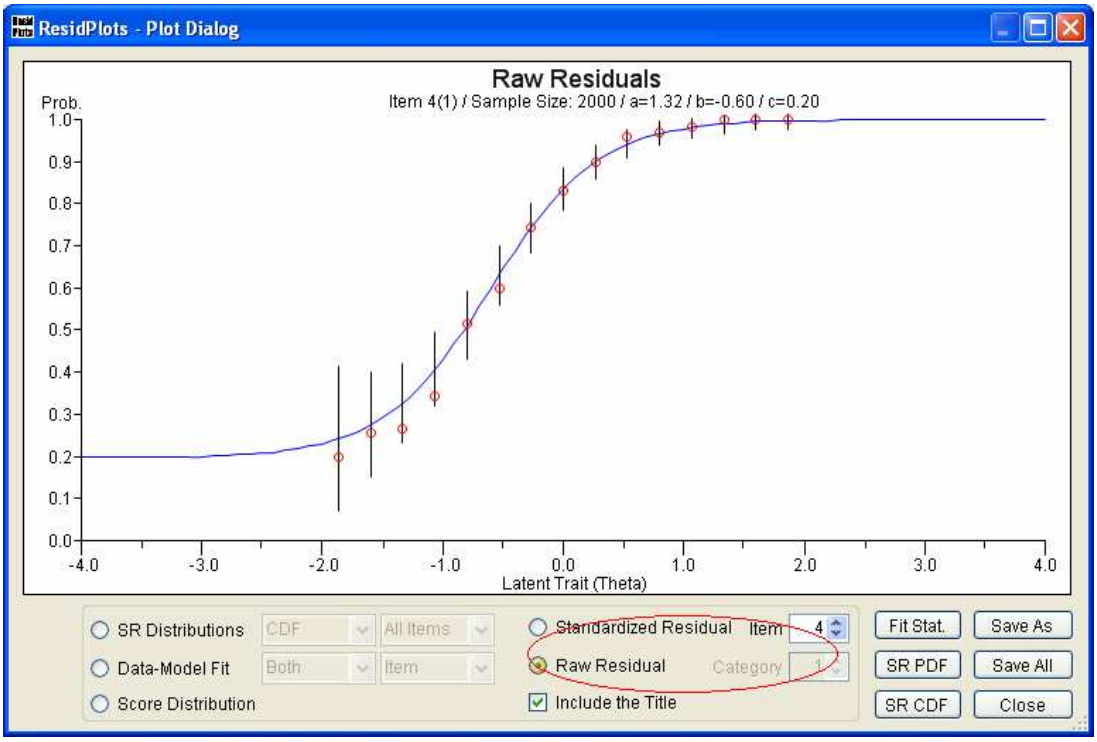
Example without a title



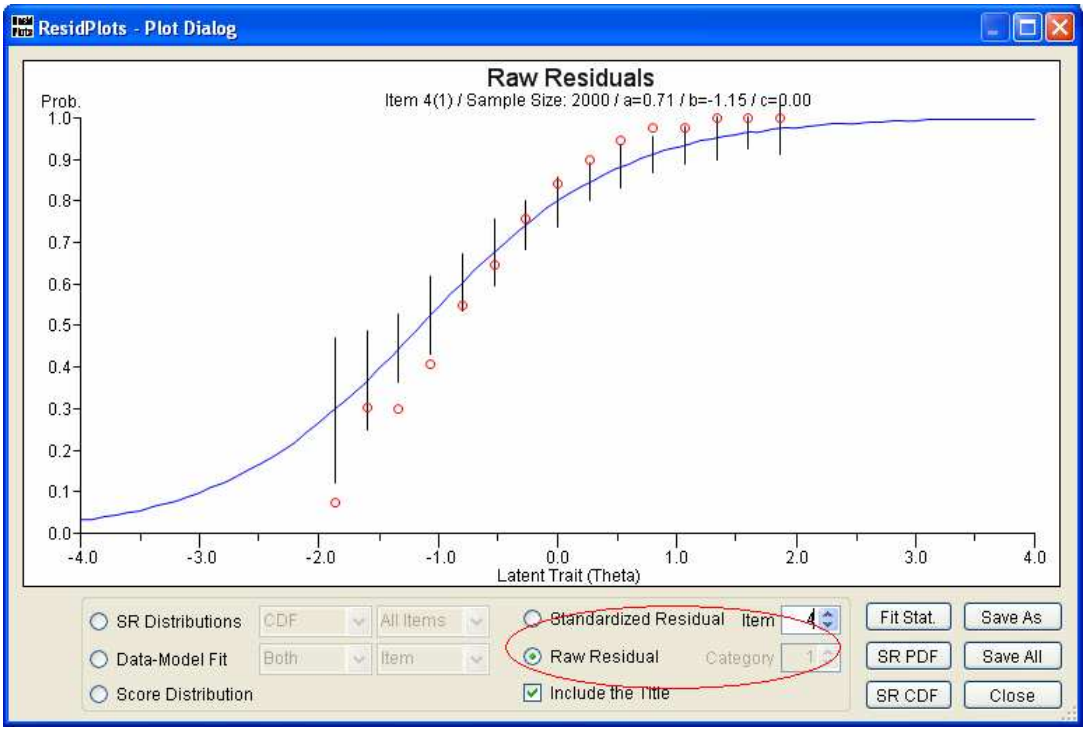
2. *Item level fit* plots

Item level fit plots include raw residual (RR) plots and standardized residual (SR) plots. *ResidPlots-2* shows both for each item and also for each score category for polytomous items. If a score category has zero examinees in it, this category is not displayed, and the category is not included in any of the statistical calculations.

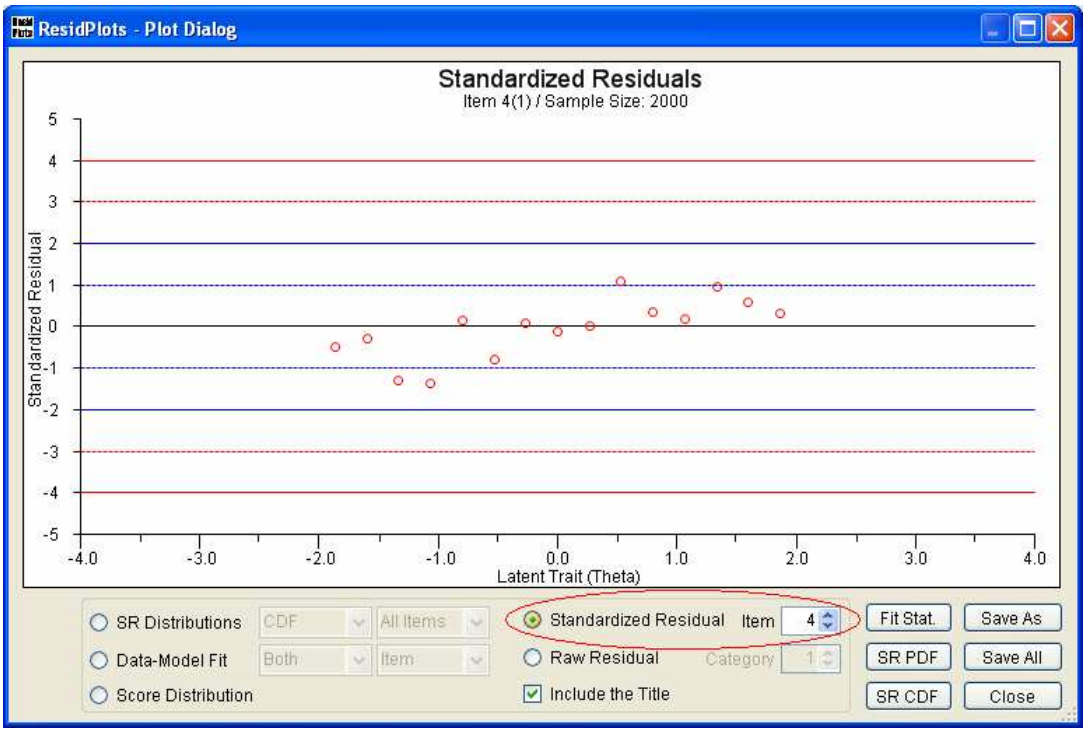
Example of raw residual plot (the item was fit by the 3PLM)



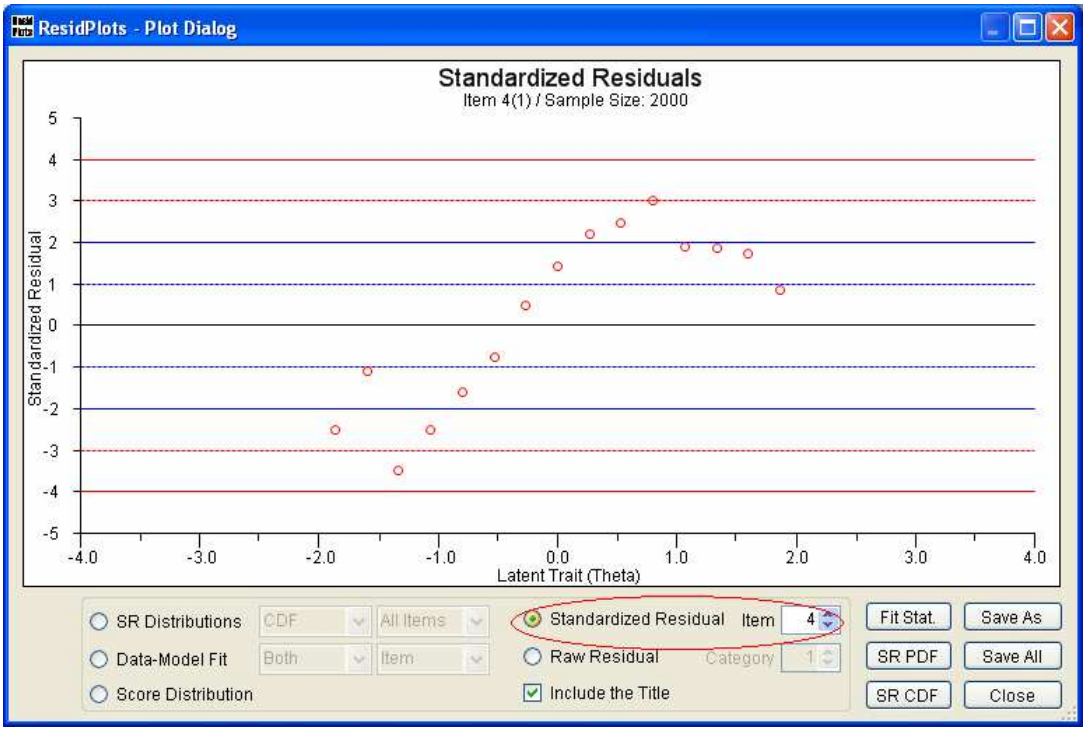
Example of a raw residual plot (same item as above was fit by the 1PLM)



Example of a standardized residual plot (the item was fit by the 3PLM)



Example of standardized residual plot (the same item was fit by the 1PLM)



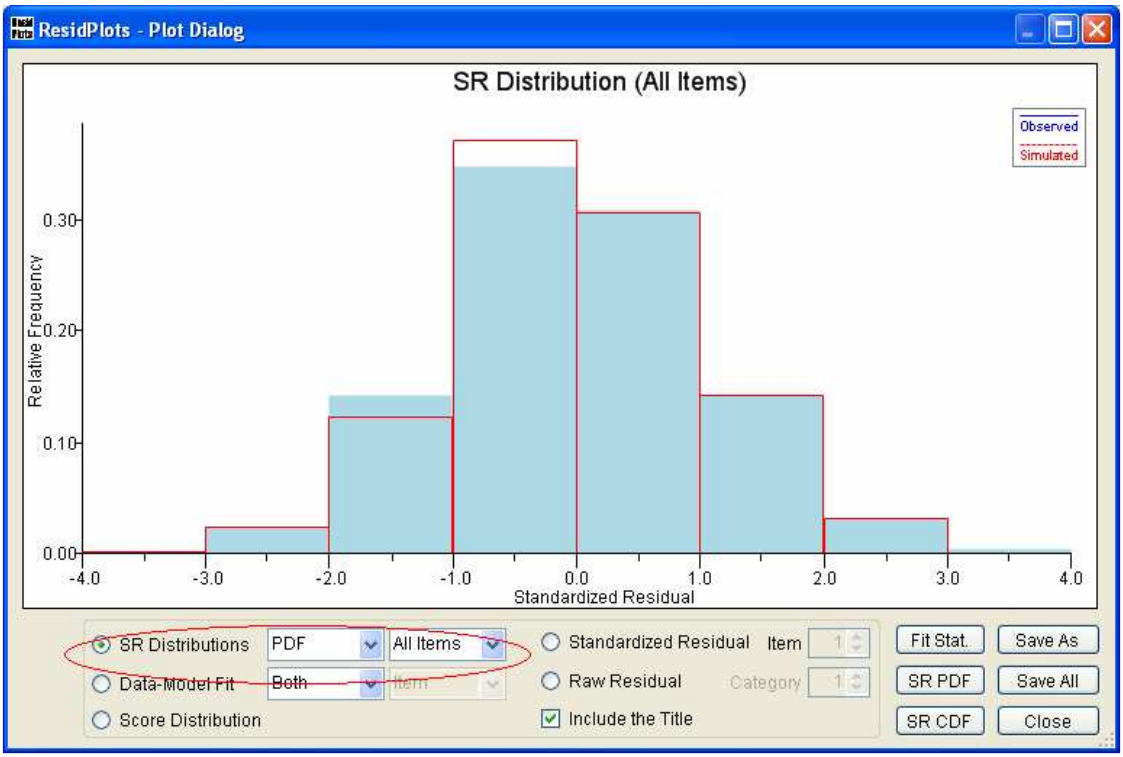
3. Fit plots at the test level

SR distributions

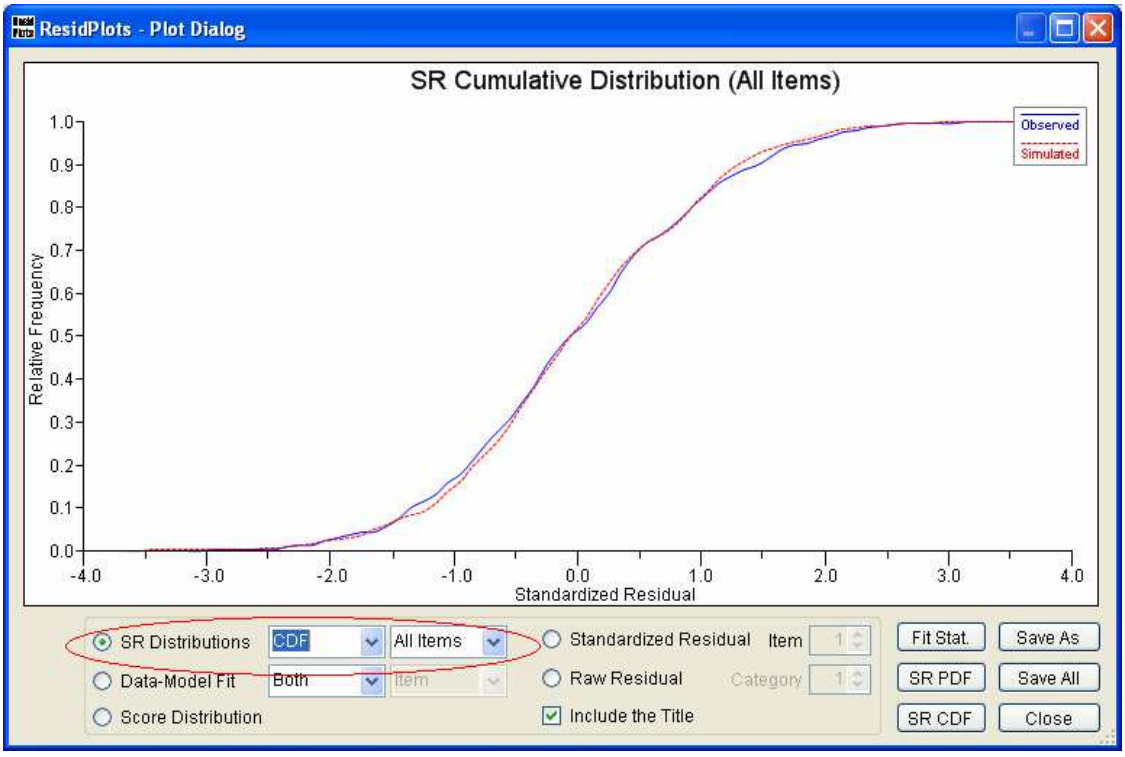
ResidPlots-2 provides standardized residual distributions (both PDF and CDF) from the observed data and/or the simulated data. What's more, if the test format is mixed, users can choose to see distributions on all items, or only dichotomous items or only polytomous items. This can be a useful feature in attempting to understand whether format of the items is differentially influencing the residuals.

If it appears from the graphical displays of the ICCs or item category score probabilities, that many expected probabilities are less than .05 or greater than .95 of the maximum score points on an item, the associated SRs are deleted to avoid a tendency of too peaked SR distributions, as might be the case for low performing candidates on difficult test items or difficult to achieve score points, it may be best to restrict the range over which residuals are studied. A better choice than looking at residuals from -3.0 to +3.0, might be to look at them from -2.0 to +2.0. We are currently working on a solution to deleting SRs associated with very low expected probabilities for success on an item or a score point. In the current version, and when the problem arises, there is a tendency to get an abundance of SRs very close to zero, making the distribution too peaked in shape and thus providing misleading information about model fit. SRs associated with low expected probabilities, however, are included in the chi-square statistics used to assess item level fit.

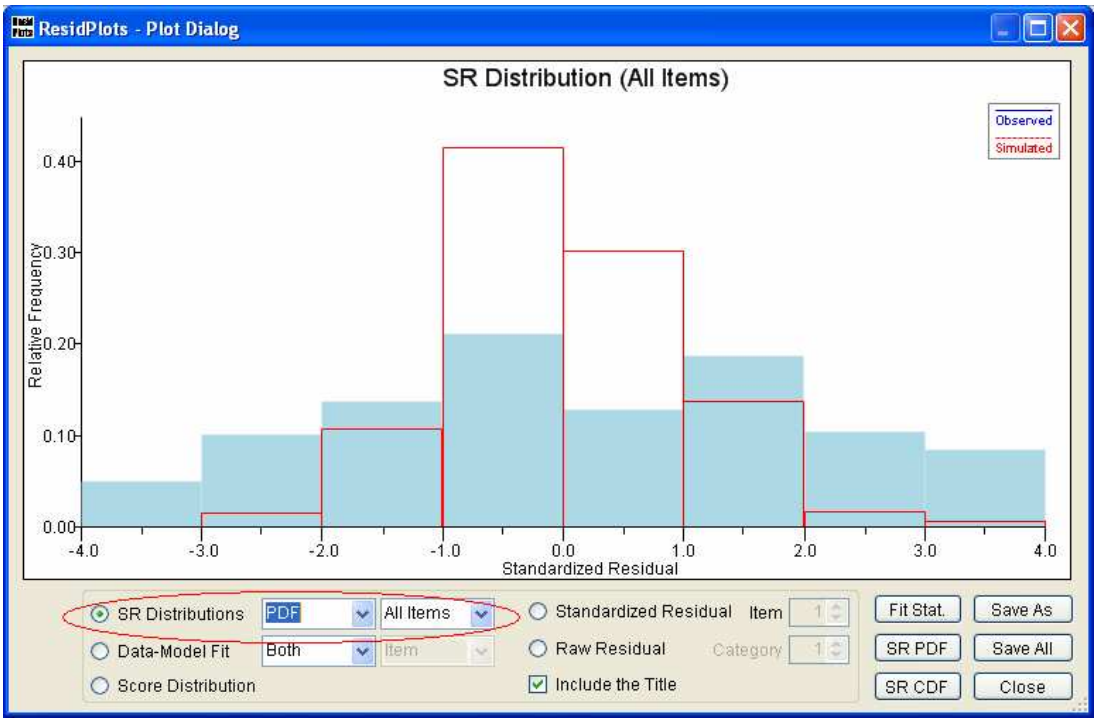
Example of SR PDF (data were fit by the 3P/GRM)



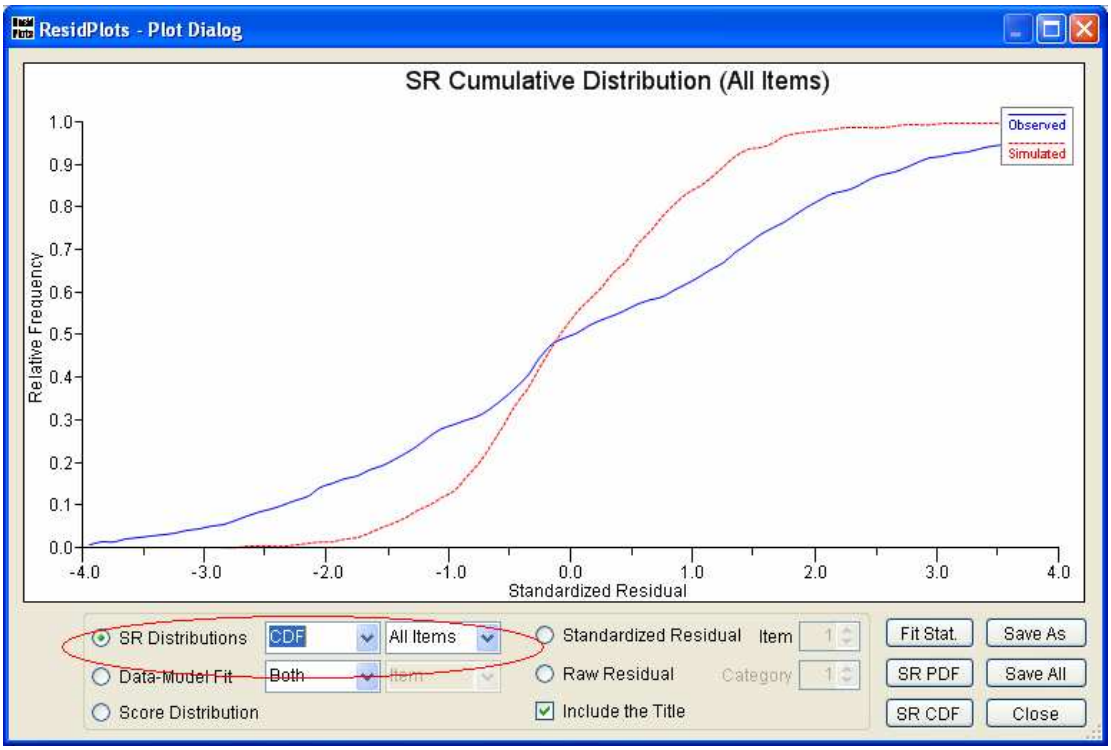
Example of SR CDF (data were fit by the 3P/GRM)



Example of SR PDF (same data were fit by the 1P/PCM)



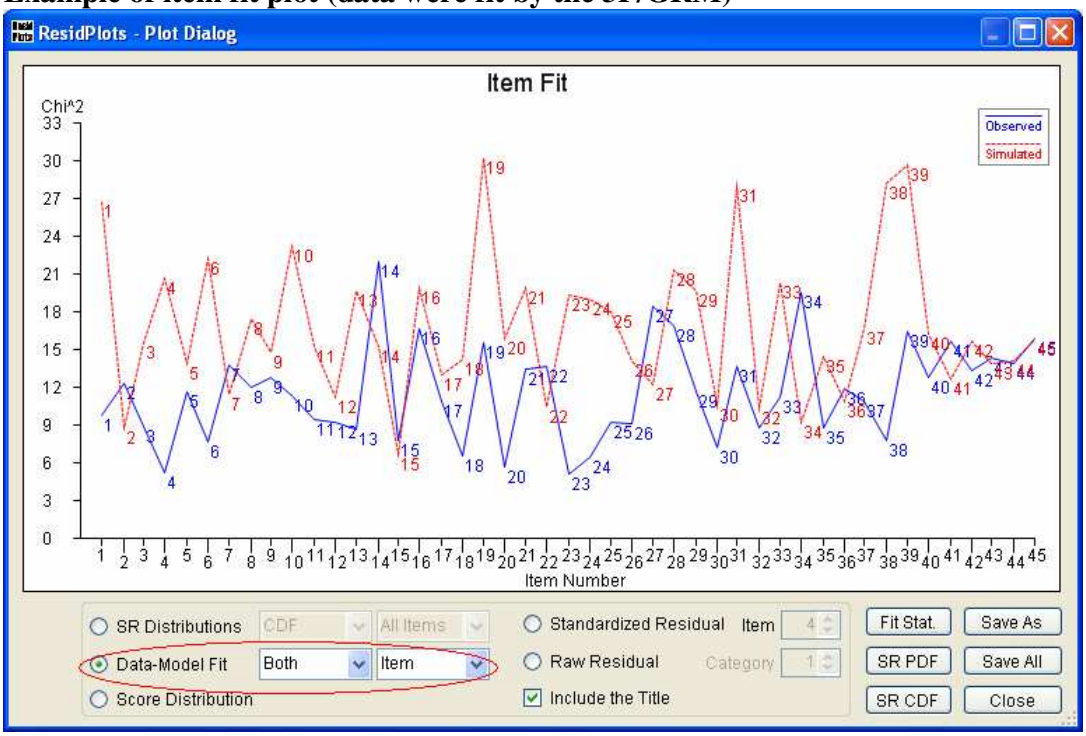
Example of SR CDF (same data were fit by the 1P/PCM)



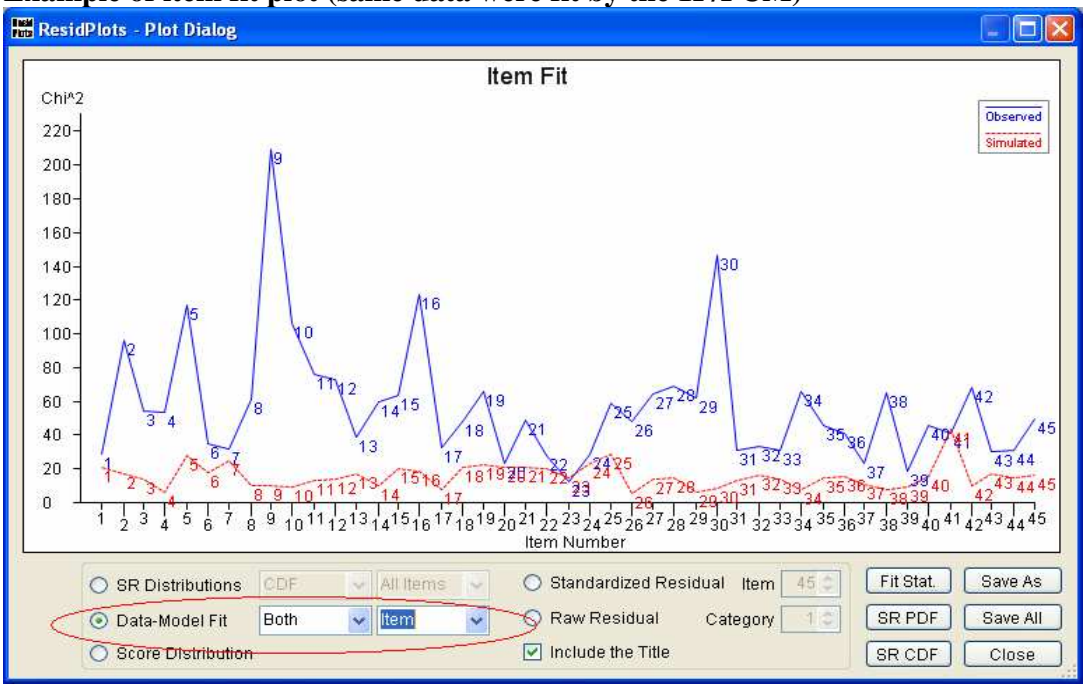
Data-model fit---item fit plot

The item fit plot gives users the chi square value of each item from the observed and/or simulated test plotted, in order, from the first to the last item in the test.

Example of item fit plot (data were fit by the 3P/GRM)



Example of item fit plot (same data were fit by the 1P/PCM)

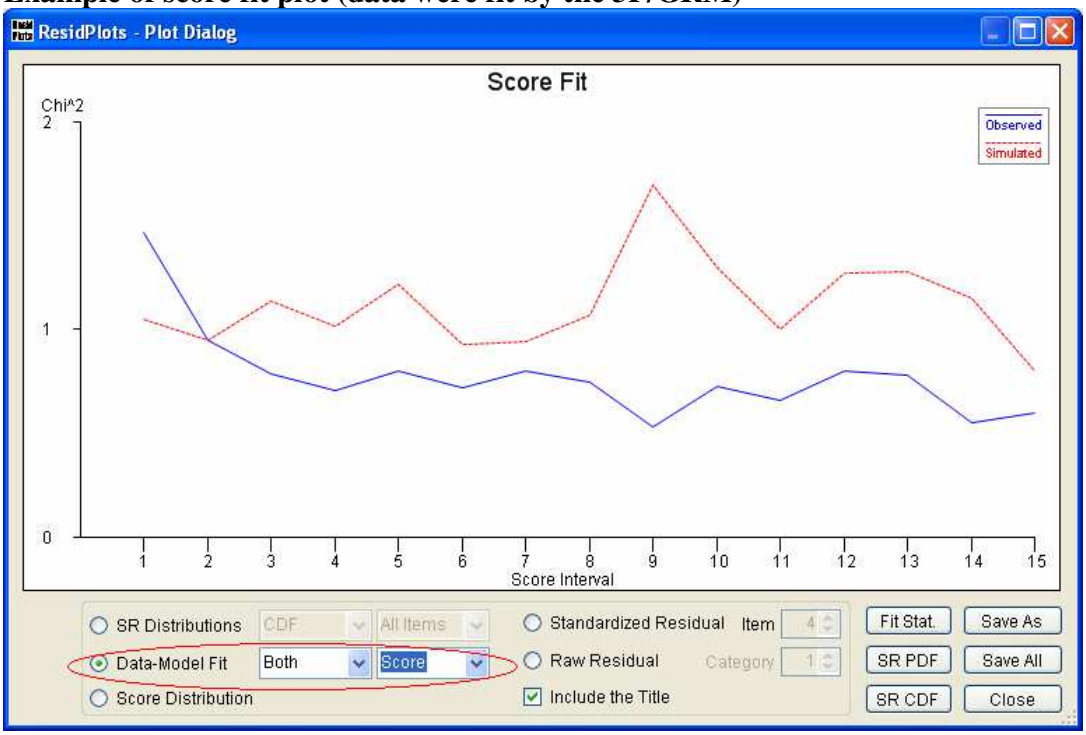


These plots can be used to identify the items which are fit least well by the model.

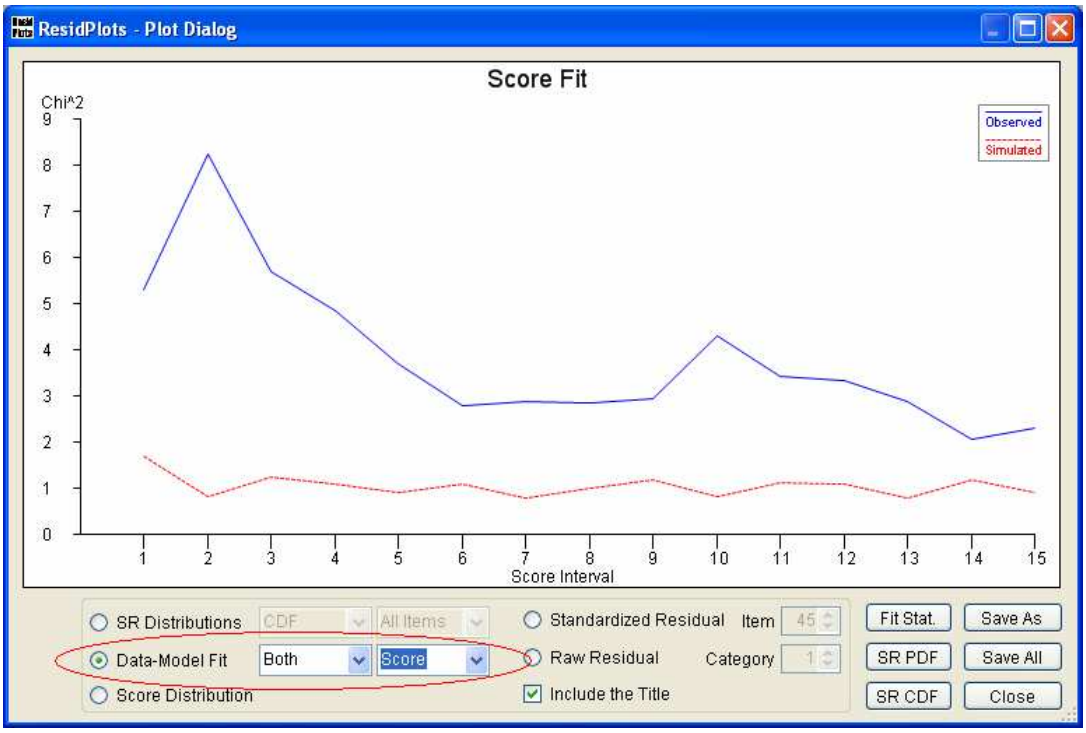
Data-model fit---score fit plot

The score fit plot provides users the chi square value of each score interval from the observed and/or simulated test (by summing over items).

Example of score fit plot (data were fit by the 3P/GRM)



Example of score fit plot (same data were fit by the 1P/PCM)



These last two displays allow the user to identify the level of misfit over the test items across the proficiency scale. The one above shows that the 1P/PCM models fit less well at the lower end of the proficiency scale. The pattern is not apparent when the 3P/GRM model was fit to the same data. See the display on the previous page.

4. Score distributions

ResidPlots-2 provides the observed and predicted score distributions. To derive the predicted score distribution, the assumption is made that the model parameter estimates for items and persons are true parameters, and then they are used to predict how the candidates would actually score on the test. To smooth out the predicted score distribution, 10 simulated distributions are generated and then averaged.

Example of frequency distribution (PDF) (data were fit by the 3PLM)

Example of cumulative distribution (CDF) (data were fit by the 3PLM)

Example of frequency distribution (PDF) (data were fit by the 1PLM)

Example of cumulative distribution (CDF) (data were fit by the 1PLM)

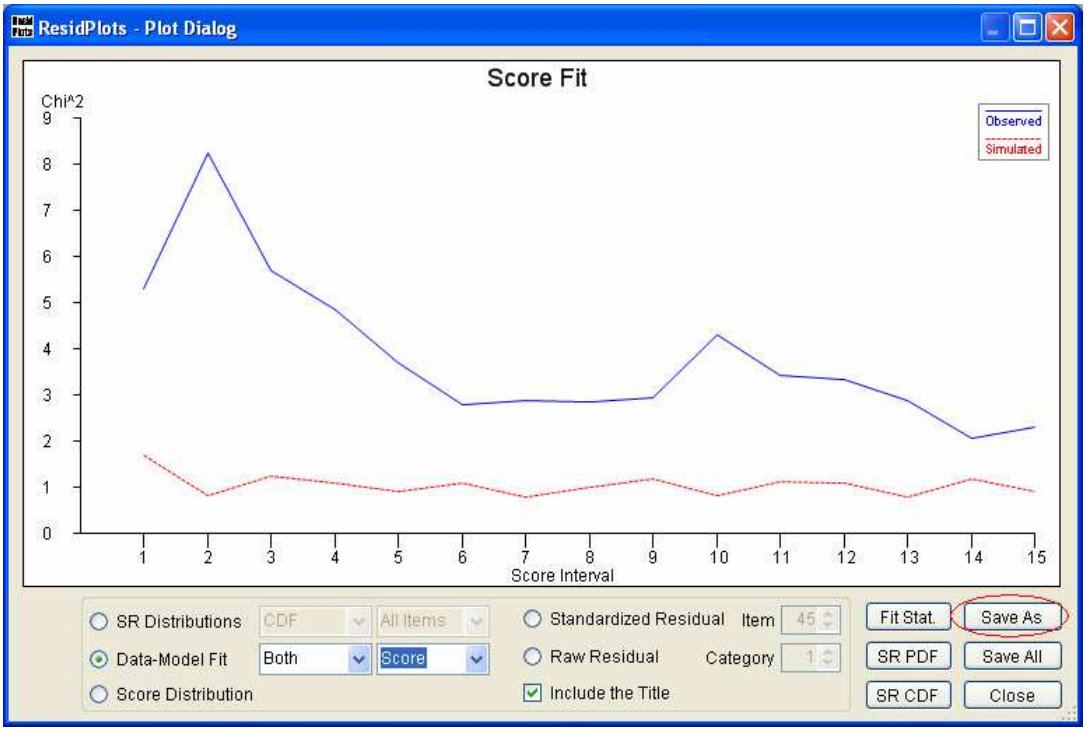
This comparison of observed and predicted score distributions is not of value with incomplete test designs such as MST and CAT since observed score distributions are based on test scores that are not comparable across examinees and the predicted score distributions are based on the full set of items available for testing, but any examinee saw only a fraction of these items. The software will do the analysis, but it is not valuable in addressing the question of model fit.

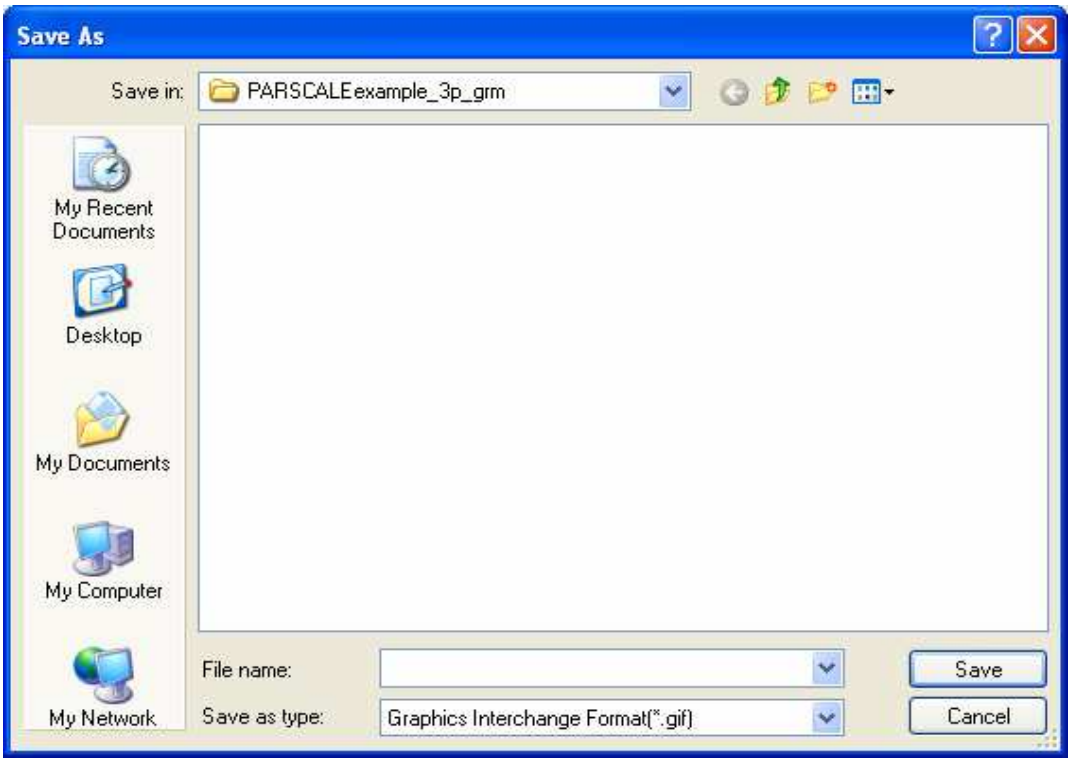
5. Plot saving

Users can save any single plot and/or save all plots to a folder specified.

Example of saving one single plot

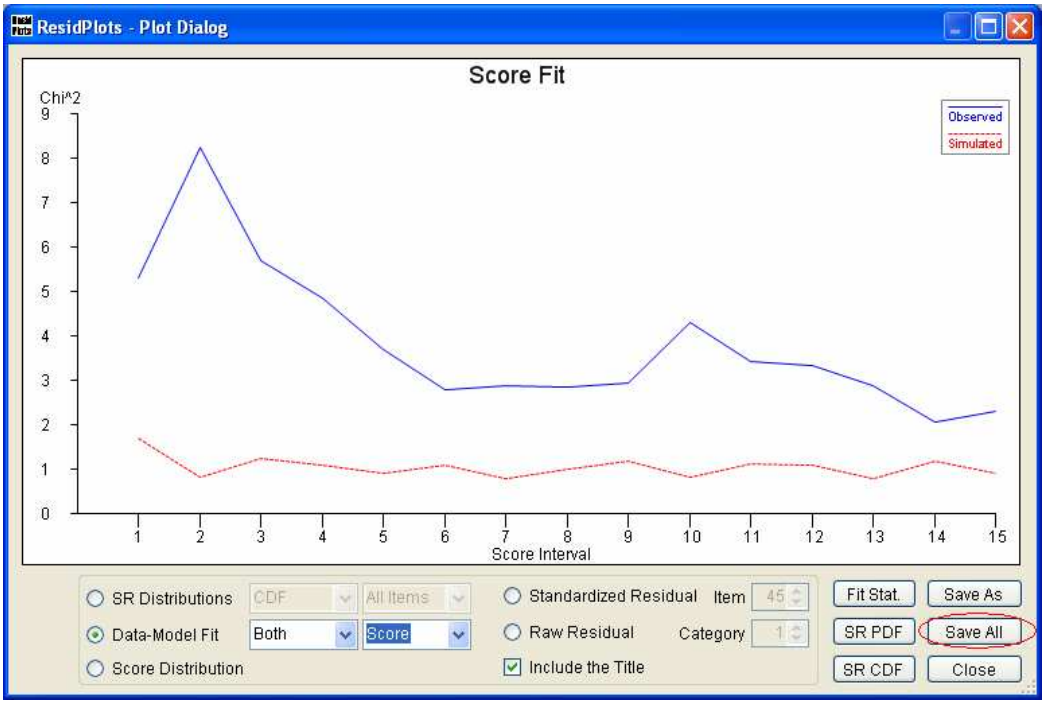
If users click 'save as', a dialog box will pop up. Users then select a folder to save the plot. Only the present plot can be saved under this option.





Example of saving all plots

If users click ‘save all’, a dialog box will pop up to let users select an existing folder or make a new folder to save all plots.

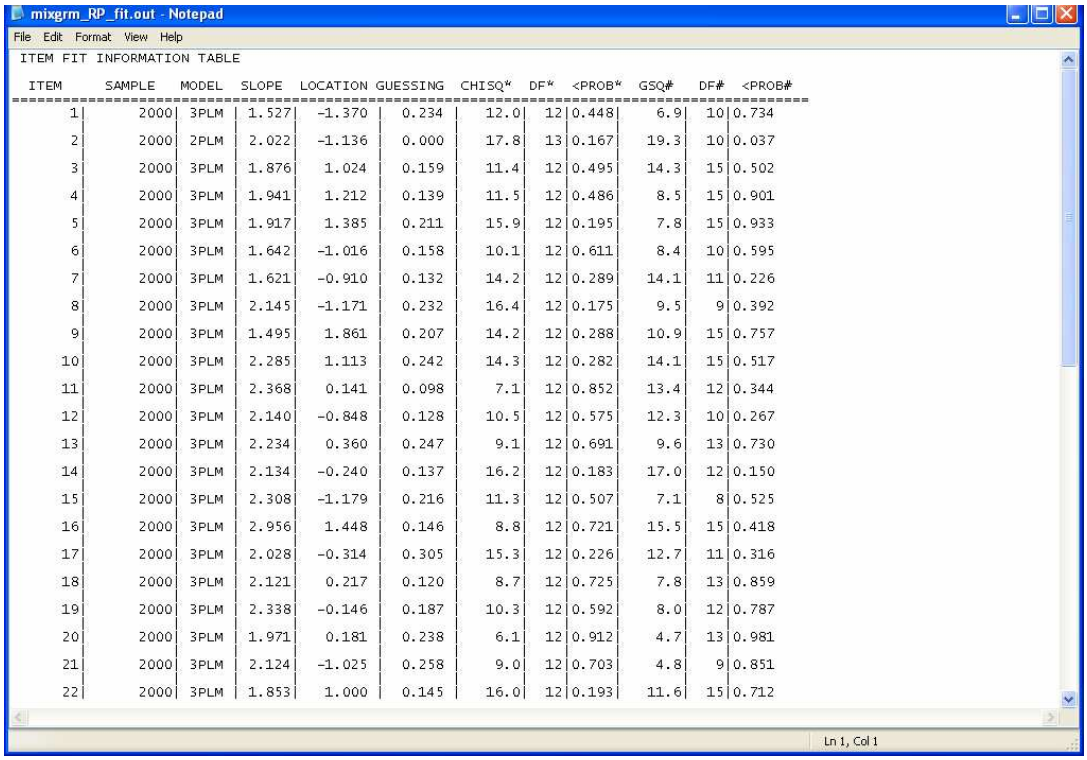


Tables

ResidPlots-2 provides four tables to report the data and fit information quantitatively. If users check "save the output files" option on the interface, all four tables (FIT STAT, SR PDF, SR CDF, REPORT) can be reached in the folder where the syntax file is located. Three of them (FIT STAT, SR PDF, SR CDF) can also be accessed on the plot screen because they provide information associated with certain plots. The following is a description of each table.

1. <u>FIT STAT table</u>

This table reports item number, sample size, item parameter estimates, chi square fit statistic (chi square value, degree of freedom, probability), G square fit statistic (G square value, degree of freedom, probability). Users may need to use this table to evaluate item fit as a reference. This table is an addition to the item fit plot.



mixgrm_RP_fit.out - Notepad

File Edit Format View Help

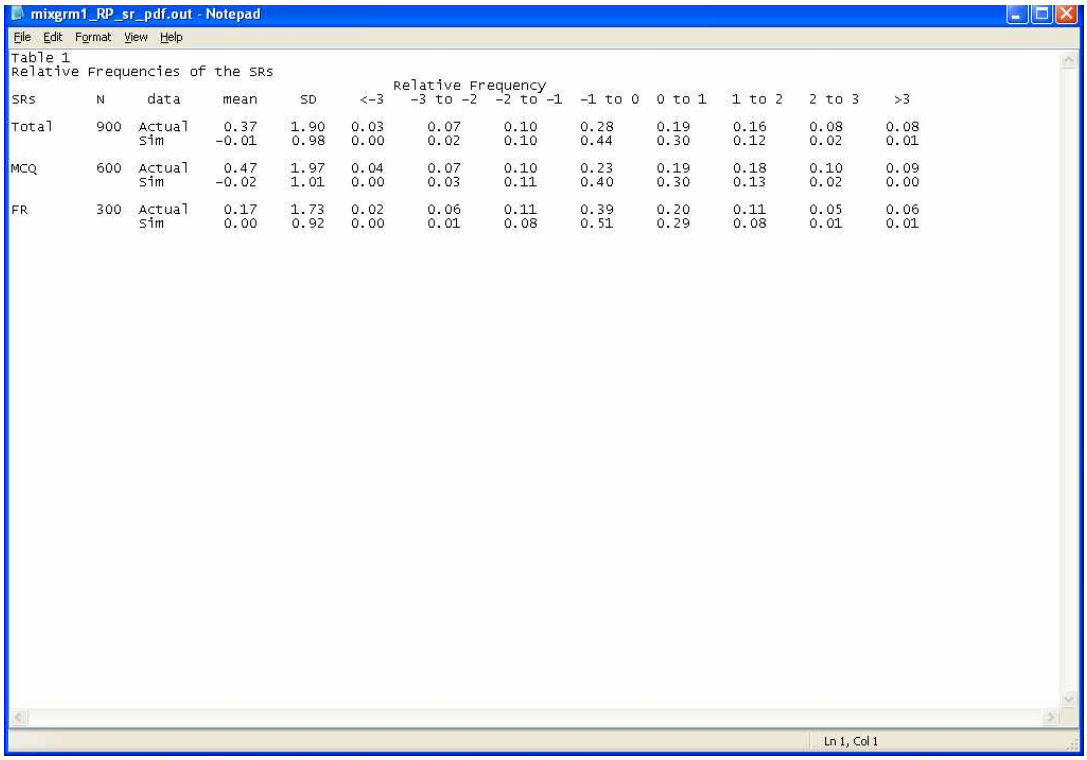
ITEM FIT INFORMATION TABLE

ITEM	SAMPLE	MODEL	SLOPE	LOCATION	GUESSING	CHISQ*	DF*	<PROB*	GSQ#	DF#	<PROB#
1	2000	3PLM	1.527	-1.370	0.234	12.0	12	0.448	6.9	10	0.734
2	2000	2PLM	2.022	-1.136	0.000	17.8	13	0.167	19.3	10	0.037
3	2000	3PLM	1.876	1.024	0.159	11.4	12	0.495	14.3	15	0.502
4	2000	3PLM	1.941	1.212	0.139	11.5	12	0.486	8.5	15	0.901
5	2000	3PLM	1.917	1.385	0.211	15.9	12	0.195	7.8	15	0.933
6	2000	3PLM	1.642	-1.016	0.158	10.1	12	0.611	8.4	10	0.595
7	2000	3PLM	1.621	-0.910	0.132	14.2	12	0.289	14.1	11	0.226
8	2000	3PLM	2.145	-1.171	0.232	16.4	12	0.175	9.5	9	0.392
9	2000	3PLM	1.495	1.861	0.207	14.2	12	0.288	10.9	15	0.757
10	2000	3PLM	2.285	1.113	0.242	14.3	12	0.282	14.1	15	0.517
11	2000	3PLM	2.368	0.141	0.098	7.1	12	0.852	13.4	12	0.344
12	2000	3PLM	2.140	-0.848	0.128	10.5	12	0.575	12.3	10	0.267
13	2000	3PLM	2.234	0.360	0.247	9.1	12	0.691	9.6	13	0.730
14	2000	3PLM	2.134	-0.240	0.137	16.2	12	0.183	17.0	12	0.150
15	2000	3PLM	2.308	-1.179	0.216	11.3	12	0.507	7.1	8	0.525
16	2000	3PLM	2.956	1.448	0.146	8.8	12	0.721	15.5	15	0.418
17	2000	3PLM	2.028	-0.314	0.305	15.3	12	0.226	12.7	11	0.316
18	2000	3PLM	2.121	0.217	0.120	8.7	12	0.725	7.8	13	0.859
19	2000	3PLM	2.338	-0.146	0.187	10.3	12	0.592	8.0	12	0.787
20	2000	3PLM	1.971	0.181	0.238	6.1	12	0.912	4.7	13	0.981
21	2000	3PLM	2.124	-1.025	0.258	9.0	12	0.703	4.8	9	0.851
22	2000	3PLM	1.853	1.000	0.145	16.0	12	0.193	11.6	15	0.712

Ln 1, Col 1

2. SR PDF table

This table reports mean, standard deviation, and relative frequency of the standardized residual distribution. "MCQ" refers to dichotomous items, "FR" refers to polytomous items, "TOTAL" refers to all items, "ACTUAL" refers to observed data, "SIM" refers to simulated data. This table is an addition to the SR distribution (PDF) plot, so users can see the specific frequency numbers from this table.



mixgrm1_RP_sr_pdf.out - Notepad

File Edit Format View Help

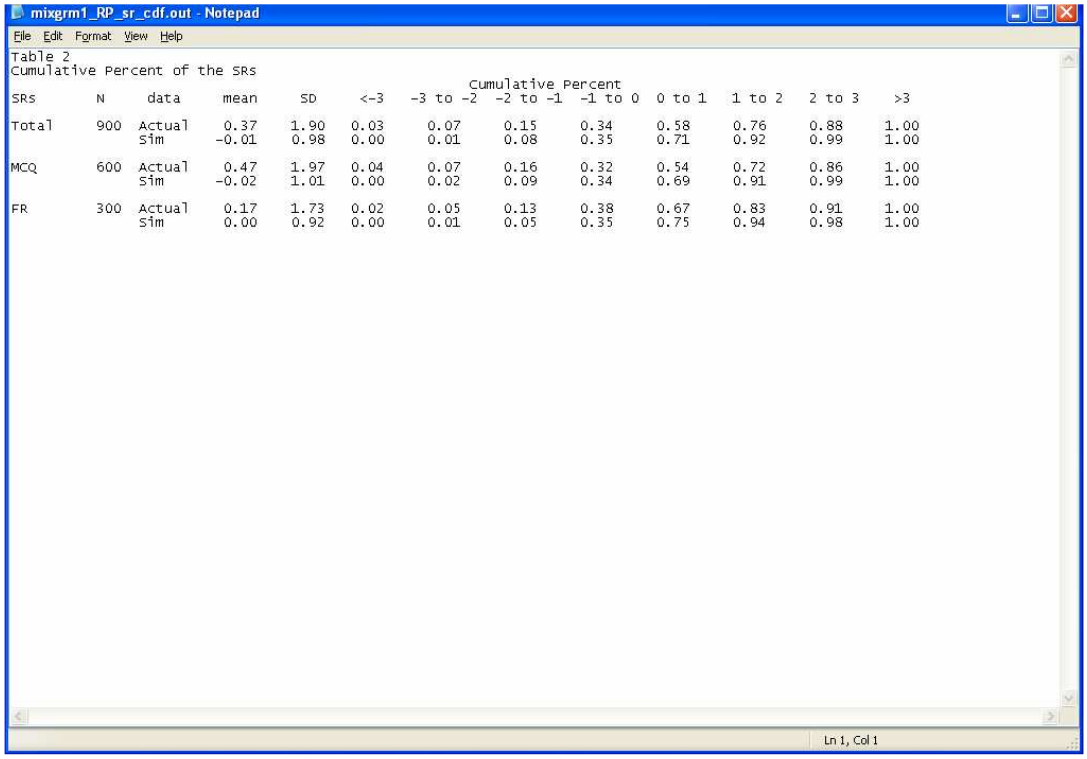
Table 1
Relative Frequencies of the SRs

					Relative Frequency						
SRs	N	data	mean	SD	<-3	-3 to -2	-2 to -1	-1 to 0	0 to 1	1 to 2	2 to 3	>3
Total	900	Actual	0.37	1.90	0.03	0.07	0.10	0.28	0.19	0.16	0.08	0.08
		Sim	-0.01	0.98	0.00	0.02	0.10	0.44	0.30	0.12	0.02	0.01
MCQ	600	Actual	0.47	1.97	0.04	0.07	0.10	0.23	0.19	0.18	0.10	0.09
		Sim	-0.02	1.01	0.00	0.03	0.11	0.40	0.30	0.13	0.02	0.00
FR	300	Actual	0.17	1.73	0.02	0.06	0.11	0.39	0.20	0.11	0.05	0.06
		Sim	0.00	0.92	0.00	0.01	0.08	0.51	0.29	0.08	0.01	0.01

Ln 1, Col 1

3. SR CDF table

This table reports the cumulative SR distribution (SR CDF) based on SR PDF.



mixgrm1_RP_sr_cdf.out - Notepad

File Edit Format View Help

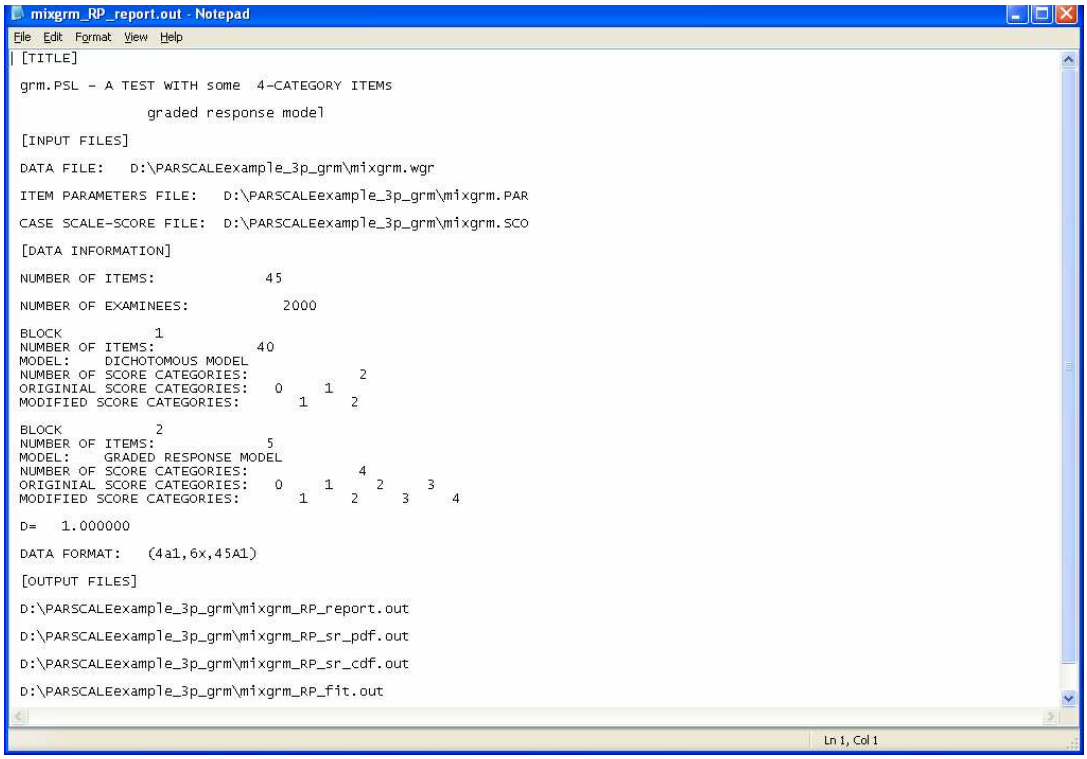
Table 2
Cumulative Percent of the SRs

					Cumulative Percent						
SRs	N	data	mean	SD	<-3	-3 to -2	-2 to -1	-1 to 0	0 to 1	1 to 2	2 to 3	>3
Total	900	Actual	0.37	1.90	0.03	0.07	0.15	0.34	0.58	0.76	0.88	1.00
		Sim	-0.01	0.98	0.00	0.01	0.08	0.35	0.71	0.92	0.99	1.00
MCQ	600	Actual	0.47	1.97	0.04	0.07	0.16	0.32	0.54	0.72	0.86	1.00
		Sim	-0.02	1.01	0.00	0.02	0.09	0.34	0.69	0.91	0.99	1.00
FR	300	Actual	0.17	1.73	0.02	0.05	0.13	0.38	0.67	0.83	0.91	1.00
		Sim	0.00	0.92	0.00	0.01	0.05	0.35	0.75	0.94	0.98	1.00

Ln 1, Col 1

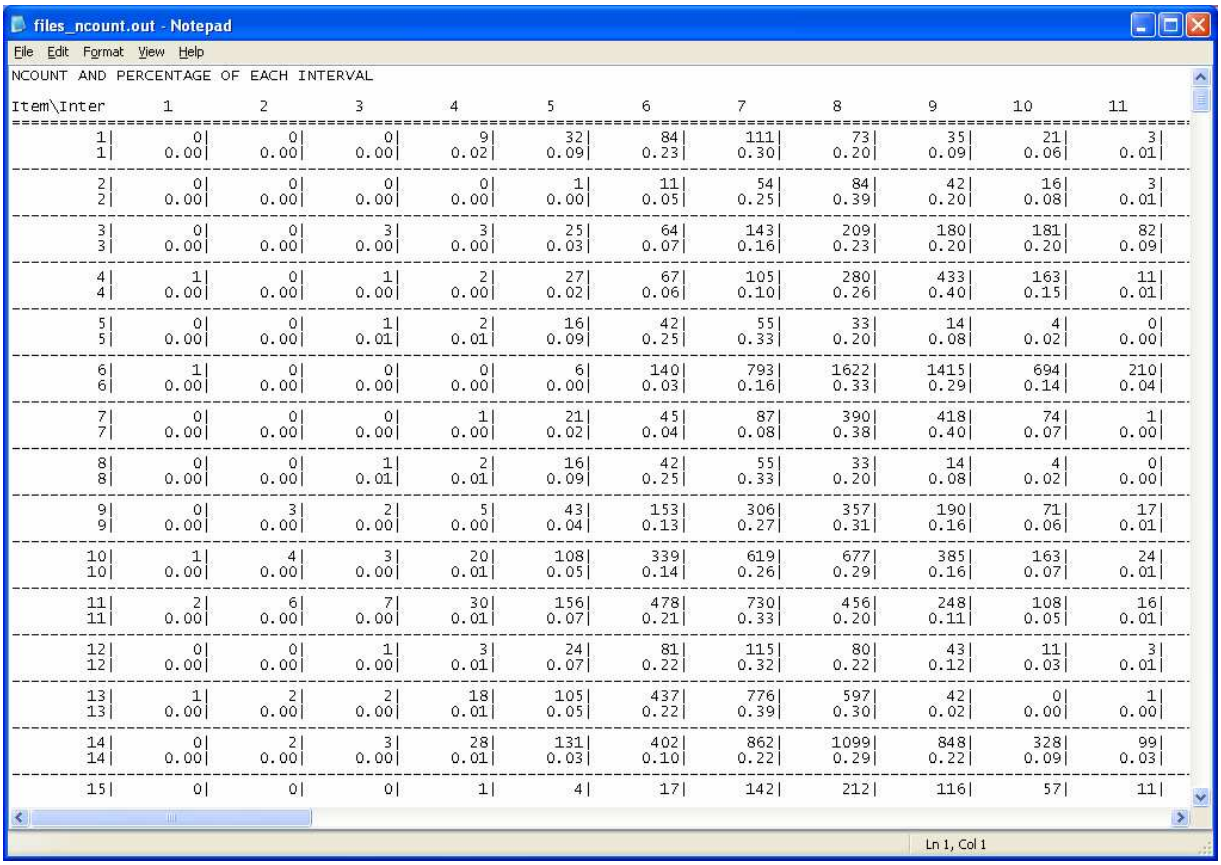
4. REPORT table

This is a check table. This table can be used to double check if the input information of *ResidPlots-2* matches the information from the syntax of PARSCALE or BILOG-MG.



5. NCOUNT table

Due to the different features of data and users' choices of intervals and score ranges, the sample size and percentage of the sample in each interval are reported in this table. The first column provides the item number, the first row provides the frequency counts and the second row provides the percentages. If there are no persons in an interval, the data point is suppressed in both the residual and standardized residual plots. This information though would not be designated in the display of data itself.



files_ncount.out - Notepad

File Edit Format View Help

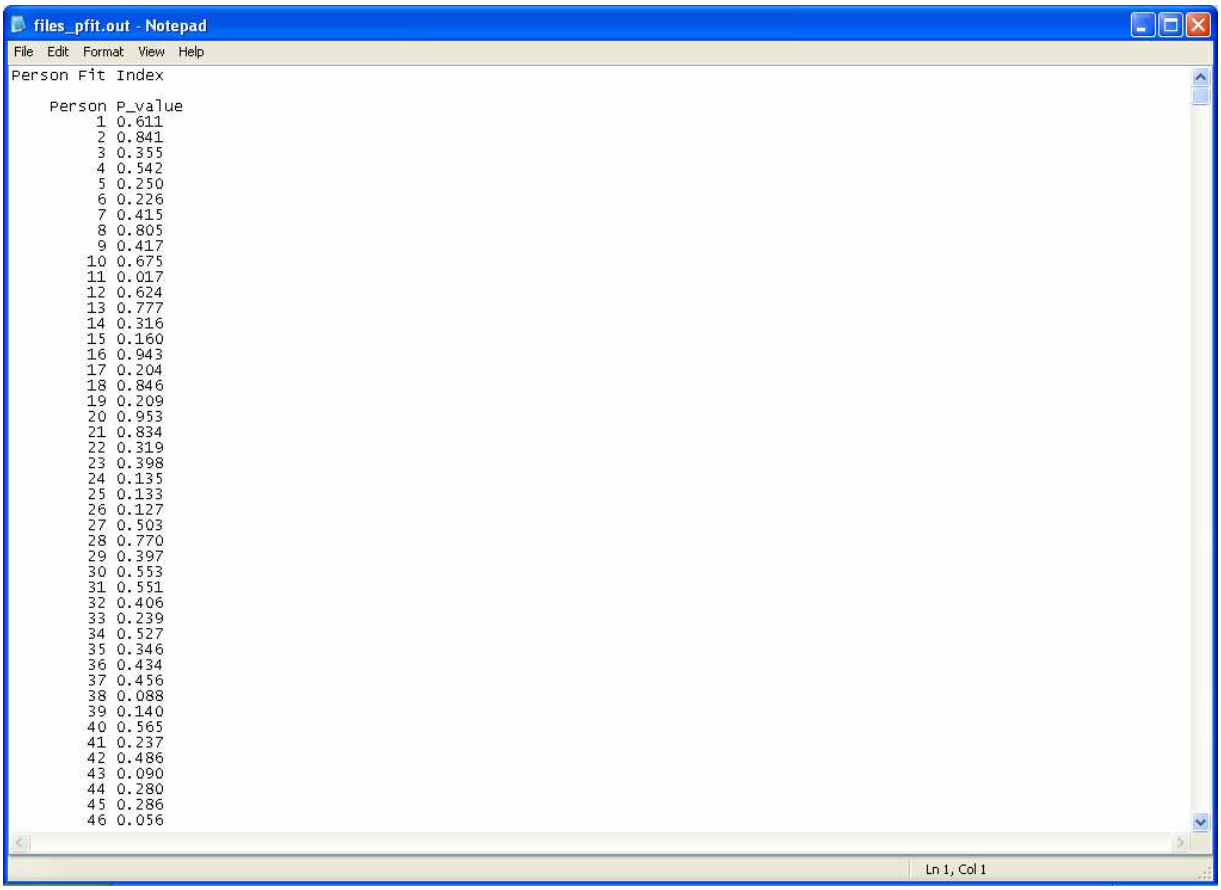
NCOUNT AND PERCENTAGE OF EACH INTERVAL

Item\Inter	1	2	3	4	5	6	7	8	9	10	11
1	0	0	0	9	32	84	111	73	35	21	3
1	0.00	0.00	0.00	0.02	0.09	0.23	0.30	0.20	0.09	0.06	0.01
2	0	0	0	0	1	11	54	84	42	16	3
2	0.00	0.00	0.00	0.00	0.00	0.05	0.25	0.39	0.20	0.08	0.01
3	0	0	3	3	25	64	143	209	180	181	82
3	0.00	0.00	0.00	0.00	0.03	0.07	0.16	0.23	0.20	0.20	0.09
4	1	0	1	2	27	67	105	280	433	163	11
4	0.00	0.00	0.00	0.00	0.02	0.06	0.10	0.26	0.40	0.15	0.01
5	0	0	1	2	16	42	55	33	14	4	0
5	0.00	0.00	0.01	0.01	0.09	0.25	0.33	0.20	0.08	0.02	0.00
6	1	0	0	0	6	140	793	1622	1415	694	210
6	0.00	0.00	0.00	0.00	0.00	0.03	0.16	0.33	0.29	0.14	0.04
7	0	0	0	1	21	45	87	390	418	74	1
7	0.00	0.00	0.00	0.00	0.02	0.04	0.08	0.38	0.40	0.07	0.00
8	0	0	1	2	16	42	55	33	14	4	0
8	0.00	0.00	0.01	0.01	0.09	0.25	0.33	0.20	0.08	0.02	0.00
9	0	3	2	5	43	153	306	357	190	71	17
9	0.00	0.00	0.00	0.00	0.04	0.13	0.27	0.31	0.16	0.06	0.01
10	1	4	3	20	108	339	619	677	385	163	24
10	0.00	0.00	0.00	0.01	0.05	0.14	0.26	0.29	0.16	0.07	0.01
11	2	6	7	30	156	478	730	456	248	108	16
11	0.00	0.00	0.00	0.01	0.07	0.21	0.33	0.20	0.11	0.05	0.01
12	0	0	1	3	24	81	115	80	43	11	3
12	0.00	0.00	0.00	0.01	0.07	0.22	0.32	0.22	0.12	0.03	0.01
13	1	2	2	18	105	437	776	597	42	0	1
13	0.00	0.00	0.00	0.01	0.05	0.22	0.39	0.30	0.02	0.00	0.00
14	0	2	3	28	131	402	862	1099	848	328	99
14	0.00	0.00	0.00	0.01	0.03	0.10	0.22	0.29	0.22	0.09	0.03
15	0	0	0	1	4	17	142	212	116	57	11

Ln 1, Col 1

6. PFIT Table

This table contains a commonly used IRT-based person fit statistic called “Lz” (Drasgow, Levine & Williams, 1985). Lz is a standardized statistic and follows a standard normal distribution. Thus, an examinee with an Lz value of -1.95 might not be considered to fit the model(s) in a test because the Lz value represents a value beyond the 0.05 error rate of -1.65. In ResidPlots-2, for users’ convenience, the probability values instead of Z values are reported. So, if the users’ probability criterion is 0.05, a p value below 0.05 in the second column indicate person misfit. As shown below, the first column and second column provide the person number and corresponding p value, respectively.



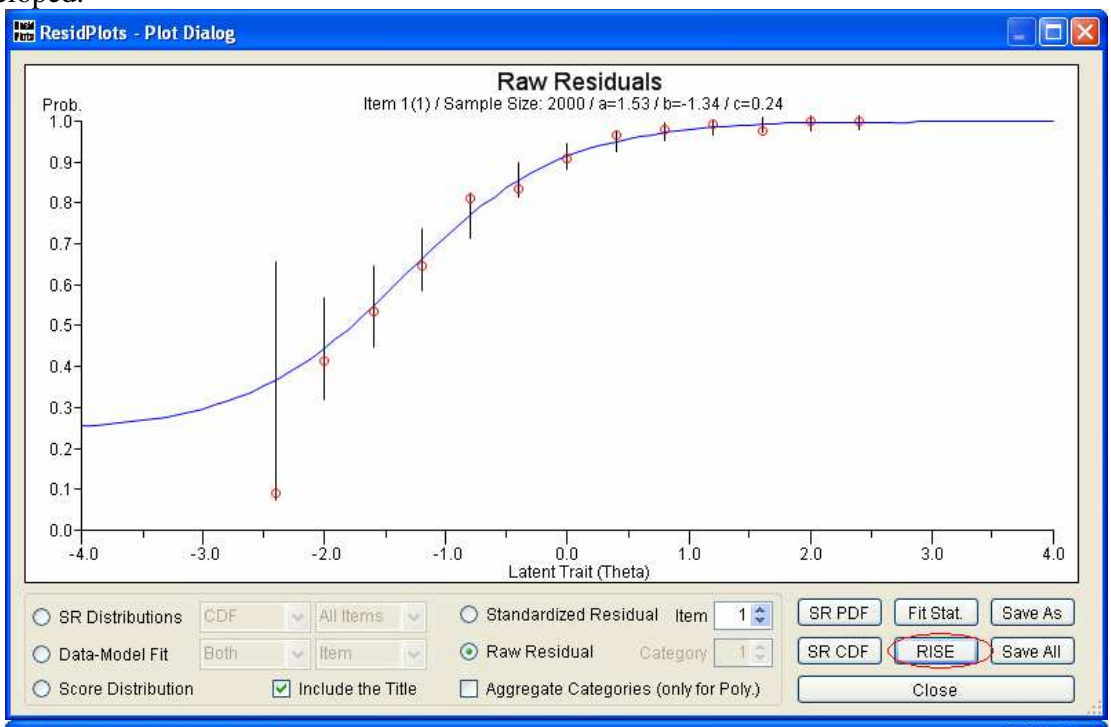
7. P_RISE Table

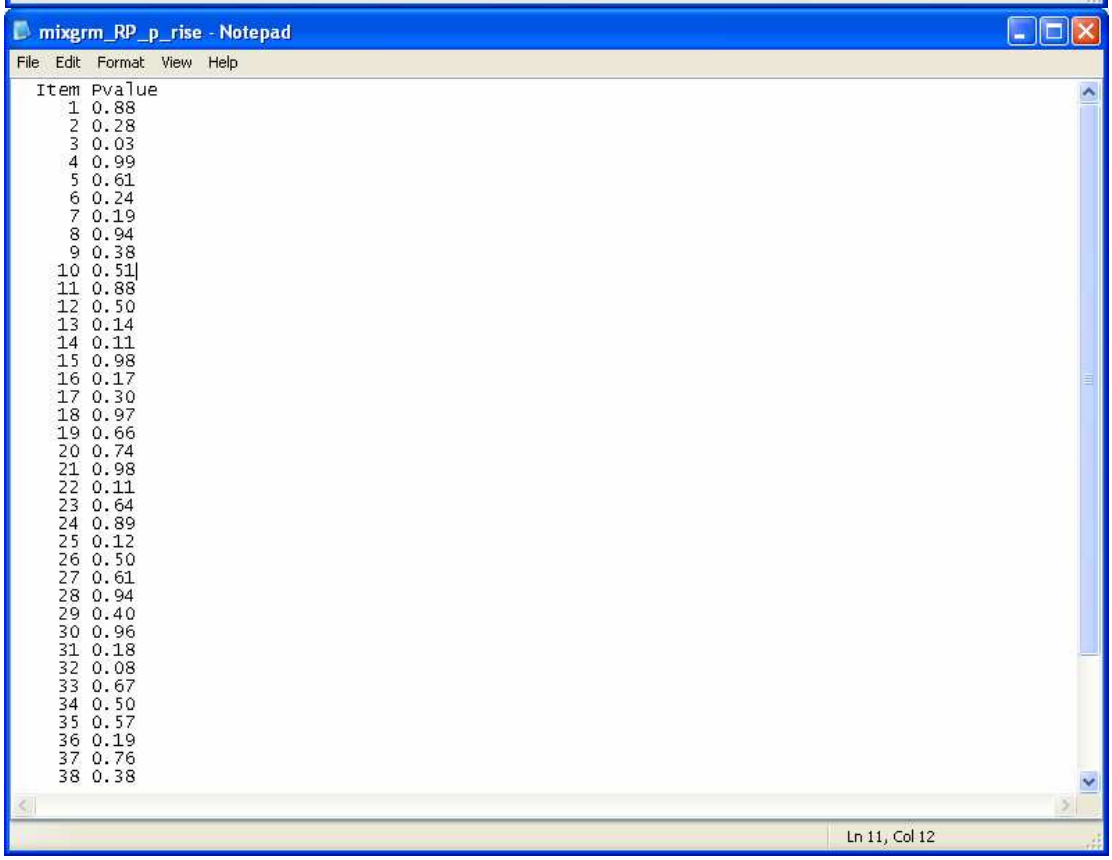
This table provides a nonparametric fit statistic called "root integrated square error" (RISE). Users can obtain this table by clicking the "RISE" button on the bottom of the plot screen (see below). In the table, the first column is the item ID, and the second column is the probability value for the item. The probability criterion to judge misfit has been set to a value of 0.05 in the software, so if an item has a p value less than 0.05, then it is identified as a misfitting item.

This nonparametric approach was originally proposed by Douglas and Cohen (2001) and has been examined closely for its statistical properties and graphical displays (Wells & Bolt, 2008, Li & Wells, 2006, Liang & Wells, in press, Liang & Wells, 2008). Research studies have demonstrated RISE can exhibit controlled Type I error rate and adequate power as well as provide convenient graphical displays to locate misfit. The promising features of RISE compared to parametric approaches have been fully explained in the referenced papers.

The current option (RISE) in ResidPlots-2 is so users can address the statistical fit of test items. However, as this new fit statistic is still at the research stage, several problems and issues need to be solved. For now, users are encouraged to use it as one more tool in their investigation of model fit. Also, in this current version of the software, only the 2PLM, 3PLM, GRM and GPCM are supported. Finally, because this fit statistic is nonparametric, it would be more

accurate to associate each p value with a nonparametric graphical display of residuals but those are not yet available in the software. The nonparametric residual plots are currently being developed.





PART III

Technical Details

1. Standardized residuals

Standardized residuals are the basis for the plot for each item, standardized residual distribution (PDF and CDF), item fit plot, and score fit plot. It is calculated as follows:

$$SRj = \frac{Oj - Ej}{\sqrt{\frac{Ej(1-Ej)}{Nj}}}$$

where Oj is the observed proportion of correct answers for examinees in a score interval, Ej is the expected (model-based) proportion of correct answers in the same score interval, Nj is the number of examinees in the same score interval.

(a) Standardized residual plot

The standardized residual in each score interval is shown on the plot. If there are no examinees in an interval, the standardized residual is not shown in the display.

(b) Standardized residual distribution (PDF and CDF)

The standardized residual frequency distribution (PDF) is based on all SRs (intervals × items) in the test excluding those with intervals with zero frequencies. The cumulative distribution (CDF) is based on the PDF.

(c) Item fit plot

Each point on this plot is a chi-square value of the item. The chi square value is the sum of squares of standardized residuals over the score intervals. This plot contains the chi-square value for each item in the FitSTAT table.

(d) Score fit plot

Each point on this plot is calculated for each score interval, and is the sum of squares of standardized residuals across all items. In order to scale them, each point is divided by the number of items and score categories for the item.

2. Chi-square statistic

This statistic is reported in the Fit STAT table. For each item, it is calculated as follows:

$$\sum_{j=1}^{K} \frac{N_j(O_{ij} - E_{ij})^2}{E_{ij}(1-E_{ij})}$$

where N_j is the number of examinees in score interval j; O_{ij} is the observed proportion of examinees in interval j who answer item i correctly; and E_{ij} is the probability based on the model in interval j answering item i correctly. Degrees of freedom equals the number of score intervals minus the number of parameters being estimated.

3. G^2 statistic from PARSCALE or BILOG-MG

This statistic is directly adopted from PARSCALE or BILOG-MG output in .PH2. It is also reported in the Fit STAT table in *ResidPlots-2*.

$$2\sum_{h=1}^{n_g}\left[r_{hi}\log\frac{r_{hi}}{N_h P_i(\overline{\theta}_h)} + (N_h - r_{hi})\log\frac{N_h - r_{hi}}{N\left[1-P_i(\overline{\theta}_h)\right]}\right]$$

where n_g indicates the number of intervals; r_{hi} represents the observed frequency of correct responses for item i in interval h; N_h is the number of examinees in interval h; and $P_i(\overline{\theta}_h)$ refers to the model-predicted proportion correct for item i at $\overline{\theta}_h$. G^2 is distributed approximately as a chi-square with degrees of freedom equal to the number of intervals.

4. Predicted score distribution

ResidPlots-2 uses the item and ability estimates from MULTILOG, PARSCALE or BILOG-MG to simulate examinee response data. We repeat this simulation in the software 10 times to get 10 observed test score distributions under the assumption of model fit, and then average the distributions to get something that better approximates the expected observed test score distribution. (Ten is an arbitrary choice, but it is large enough to smooth out the expected test score distribution from a single simulation and provides a good estimate of the observed test score distribution, under the assumption that the model and the estimated item parameters are true.) The average density and cumulative distributions can then be compared to the observed test score density and cumulative distributions obtained from the real data. When they are close, it can be said that the best fitting IRT model closely recovers or predicts the actual test score distribution for the examinees who were administered the test. When they are not close, model fit can be questioned. It's a judgment as to how close the distributions need to be to establish model fit. Interpretation is enhanced by comparing the fit for more than one model to provide a basis for interpreting the results. There are both parametric as well as non-parametric statistical tests that can be used to compare the distributions, but to date, we have not included them in the software.

5. Lz (Drasgow, Levine and Williams, 1985)

This statistic is reported in the PFIT table. For each person, the general formula for both dichotomous and polytomous models is calculated as follows:

$$Lz = \frac{L_0 - E(L_0)}{[\mathrm{var}(L_0)]^{1/2}}$$

where L_0, $E(L_0)$, $\mathrm{var}(L_0)$ are defined as the log of the peak of the likelihood function, the expectation of L_0, the variance of L_0, respectively:

$$L_0 = \sum_{i=1}^{k} \ln P_i(\theta)$$

$$E(L_0) = \sum_{i=1}^{k} \sum_{x=0}^{m} P_i(x \mid \theta) \ln P_i(x \mid \theta)$$

$$Var(L_0) = \sum_{i=1}^{k} [\sum_{x=0}^{m} \sum_{y=0}^{m} P_i(x \mid \theta) P_i(y \mid \theta) \ln P_i(x \mid \theta) \ln(\frac{P_i(x \mid \theta)}{P_i(y \mid \theta)})]$$

This statistic can be used to identify examinees for whom the current IRT model is inadequate to account for their item response data.

6. <u>Item Level Plots for Polytomously-Scored Items</u>

With polytomously-scored items and with small samples, and/or with the desire to simplify the model fit output, it is sometimes of interest to look at model misfit at the item level (summing over probability curves for possible score points). At the item level too, the expected score over the theta scale looks a lot like an ICC, and with the actual scores (conditional average item level scores) plotted at intervals along the ability continuum, the display looks similar to the displays for ICCs. This feature has been included only for polytomously-scored items fitted by GRM and GPCM in the software. The calculation of expected item score is shown below:

$$E_i = \sum_{h=0}^{m_j} h P_{ijh}$$

where i indicates each theta level, j represents each item, h is for each category ranging from 0 to m_j, P_{ijh} is the model-based probability at each theta for each score category h for item j. What is being plotted in the display is the item mean for the examinees at each theta level and it can be compared to the expected score level for judging model fit.

PART-IV

Future Features of *ResidPlots-2*

We anticipate producing several updates of the software—both major and minor. Our initial list of planned changes is summarized below:

- We recognize the interest in item and test statistics for model fit. In future versions we

will be studying the work of Drasgow et al. (1995), Sinharay (see, for example, 2005), and van der Ark (2001) for additional ideas for displays that would be useful to add to the software.

- We anticipate making it possible for users to use the quadrature points and the associated frequencies from commercial software (e.g., PARSCALE) for graphing fit information.

- We will expand the model fit software next to include the rating scale model. Other models will be added as there seems to be a need for them.

Users with other ideas for expanding and/or improving the software should contact Ms. Tie Liang at tliang@educ.umass.edu or Ronald Hambleton at rkh@educ.umass.edu.

References

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51.

Douglas, J. & Cohen, A. S. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement, 25*, 234-243.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19*, 143-165.

Drasgow, F., Levine, M. V., Williams, E. A. (1985). Appropriateness measurement with polytomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67-86.

Hambleton, R. K., & Han, N. (2005). Assessing the fit of IRT models to educational and psychological test data: A five step plan and several graphical displays. In W. R. Lenderking & D. Revicki (Eds.), *Advances in health outcomes research methods, measurement, statistical analysis, and clinical applications* (pp. 57-78). Washington: Degnon Associates.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Li, S., & Wells, C. S. (2006, April). *A model fit statistic for Samejima's graded response model.* Paper presented at the meeting of the National Council on Measurement in Education, San Francisco.

Liang,T., & Wells, C.S. (in press). A model fit statistic for generalized partial credit model. *Educational and Psychological Measurement.*

Liang, T., & Wells, C. S. (2008, October). *A nonparametric approach for assessing model fit in a mixed format test.* Paper presented at the meeting of the National Council on Measurement in Education, NY.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement*, *9*, 49-57.

Muraki, E., & Bock, R. D. (2003). *PARSCALE 4: IRT item analysis and test scoring for rating-scale data* [computer program]. Chicago, IL: Scientific Software.

Orlando, M., & Thissen, D. (2000). New item fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*, 50-64.

Rogers, H. J., & Hattie, J. A. (1987). A Monte Carlo investigation of several person and item fit statistics for item response models. *Applied Psychological Measurement, 11*, 47-57.

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement, 42*(4), 375-394.

Swaminathan, H., Hambleton, R. K., & Rogers, H. J. Assessing the fit of item response theory models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics: Psychometrics* (pp. 683-718). Amsterdam: Elsevier.

Thissen, D. (2003). *MULTILOG* 7: *Multiple categorical item analysis and test scoring using item response theory* [computer program]. Chicago, IL: Scientific Software.

van der Ark, A. L. (2001). Relationships and properties of polytomous item response theory models. *Applied Psychological Measurement, 25*(3), 273-282.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York: Springer-Verlag.

Wells, C. S., & Bolt, D. M. (2008). Investigation of a nonparametric procedure for assessing goodness-of-fit in item response theory. *Applied Measurement in Education*, *21*, 22-40.

Yen, W. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245-262.

Zhao, Y. (2008). *Approaches for addressing the fit of item response theory models to educational test data*. Unpublished doctoral thesis, University of Massachusetts Amherst.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3* [computer program]. Chicago, IL: Scientific Software.