MEASUREMENT: INTERDISCIPLINARY RESEARCH AND PERSPECTIVES
2019, VOL. 17, NO. 1, 48–59
https://doi.org/10.1080/15366367.2018.1539558

Routledge
Taylor & Francis Group

SOFTWARE REVIEW

Software Review of IRTEQ, STUIRT, and POLYEQUATE for Item Response Theory Scale Linking and Equating

Jaime Malatesta and Won-Chan Lee

Center for Advanced Studies in Measurement and Assessment (CASMA), University of Iowa

ABSTRACT

This article reviews several software programs designed to conduct item response theory (IRT) scale linking and equating. The programs reviewed include IRTEQ, STUIRT, and POLYEQUATE. Features and functionalities of each program are discussed and an example analysis using the common-item non-equivalent groups design in IRTEQ is provided.

KEYWORDS

Item response theory; scale linking; equating; common-item non-equivalent groups design; IRTEQ; STUIRT; POLYEQUATE

Introduction

For many testing programs it is often necessary to administer alternate forms of a test due to circumstances such as multiple testing dates and threats to test security. As a result, new forms are continuously being built and administered to new groups of examinees. A data collection design in which different forms of a test are administered to different groups of examinees is often referred to as the common-item non-equivalent groups design (CINEG; Kolen & Brennan, 2014). With this design (and others), it is desirable that scores from the new forms can be directly compared to scores on the old form.

However, under the CINEG design, two issues need to be addressed before direct score comparisons can be made. First, the effect of different ability distributions between testing cohorts needs to be adjusted using some form of *scale linking*. Second, differences in form difficulties need to be adjusted using some form of *equating* methodology. Over time, various item response theory (IRT) scale linking and equating methods have been developed to address these needs. While a thorough treatment of these methods is beyond the scope of this article, interested readers can refer to Kolen and Brennan (2014) or Dorans, Pommerich, and Holland (2007).

The focus of this article is to review three software programs that can be used to conduct IRT scale linking and/or IRT equating: STUIRT (Kim & Kolen, 2004), the Windows console version of POLYEQUATE (Kolen, 2004), and IRTEQ (Version 1.04) (Han, 2009a). STUIRT stands for scale transformation under unidimensional IRT models and can implement several IRT scale linking methods. POLYEQUATE is a software program designed to conduct IRT true score equating (TSE) and IRT observed score equating (OSE). In this article, STUIRT and POLYEQUATE are often discussed together because while the two are stand-alone programs, a significant portion of the input required by POLYEQUATE is purposely generated by STUIRT. Unlike the first two software programs, IRTEQ can conduct various IRT scale linking methods, along with TSE. In this article, we compare the functions and capabilities of each program and provide a basic example of IRT scale linking with TSE under the CINEG design. Technically, these software programs are not restricted to the CINEG design and can be used with essentially any data collection design. However, the CINEG design is highlighted in this paper because scale linking is most commonly associated with it.

CONTACT Jaime Malatesta Jaime-malatesta@uiowa.edu Center for Advanced Studies in Measurement and Assessment (CASMA), University of Iowa, Iowa City, IA 52242.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hmes.

Features and functions

In this section, we compare STUIRT, POLYEQUATE, and IRTEQ with respect to their scope of IRT models and methods, technical options, program inputs, outputs, documentation, and user-friendliness. Table 1 provides an overview of each program's main features.

Capabilities

All of the software programs discussed in this article assume test data are unidimensional and primarily measure a single latent trait. As a result, the scale linking and equating methods built into each program are specific to unidimensional IRT (UIRT) models. Each program offers the benefit of being able to handle mixed-format test data frequently used by large-scale testing programs (Kolen & Brennan, 2014; Kolen & Lee, 2011, 2012, 2014, 2016, 2018). Even though mixed-format data tend to be somewhat multidimensional (Lee & Brossman, 2012; Tate, 2000; Yao & Boughton, 2009), UIRT models and UIRT scale linking and equating methods are most often used. Fortunately, UIRT equating results have been found to be robust to moderate violations of the unidimensionality assumption (Bolt, 1999; Camilli, Wang, & Fesq, 1995; Cook, Dorans, Eignor, & Petersen, 1985; de Champlain, 1996; Dorans & Kingston, 1985; Yen, 1984). Even so, the UIRT unidimensionality assumption should hold reasonably well before STUIRT, POLYEQUATE, or IRTEQ are used.

In general, the most common UIRT models can be handled by STUIRT, POLYEQUATE, and IRTEQ. These include the one-, two-, or three-parameter logistic (3PL; Birnbaum, 1968) models for dichotomous items and the graded response (GR; Samejima, 1969) and generalized partial credit (GPC; Muraki, 1992) models for polytomous items. IRTEQ can also handle Masters and Wright's (1997) partial credit (PC) model, which is a special case of the GPC model. While, STUIRT and POLYEQUATE do not accommodate the PC model directly, it can be modeled indirectly using the GPC model. Furthermore, STUIRT and POLYEQUATE can handle the nominal response (NR;

Table 1. Overview of program features and functions.

Software Feature	STUIRT	POLYEQUATE	IRTEQ
Capabilities			
Modality	Syntax and command-line	Syntax and command-line	Syntax or point-and-click
# examinees	Unlimited	Unlimited	Unlimited
# items	Unlimited	Unlimited	Unlimited
# item response categories	Unlimited	Unlimited	Unlimited
Item scoring functions	Program- or user-supplied	Program- or user-supplied	Program-supplied
IRT models	1PL, 2PL, 3PL, GR, NR, GPC, MC	1PL, 2PL, 3PL, GR, NR, GPC	1PL, 2PL, 3PL, GR, PC, GPC
Scale	Logistic, Normal	Logistic, Normal	Logistic, Normal
Linking methods	MS, MM, HA, SL	N/A	MS, RMS, MM, HA, SL
Equating methods	N/A	TSE, OSE	TSE
Data collection design	CINEG	Any	CINEG
Outputs			
Old form equivalents	N/A	Unrounded raw scores, rounded raw scores, unrounded scale scores, rounded scale scores	Unrounded raw scores
Equated examinee scores	No	N/A	Yes
Marginal distributions for old and new form scores	N/A	Yes	No
Summary moments for the old and new form scores	N/A	Yes	No
# of decimal places	5	5	2 and 3

Bock, 1972), of which, the GPC model is a special case. STUIRT is the only software program that can also accommodate Thissen and Steinberg's (1984) multiple-choice (MC) model. Each program can handle a mixture of dichotomous and polytomous UIRT models in a single analysis, and allows users to specify the scaling constant, *D*, as 1 or 1.7 to obtain results in the logistic or normal metric, respectively.

STUIRT and IRTEQ can each implement the Mean/Sigma (MS; Marco, 1977), Mean/Mean (MM; Loyd & Hoover, 1980), Haebara (HA; Haebara, 1980), and Stocking-Lord (SL; Stocking & Lord, 1983) scale linking methods. In addition to these methods, IRTEQ can implement the Robust Mean/Sigma (RMS; Linn, Levine, Hastings, & Wardrop, 1981) linking method.

Although, STUIRT and IRTEQ offer a similar set of scale linking methods, STUIRT offers a variety of technical options that control different aspects of the test characteristic curve (TCC) methods (e.g., HA and SL). For example, using keywords, the user can control the following: (a) starting values for the slope and intercept, (b) maximum number of iterations, (c) proficiency distributions and weights of the old and new form groups for the criterion functions, (d) standardization of the criterion functions, (e) whether the criterion functions are defined on the old scale, new scale, or both, and (f) local minimum search. IRTEQ also allows users to modify the proficiency distribution and weights of the old and new form groups used in criterion function but in a limited number of ways compared to STUIRT. For example, in IRTEQ the criterion functions are weighted using one of three proficiency distributions: (a) a uniform distribution with user-specified endpoints, (b) a normal distribution with user-specified mean and standard deviation, or (c) the actual examinee distribution. In STUIRT, the criterion functions can be weighted in one of the six ways outlined in Kolen and Brennan (2014, p. 186).

There are additional ways that STUIRT and IRTEQ differ with respect to how IRT scale linking is performed. First, for the moment methods, IRTEQ allows for added control over how the rescaled estimates of the common-items (CIs) on the new form are computed. For example, users can specify whether the final rescaled CI estimates reflect an average of the old form and rescaled new form estimates (Hambleton, Swaminathan, & Rogers, 1991, pp. 137–138), or just the rescaled estimates (the latter is comparable to STUIRT). Second, STUIRT allows users to choose which response categories are included in the computation of the scale transformation coefficients (i.e., the slope and intercept). For example, for a polytomous item with five response categories, the first and last response category may be deselected using the subkeyword **NA** (not applicable).

Once the item and examinee estimates from the new form are placed on the scale of the old form using one of the aforementioned scale linking methods, IRT equating can be used to obtain old form equivalent scores for examinees who took the new form. IRTEQ has the added benefit of being able to conduct both IRT scale linking and equating within the same run, whereas STUIRT is only equipped to perform scale linking. To get equating results after using STUIRT, additional computations are required. POLYEQUATE appears to be a logical choice to obtain these calculations given that required inputs for it can be directly extracted from the main STUIRT output file. For this reason, the IRT equating capabilities of IRTEQ are compared with those of POLYEQUATE (Windows console version).

The biggest difference between IRTEQ and POLYEQUATE is that the former can be used to conduct just TSE whereas the latter can conduct both TSE and OSE. Furthermore, POLYEQUATE provides equating results for raw scores, unrounded scale scores, and rounded scale scores, whereas IRTEQ provides results for just raw scores. POLYEQUATE is compatible with the same set of IRT models as STUIRT, with the exception of the MC model. When the 3PL model is used, old form equivalents are undefined for new form raw scores below the sum of the *c*-parameters. When this happens old form equivalents are found in POLYEQUATE using the ad hoc interpolation method described in Kolen (1981). Similarly, some form of linear interpolation is also used in IRTEQ.

Installation

The installation of each software program requires minimal effort by the user. STUIRT and POLYEQUATE can be freely downloaded as .zip files from the website for the Center for Advanced Studies in Measurement

and Assessment (CASMA) at the University of Iowa (https://education.uiowa.edu/centers/casma). Inside each programs' .zip file are example input and output files, the manual, and executable program file.

IRTEQ was built for the Windows operating system and can be freely downloaded from the author's website (www.hantest.net) or from the Research, Educational Measurement, and Psychometrics homepage at the University of Massachusetts at Amherst (https://www.umass.edu/remp/main_software.html). New users of IRTEQ will need to download three files: 1) a .zip file that contains the application files, executable setup file, and executable program file, 2) a separate .zip file that contains example files, and 3) the .pdf manual. In order for the example files to run properly, they must be extracted to the user's C: drive, in the folder, "C:\IRTEQ\EXAMPLES."

User interface (UI) and required inputs

IRTEQ and STUIRT/POLYEQUATE differ quite a bit with respect to how the user interacts with them. For example, IRTEQ can be completely run through a simple graphical UI (GUI) or through a syntax-based control file. IRTEQ also allows users to set up analyses in the GUI and then save it as a .syn (syntax) file which can be run later. In addition, IRTEQ has a built-in capacity for running multiple syntax files at one time with a cue file (similar to a batch file in the DOS system). This last feature makes running simulation-type studies very easy to carry out.

By comparison, STUIRT and POLYEQUATE do not have a GUI and instead are each run through command-line prompts and control cards. For STUIRT, the user opens the executable program file and is prompted to enter the name of the input and output files (contents of each will be described later). Upon opening the executable POLYEQUATE file, the user is prompted to enter the control card filename followed by an "enter" to exit the program. The POLYEQUATE control card file contains keywords and filenames for the necessary input files. STUIRT and POLYEQUATE can also be used for simulation-like studies but users need to create a separate batch file and have some familiarity with DOS commands.

The number of input files required by IRTEQ is fewer than that required by the combination of STUIRT and POLYEQUATE. For example, to conduct IRT scale linking and equating using IRTEQ, users must prepare separate item parameter files for the old and new forms, in either .par (PARSCALE; Muraki & Bock, 2003) or .wgi (WinGen; Han, 2007) file formats. The .par file format must be used in order to save the rescaled item parameter estimates in the PARSCALE format. The user must also create a linking item list (.lil) file if they do not want to enter them manually inside the GUI. If the user wants to save the equated scores for each new form examinee, they must also provide a score file containing examinee theta estimates, in either .sco (PARSCALE) or .wge (WinGen) file formats. Thus, the number of IRTEQ input files ranges from two to four.

STUIRT is executed using one input file that contains item parameter estimates for the old form, item parameter estimates for the new form, a list of CIs, and optional keywords and specifications that primarily control the settings of the TCC methods. The STUIRT input file does not need to be in the same folder as the executable program file (unlike IRTEQ).

Creating the input files required to run POLYEQUATE is more laborious compared to STUIRT and IRTEQ. POLYEQUATE requires six input files: a control card containing the other five filenames, separate item parameter files for the old form and the rescaled new form, quadrature points and weights for the old form and the rescaled new form ability distributions, and a raw to scale score conversion table for the old form. The setup of these files is beyond the scope of this paper and is thoroughly described in the POLYEQUATE manual. However, the old and new form item parameter files and transformed new form ability quadrature distribution are provided in the STUIRT output file, and hence can be copied and pasted.

Program outputs

In general, IRTEQ and STUIRT/POLYEQUATE provide the same basic linking and equating output such as linking coefficients, transformed item parameter estimates, and old form equivalents for new

form scores. One advantage of using IRTEQ is that it provides various graphical results, whereas STUIRT/POLYEQUATE do not. However, STUIRT and POLYEQUATE provide additional results that are not given in IRTEQ (these are discussed below).

IRTEQ produces one main output file that contains the following: filenames for the inputted old and new form item parameter estimates, descriptive statistics for the CIs for the old and new forms, scale transformation coefficients (labeled as equating coefficients), file locations for the rescaled item parameter estimates and conversion tables, and a message indicating whether the program successfully finished. Two additional output files are created by default – one contains the rescaled item parameter estimates for the new form (.wgi or .par file) and the second contains number-correct old form equivalents from the TSE procedure (.con file). If requested by the user, IRTEQ can also save number-correct old form equivalents for all examinees (.ncs file). A nice feature of IRTEQ is that it automatically saves old form equivalents and rescaled item parameter estimates for each of the selected scale linking methods, in separate files. Whereas, in STUIRT, rescaled item parameter estimates and quadrature distributions are saved for only one linking method at a time. Furthermore, in IRTEQ, users can save TCCs for the CI set and total number-correct scores for the old and new form groups. Inside the GUI, users can also inspect CI abc-plots to check for any noticeable outliers.

For STUIRT, all linking output is contained in a single file, named by the user. The main output file starts out by first displaying the entire input control card. If users specify the **OP** keyword, options and defaults are displayed next for the settings that control the TCC methods. Descriptive summary statistics for the CIs are provided next, followed by linking coefficients from all four linking methods (provided by default). The last set of results contain the rescaled new form item parameter estimates, old form item parameter estimates, and rescaled new form ability quadrature distribution, which are each required as inputs to POLYEQUATE.

An advantage of STUIRT is that for the TCC methods, it provides initial solutions and detailed diagnostic information regarding the convergence criteria for the final solution (IRTEQ provides the minimized loss function value). The results provided by STUIRT also have more precision (i.e., five to six decimal places) than those provided by IRTEQ (typically three decimal places).

IRTEQ output specific to TSE has already been described above but primarily consists of a .con file with number-correct scores on the new form, the corresponding theta estimates, and corresponding unrounded old form number-correct score equivalents. POLYEQUATE output is contained in a single output file, and starts out by first displaying the complete set of input parameters, followed by IRT TSE and OSE results, followed by marginal distributions and summary statistics for raw scores, unrounded scale scores, and rounded scale scores for both forms.

Usability and documentation

The manuals for STUIRT and POLYEQUATE are very detailed and leave little to be questioned regarding the technical procedures and algorithms used. Because the STUIRT and POLYEQUATE syntax is program-specific, a good portion of each manual is dedicated to syntax descriptions. However, in general, the syntax is simple enough that learning it does not take a great deal of effort. An added benefit of using these programs is that many of the software examples correspond to examples in Kolen and Brennan (2014), thus providing an additional avenue for new users to check their understanding of the programs.

STUIRT and POLYEQUATE typically (but not always) provide error messages in the command-prompt window if syntax is unrecognizable. If the analysis is unsuccessful and no error message appears, users can open the output file to determine which set of inputs were last read in correctly to get clues on syntax errors. For example, if only the new form item parameter estimates appear in the output file, users are clued to inspect the next set of input parameters (e.g., old form item parameter estimates) for syntax errors. The inspection of input parameters in the output file is probably the most useful diagnostic tool users have for debugging their syntax in STUIRT and POLYEQUATE.

The GUI and point-and-click mode associated with IRTEQ make it very accessible to new users. The GUI contains a "Directions" dialog box that gives abbreviated instructions to the user. Furthermore, the necessary options are well-labeled and described with sufficient detail in the manual. The fact that IRTEQ can also be run with syntax and has a built-in simulation mode makes it appealing to researchers with a wide range of technical skills. Due to the somewhat limited functions available in IRTEQ, the manual is brief and omits many of the technical specifications detailed in the STUIRT and POLYEQUATE manuals. However, even some basic information, such as IRT model compatibility, is not provided in the manual. Luckily, some of these omitted details can be found in Han (2009b). Furthermore, some file extensions (i.e., .wgi and .wge; .lil and .lkl) are used interchangeably in the manual but are not treated as such by the program (correct file extensions include: .wgi or .par for item parameter estimates, .wge for person score file, and .lil or .txt for item linking file).

Debugging syntax errors and input files for IRTEQ requires a similar amount of effort as with STUIRT and POLYEQUATE. The error messages provided by IRTEQ give slightly more diagnostic details such as "error in 10th line" and "error found with the linking item list." However, even with these messages, additional detective work is typically required by the user.

Summary

In general, IRTEQ offers an all-inclusive IRT linking and equating functionality that is not found in STUIRT or POLYEQUATE. IRTEQ is a versatile program in the sense that it can be run using its GUI, with syntax, or with cue files that facilitate simulation-like studies. As a whole, IRTEQ requires fewer input files in comparison with STUIRT/POLYEQAUTE, which places less burden on the user. The linking and equating results reported are sufficient for general purposes but are not complete. The most noticeable drawbacks of IRTEQ are the limited number of technical options (including its ability to only conduct TSE) and its brief documentation. Overall, the program is a convenient tool for researchers and practitioners with various degrees of technical skills and psychometric knowledge.

By comparison, STUIRT and POLYEQUATE can be viewed as more technical programs that demand more from the user with respect to input parameters. Some advantages to using STUIRT and POLYEQUATE are (a) results are reported with greater precision (i.e., five decimal places), (b) users have more control over the TCC linking methods, (c) users have the option to use IRT OSE, (d) results for scale score equivalents are provided in addition to raw score equivalents, (e) additional equating results such as marginal distributions and summary statistics are provided, and (f) each has thorough documentation. The most noticeable drawbacks for STUIRT/POLYEQUATE are the lack of built-in graphing capabilities and the fact that file formats are not compatible with any other IRT software programs.

IRTEQ example

This section provides a basic example of using IRTEQ for IRT scale linking and equating under the CINEG design. A decision was made to exclude examples for STUIRT and POLYEQUATE based on fact that documentation for those programs includes detailed examples that would be redundant with the example presented here.

Data

Two groups of 8,000 examinees were drawn at random (without replacement) from a large-scale placement exam. Groups one and two (labels used by IRTEQ) are referred to as the old and new form groups, respectively. The full-length exam consisted of 65 MC and 4 free-response (FR) items (scored 0–5). Using the full-length exam, two pseudo-forms were created that each consisted of 35 MC and 2 FR items, of which, 10 MC and 1 FR items were common to both forms. MC and FR items were modeled using the 3PL and GR models, respectively. Item calibration was conducted in *flexMIRT* (Cai, 2017) using the default settings with the addition of a *c*-parameter prior distribution

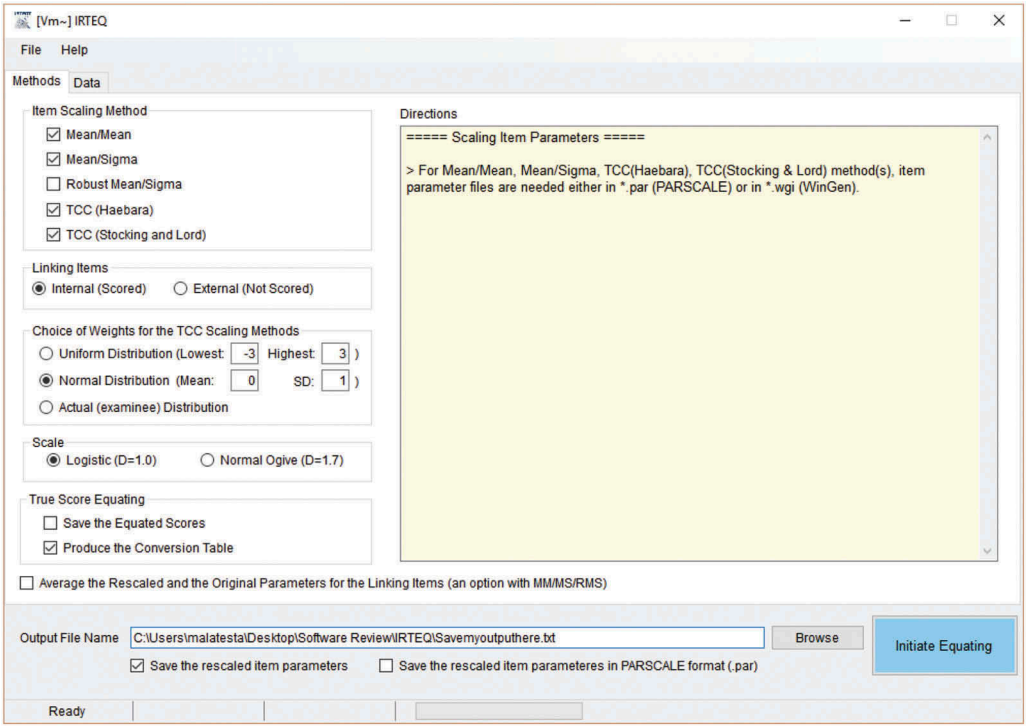

Figure 1. Methods tab of IRTEQ GUI.

(beta (1, 4)). Minimum and maximum number-correct scores were 0 and 45, respectively. Because the old and new form groups were sampled from a single population, scale linking was technically unnecessary as the same IRT calibration settings were used for both groups. However, both linking and equating were performed in this example for illustrative purposes.

IRTEQ GUI and input files

Upon opening the IRTEQ GUI, the Methods tab is displayed and allows users to select various IRT scale linking method(s) and various options related to the scaling methods and TSE method (see Figure 1).

In this example, all linking methods except RMS were selected and a standard normal proficiency distribution was used to weight the criterion functions of the TCC methods. Response probabilities and item parameter estimates are expressed on the logistic scale (D = 1.0) and CIs count toward examinees total number-correct scores. Under the TSE options, the conversion table that contains old form equivalents, is requested. Once all options are selected in the Methods tab, the Data tab is selected next and is used to read in the item parameter estimates for the old and new form groups (see Figure 2).

If the item parameter input files were read in correctly, item parameter estimates will appear in the GUI windows in neatly aligned columns. Item parameter files in the .wgi format must be tab delimited. An example item parameter input file for the old form can be found in Figure 3 (the same layout is used for the new form).

Next, users can choose to read in an item linking file (.lil) or manually enter the CIs into the "Linking item list" section of the Data tab. Users can verify whether the CIs and item parameter estimates were read in correctly by selecting the "Plot TCC" button. If read in correctly, the left-side

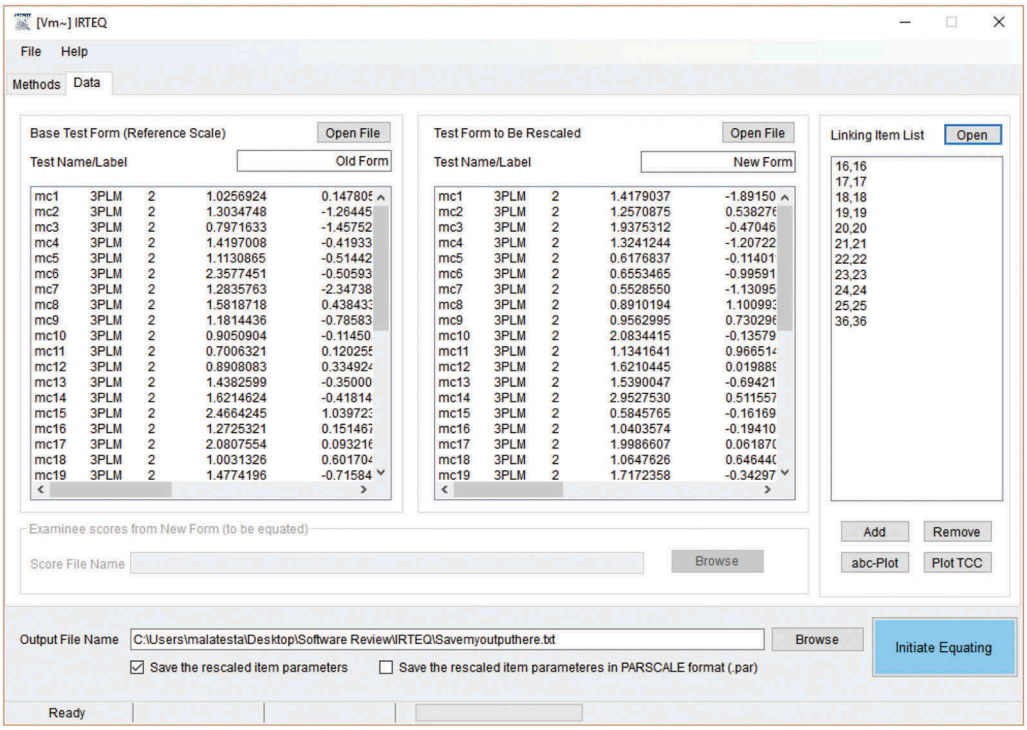

Figure 2. Data tab of IRTEQ GUI.

```
mc1  3PLM  2  1.0256924      0.1478056      0.2063826
mc2  3PLM  2  1.3034748     -1.2644552      0.2515225
mc3  3PLM  2  0.7971633     -1.4575278      0.2057610
.
.
.
fr1   GRM  6  2.0095877  -3.8329845  -2.3128350  -0.8643710  0.4245285  1.4565043
fr2   GRM  6  2.0602363  -3.2469946  -2.2535930  -0.9179987  0.3934798  1.4577879
```

Figure 3. Item parameter input file for the old form.

y-axis will range from 0 to the maximum number-correct score for the CIs and the right-side y-axis will range from 0 to the maximum number-correct score for the entire test. This added check is particularly useful because error messages for the linking item list are not always provided. The CI set can be inspected for outliers by clicking on the button labeled, "abc-Plot" (Figure 4). Any items with large deviations from the 45-degree reference line can be inspected further and dropped from the CI set if deemed appropriate by the user.

Once all fields are populated, the "Initiate Equating" button is used to run the linking and equating procedures.

Output files

Once IRTEQ finishes the linking and equating procedures, the main IRTEQ output file will appear. The output file is divided into five sections: (1) title and date and time of the analysis, (2) path and filenames for the item input files and the number of CIs, (3) descriptive statistics for the CI-set for the old and new form groups, (4) solutions for the slope and intercept for the selected scale linking

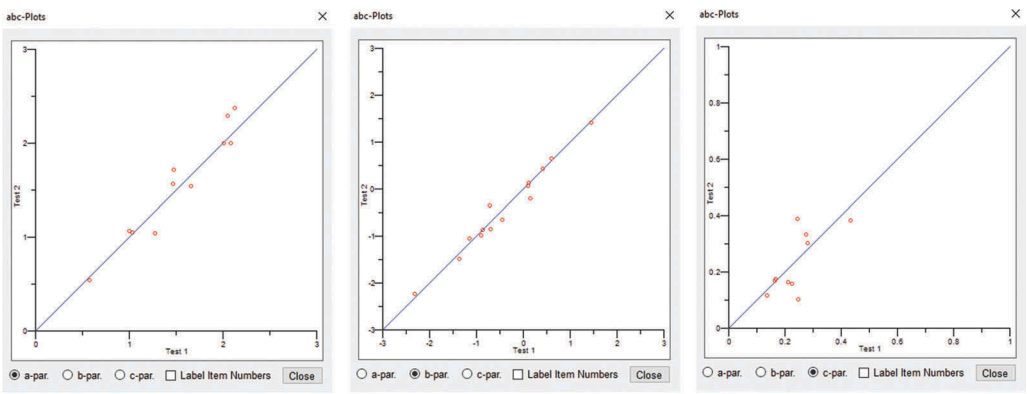

Figure 4. Example of abc-plots.

```
========================================================================

mean discrimination of linking items in old form:          1.523
mean difficulty of linking items in old form:              -0.630
standard deviation of difficulty parameters of linking items in old form: 1.275

mean discrimination of linking items in new form:          1.563
mean difficulty of linking items in new form:              -0.664
standard deviation of difficulty parameters of linking items in new form: 1.295

Correlation coeff. for discrimination of linking items:  0.966
Correlation coeff. for difficulty of linking items:      0.992
Correlation coeff. for guessing of linking items:        0.811

========================================================================

Equating coefficient A with Mean-Mean method:              1.026
Equating coefficient B with Mean-Mean method:              0.051

Equating coefficient A with Mean-Sigma method:             0.985
Equating coefficient B with Mean-Sigma method:             0.024

Equating coefficient A with TCC (Haebara) method:                1.02
Equating coefficient B with TCC (Haebara) method:                0.00
Minimized Loss Function Value with TCC (Haebara) method:         0.00064

Equating coefficient A with TCC (Stocking & Lord) method:        1.02
Equating coefficient B with TCC (Stocking & Lord) method:        0.01
Minimized Loss Function value with TCC (Stocking & Lord) method: 0.00005

========================================================================
```

Figure 5. Portion of main output file for IRTEQ.

methods, and (5) the location of the output files. Due to space constraints, only sections 3 and 4 of the main output file are displayed in Figure 5.

The descriptive statistics for the CI-set can be used to evaluate how similar the abilities are in the two groups of examinees. Because the old and new form groups were drawn from the same population, the descriptive statistics for the old and new form groups are very similar, suggesting the two groups represent similar ability levels. The values of the slope (i.e., A) and intercept (i.e.,

Test2	Theta	Test1 equivalent
0	*ind.	0.000
1	*ind.	1.009
2	*ind.	2.019
3	*ind.	3.028
4	*ind.	4.037
5	*ind.	5.046
6	*ind.	6.056
7	*ind.	7.065
8	-4.506	8.074
9	-3.676	9.287
10	-3.132	10.593
.		
.		
.		
41	1.883	41.625
42	2.173	42.558
43	2.583	43.473
44	3.309	44.324
45	*inf.	37.000

Figure 6. Old form number-correct score TSE equivalents for new form group.

B) for the various linking methods are close to identity (e.g., 1 and 0). This indicates that a very small transformation is needed to place the new form estimates onto the old form scale.

The primary TSE output file of interest is the IRTEQ conversion table that contains the number-correct scores for the new form (labeled "Test 2"), the unrounded equivalent scores on the old form scale (labeled "Test 1 equivalent"), and their corresponding theta estimate (not in that order). An abbreviated version of the conversion table can be seen in Figure 6.

Referring to Figure 6, an examinee in the new form group who received a raw score of 10 would be expected to receive a score of 11 on the old form. Put another way, a score of 10 on the new form is equivalent to a score of 11 on the old form. Because our example included the 3PL model, theta estimates below the sum of the *c*-parameters are undefined. As a result, an ad hoc method is required to obtain the old form equivalents for scores in this range. It appears that some form of linear interpolation is used in IRTEQ. Furthermore, theta estimates for perfect scores are set to positive infinity in IRTEQ and the corresponding old form equivalent is set equal to the number of test items (not to the maximum number-correct score). Therefore, to come up with an accurate old form equivalent score, users need to manually adjust the maximum old form equivalent score to match that of the maximum number-correct score.

A useful feature of IRTEQ is the TCC plot that allows the user to compare the TCCs for the new form (both before and after rescaling) to the old form. Users can choose whether to include TCCs at the test-level, for only the CI-set, or for both (see Figure 7). In our example, the new form test-level TCCs are nearly identical before and after rescaling and therefore, only the rescaled one is included. This last observation is consistent with the fact that the estimated linking coefficients were close to identity.

In summary, this example entailed IRT linking and equating of two pseudo-test forms under the CINEG design. The old and new form examinee groups were sampled from a single test administration and had very similar ability levels. This similarity was evident in the values of the linking coefficients and the transformed TCCs. After the new form estimates were rescaled to the old form scale, IRT TSE was conducted to adjust for differences in form difficulty. In practice, old form equivalents for examinees who took the new form would be reported and could be directly compared to scores obtained by examinees who took the old form.

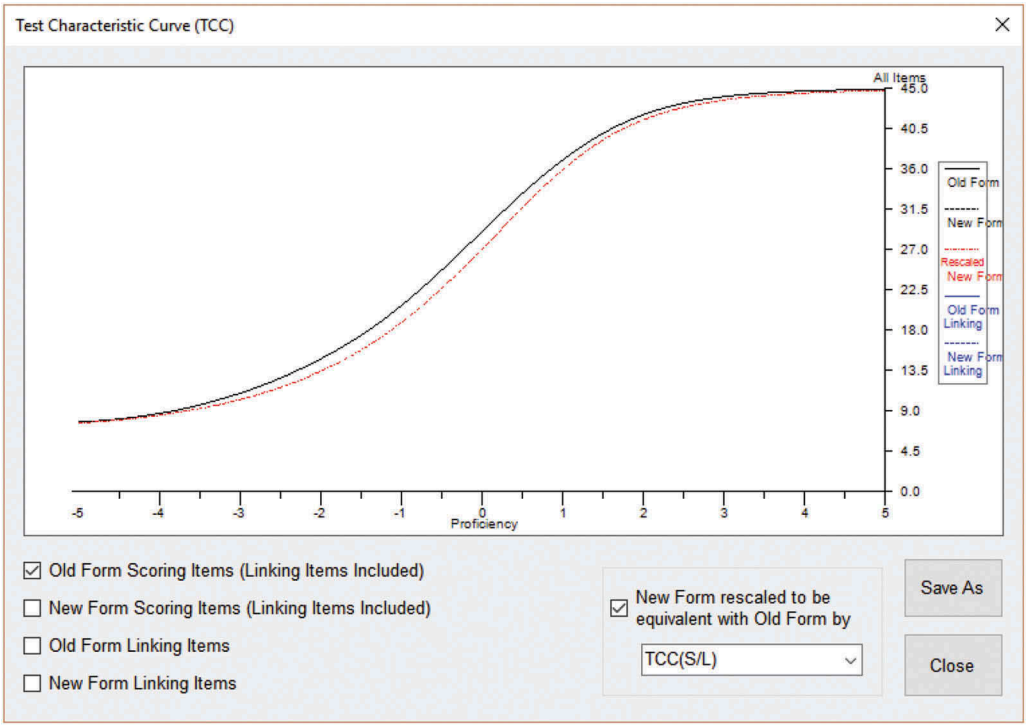

Figure 7. TCC plot after executing IRTEQ.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51. doi:10.1007/BF02291411

Bolt, D. M. (1999). Evaluating the effects of multidimensionality on IRT true-score equating. *Applied Measurement in Education*, *12*, 383–407. doi:10.1207/S15324818AME1204_4

Cai, L. (2017). *flexMIRT®: Flexible multilevel multidimensional item analysis and test scoring (version 3.51) [Computer software]*. Chapel Hill, NC: Vector Psychometric Group, LLC.

Camilli, G., Wang, M., & Fesq, J. (1995). The effects of dimensionality on equating the law school admission test. *Journal of Educational Measurement*, *32*, 79–96. doi:10.1111/jedm.1995.32.issue-1

Cook, L. L., Dorans, N. J., Eignor, D. R., & Petersen, N. S. (1985). *An assessment of the relationship between the assumption of unidimensionality and the quality of IRT true-score equating* (ETS Research Report 85-30). Princeton, NJ: Educational Testing Services.

de Champlain, A. F. (1996). The effect of multidimensionality of IRT true-score equating for subgroups of examinees. *Journal of Educational Measurement*, *33*, 181–201. doi:10.1111/j.1745-3984.1996.tb00488.x

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement*, *22*, 249–262. doi:10.1111/jedm.1985.22.issue-4

Dorans, N. J., Pommerich, M., & Holland, P. (Eds). (2007). *Linking and aligning scores and scales*. New York, NY: Springer.

Haebara, T. (1980). Equating logistic proficiency scales by a weighted least squares method. *Japanese Psychological Research*, *22*, 144–149. doi:10.4992/psycholres1954.22.144

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

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement*, *31*(5), 457–459. doi:10.1177/0146621607299271

Han, K. T. (2009a). *IRTEQ: IRT equating software*. Amherst, MA: The University of Massachusetts-Amherst. Retrieved from https://www.umass.edu/remp/main_software.html
Han, K. T. (2009b). IRTEQ: Windows application that implements item response theory scaling and equating. *Applied Psychological Measurement*, *33*(6), 491–493. doi:10.1177/0146621608319513
Kim, S., & Kolen, M. J. (2004). *STUIRT: A computer program for scale transformation under unidimensional item response theory models*. Iowa City, IA: Iowa Testing Programs, The University of Iowa. Retrieved from https://education.uiowa.edu/centers/casma
Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement*, *18*(1), 1–11. doi:10.1111/jedm.1981.18.issue-1
Kolen, M. J. (2004). *POLYEQUATE (Windows console version)*. Iowa City, IA: The University of Iowa. Retrieved from https://education.uiowa.edu/centers/casma
Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York, NY: Springer.
Kolen, M. J., & Lee, W. Eds., (2011). *Mixed-format tests: Psychometric properties with a primary focus on equating*.Vol. 1 (CASMA Monograph No. 2.1.). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.
Kolen, M. J., & Lee, W. Eds., (2012). *Mixed-format tests: Psychometric properties with a primary focus on equating*.Vol. 2 (CASMA Monograph No. 2.2.). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.
Kolen, M. J., & Lee, W. Eds., (2014). *Mixed-format tests: Psychometric properties with a primary focus on equating*.Vol. 3 (CASMA Monograph No. 2.3.). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.
Kolen, M. J., & Lee, W. Eds., (2016). *Mixed-format tests: Psychometric properties with a primary focus on equating*.Vol. 4 (CASMA Monograph No. 2.4.). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.
Kolen, M. J., & Lee, W. Eds., (2018). *Mixed-format tests: Psychometric properties with a primary focus on equating*.Vol. 5 (CASMA Monograph No. 2.5.). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.
Lee, W., & Brossman, B. G. (2012). Observed score equating for mixed-format tests using a simple-structure multidimensional IRT framework. M. J. Kolen & W. Lee Eds., *Mixed-format tests: Psychometric properties with a primary focus on equating* (Vol. 2, pp. 115–142) (CASMA Monograph No. 2.2.). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.
Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement*, *5*, 159–173. doi:10.1177/014662168100500202
Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, *17*, 179–193. doi:10.1111/jedm.1980.17.issue-3
Marco, G. L. (1977). Item characteristic curve solution to three intractable testing problems. *Journal of Educational Measurement*, *14*, 139–160. doi:10.1111/j.1745-3984.1977.tb00033.x
Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101–121). New York, NY: Springer.
Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176. doi:10.1177/014662169201600206
Muraki, E., & Bock, R. D. (2003). PARSCALE (Version 4.1) [Computer Program]. Mooresville, IN: Scientific Software.
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society.
Stocking, M., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 207–210. doi:10.1177/014662168300700208
Tate, R. (2000). Performance of a proposed method for the linking of mixed format tests with constructed response and multiple choice items. *Journal of Educational Measurement*, *37*, 329–346. doi:10.1111/jedm.2000.37.issue-4
Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, *49*, 501–519. doi:10.1007/BF02302588
Yao, L., & Boughton, K. (2009). Multidimensional linking for tests with mixed item types. *Journal of Educational Measurement*, *46*, 177–197. doi:10.1111/jedm.2009.46.issue-2
Yen, W. M. (1984). Effects of local dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*, 125–145. doi:10.1177/014662168400800201