FlexCCT v2.5: Help File

Contents

Copyright

Introduction

FlexCCT (Flexible Cultural Consensus Theory) provides a flexible, easy to use toolkit for analyzing ratings data. FlexCCT utilizes techniques from the academic discipline of "Cultural Consensus Theory" (CCT) and from machine learning. CCT can be thought of as a data pooling or aggregation technique that uses the "wisdom of the crowd" to both create better ratings and characterize raters with respect to competency and bias. This document will give an overview of the features present in the FlexCCT package, along with illustrated examples.

For technical details of the methodology please see the following papers:

France, S.L., Batchelder, W. H. (2015). A Maximum Likelihood Item Easiness Model for Test Theory Without An Answer Key, Educational and Psychological Measurement, 75(1), 57-77.
<preprint> <reprint>

France, S.L., Batchelder, W. H. (2014). Unsupervised Consensus Analysis for On-line Review and Questionnaire Data, Information Sciences, 281, 241-257. (PN016)
<preprint> <reprint>

FlexCCT also implements minimum residual factor analysis, which is described in the following paper:

Comrey, A. L. (1962). The minimum residual method of factor analysis. *Psychological Reports, 11*(1), 15-18.

Main Application

Double clicking on the FlexCCT.exe icon will load the main FlexCCT application. This will bring up the main FlexCCT form, which is shown below.

The form consists of the following sections:

Data Input

This frame defines the input file details. The input file is a csv file with subjects as rows and the items to be rated as columns. The data file can optionally contain row names and/or column names. Each entry to the file is a rating. FlexCCT allows for missing rating values, which default to -1. Each item can have ratings for multiple traits. For example, an essay grading item could have attributes for grammar, consistency, exposition, and content. Consider a situation with 5 users, 6 items and 2 traits per item. Let R(i,j,a) signify the rating for the ith user rating the jth item on the ath trait. The file configuration is given below.

R(1,1,1),R(1,1,2), R(1,2,1),R(1,2,2),R(1,3,1),R(1,3,2), R(1,4,1),R(1,4,2), R(1,5,1),R(1,5,2), R(1,6,1),R(1,6,2)

R(2,1,1),R(2,1,2), R(2,2,1),R(2,2,2),R(2,3,1),R(2,3,2), R(2,4,1),R(2,4,2), R(2,5,1),R(2,5,2), R(2,6,1),R(2,6,2)

R(3,1,1),R(3,1,2), R(3,2,1),R(3,2,2),R(3,3,1),R(3,3,2), R(3,4,1),R(3,4,2), R(3,5,1),R(3,5,2), R(3,6,1),R(3,6,2)

R(4,1,1),R(4,1,2), R(4,2,1),R(4,2,2),R(4,3,1),R(4,3,2), R(4,4,1),R(4,4,2), R(4,5,1),R(4,5,2), R(4,6,1),R(4,6,2)

R(5,1,1),R(5,1,2), R(5,2,1),R(5,2,2),R(5,3,1),R(5,3,2), R(5,4,1),R(5,4,2), R(5,5,1),R(5,5,2), R(5,6,1),R(5,6,2)

Column/row names are given by adding a column/row before the first ratings. For example, the following figure gives the first four rows of a file with both raters and items. The raters are analysts rating NFL Teams' drafts. The items are the actual NFL teams.

```
,Arizona,Atlanta,Baltimore,Buffalo,Carolina,Chicago,Cincinnati,Cleveland,Dallas,Denver,Detroit,Green Bay,Hc
ESPN's MEL KIPER,8,6,9,8,8,7,11,8,7,7,9,9,9,11,6,7,6,9,8,5,7,6,5,12,9,10,7,5,8,12,7,7
CBS SPORTS,7,6,10,9,10,6,13,5,10,7,7,9,9,12,8,6,10,10,8,1,9,5,3,10,12,8,10,5,7,9,8,8
FOX SPORTS,10,6,9,9,9,5,12,6,10,8,8,9,8,11,6,10,6,8,12,3,9,6,3,9,13,9,7,9,12,10,6,11
NBC SPORTS,9,5,10,9,6,10,12,9,9,6,10,11,9,13,3,8,9,7,10,3,12,8,3,10,12,10,7,8,10,12,7,11
```

Data File

Click the Browse button and choose the .csv data file using the standard Microsoft Windows file dialog box.

Number of Traits

Define the number of traits used in the file. This defaults to 1. The number of columns must be divisible by the number of traits. For example, 60 columns and 6 traits would indicate a dataset that has 10 items with 6 traits per item

Row Headers

The FlexCCT import routine will attempt to guess whether or not row headers are included in the data file, based upon the data type of the first column. However users can explicitly state that row headers are included. This feature is useful if the row headers are numeric.

Column Headers

The FlexCCT import routine will attempt to guess whether or not column headers are included in the data file, based upon the data type of the first row. However users can explicitly state that column headers are included. This feature is useful if the column headers are numeric.

Ground Truth

If the first row of numeric data contains true/correct values for the data, i.e., a "ground truth", then select this checkbox. A ground truth can be incorporated when there are actual true values for the data; e.g., users estimate distance data and the true distances are actually available.

Preprocessing

The preprocessing frame allows users to specify how the raw imported data is processed. There are two preprocessing options.

Data Combination

The CCT methods implemented in this package assume a single trait. If the data have multiple traits then there are two options for combining the traits into a single meta-trait:

- **Add Traits:** This option utilizes the assumptions of classical test theory. Add the values of the traits together. For example, if an essay scored 3 out of 6 for grammar, 4 out of 6 for consistency, 3 out of 6 for exposition, and 4 out of 6 content then the overall score for the essay would be 14/24.
- **Correspondence Analysis:** The correspondence analysis option assumes that the individual traits are measured on an ordinal scale. The method of multiple correspondence analysis is then used to create a continuous scale measure that captures as much structure as possible from the ordinal variables.

Data Transformation

CCT assumes that all items are measured on the same scale. If this is not true then this will lead to a violation of the CCT error assumptions and a possibly incorrect analysis. One way of getting around this problem is to rescale the data. The options are as follows:

- **None:** There is no transformation. The raw scores are used. This option should be used if all items are measured on the same scale and no error heterogeneity is suspected.
- **Standardization:** The scores are standardized by subtracting the item means and dividing by the item standard deviations. This is an option when items are measured on multiple scales.
- **Range Scale:** The scores are range scaled by putting all scores into the range [0,1]. This is achieved by subtracting the minimum item score from each score and then dividing by the range (maximum – minimum). This is an option when items are measured on multiple scales.
- **Log Transformation:** The log transformation is utilized when items are measured on the same scale, but there is possible error heterogeneity. An example of this is a situation where raters are estimating distances. All items are measured on the same scale, but based on the Weber-Fechner psychophysical law, large distances are liable to have larger error variances than do smaller differences. This can be rectified by using a log transformation on the data.

Model Settings

The model settings frame determines the CCT model settings. FlexCCT implements a range of models and model features.

Base Model

The Base Model dialog determines the basic CCT model used. Models range from simple aggregation models to complex likelihood models incorporating item easiness/difficulty. The options are as follows:

- **Simple Average:** A baseline simple averaging model. It calculates the mean of all of the ratings. Assumes even competencies.
- **Factor Analysis:** Implements a principal axes factor analysis, as per Comrey (1966).
- **ML Model**: The basic maximum likelihood model. Outputs a set of user competencies **d**, where $\sigma_i^2(\varepsilon_{ik}) = 1/d_i$, along with a competency weighted item answer key **z**.
- **IE Multiply**: The multiplicative item easiness model. Assumes that for rater i and item k, the item competencies are affected by multiplicative scalar item easiness values, so that the error variance $\sigma_{ik}^2(\varepsilon_{ik}) = 1/(d_i\beta_k)$.
- **IE Add**: The additive item easiness model. Assumes that for rater i and item k, the item competencies are affected by additive scalar item easiness values, so that the error variance $\sigma_{ik}^2(\varepsilon_{ik}) = 1/(d_i+\beta_k)$.

Bias Type

Biases can be defined for each model. Both additive and multiplicative biases can be added. The biases are defined, so that for user i rating item k, the item scores are defined as $x_{ik} = b_{Mi}z_k + b_{Ai}$. Here b_{Mi} is the multiplicative bias for item i and b_{Ai} is an additive bias for item i. The options are as follows:

- **No Bias:** No biases are included in the model. Thus, for each i, $b_{Mi} = 1$ and $b_{Ai} = 0$.
- **Additive Bias:** Only additive biases are included in the model. Thus for each i, $b_{Mi} = 1$.
- **Multiplicative Bias:** Only multiplicative biases are included in the model. Thus for each i, $b_{Ai} = 0$.

- **Add & Mult Bias:** Both additive and multiplicative biases are included in the model.

All four bias types can be utilized with every model except for the principal axes factor analysis model, which does not allow bias to be modeled.

Missing Value Indicator

The missing value indicator allows the specification of the value used for missing data. The default is "-1". If the ratings values can include negative numbers then this number should be changed to a number outside of the range of the ratings.

Optimization Settings

The optimization settings frame allows the specification of the optimization procedure for the implemented FlexCCT model.

Optimization Method

This dropdown list specifies the method by which the model is optimized (i.e., the likelihood function is optimized). **The default method is the gradient method and this method is to be recommended for non-technical users**.

Each option has advantages and disadvantages. A technical explanation is given as follows:

- **Gradient:** This procedure utilizes first order derivatives of the model likelihood function to maximize the likelihood using a gradient descent procedure. All models specified in the Model Setting frame and all advanced settings can be utilized with this procedure.
- **Derivative Free:** This procedure utilizes a gradient descent procedure to maximize the likelihood function, but does not utilize function derivatives. All models specified in the Model Setting frame and all advanced settings can be utilized with this procedure. This method is significantly slower than the gradient descent procedure.
- **Fixed Point:** This procedure alternatively optimizes the first order conditions (i.e., the derivatives set to 0) for the different variable sets (**d**, **z**, **b**$_A$, **b**$_M$, and **β**). This is by far the fastest optimization option, but cannot implement every option. The fixed point procedure cannot implement the "IE Add" model type or the "Fix Average" identifiability options.

Converge

The convergence criterion for the optimization algorithm. The algorithm completes and is considered to "converge" when at two subsequent iterations the change in objective function is less than this amount. This value defaults to "1e-6" or 1×10^{-6}.

d Max

The maximum value for the user competency. Setting the maximum value for the competency prevents competencies from going to infinity (i.e., to an asymptote of the likelihood function).

IE Max

The maximum value for item easiness. As per "d Max", setting the maximum value for the item easiness prevents the item easiness values from going to infinity.

Form Buttons

The main form has five buttons, which are situated at the top right hand side of the form. The functionality for each button is described below.

Run

This option is only available if a csv file has been specified in the Data Input frame. Load in the data and run the specified FlexCCT model. Run the FlexCCT optimization procedure and output the model parameters and model fit statistics in the CCT Results Output form.

Default

Return all of the form fields to their default settings.

Advanced

Specify advanced features. The advanced features include the ability to specify item identifiability options, define confidence intervals, and set the number of cultures/clusters.

Scree

This option is only available if a csv file has been specified in the Data Input frame. Shows the scree plot form, which can be used to guess the number of clusters/cultures present in the data.

Validation

Specify validation settings. The validation features give the ability to evaluate a model either by comparing the answer to the ground truth or by performing hold-k-out validation.

The Scree Plot
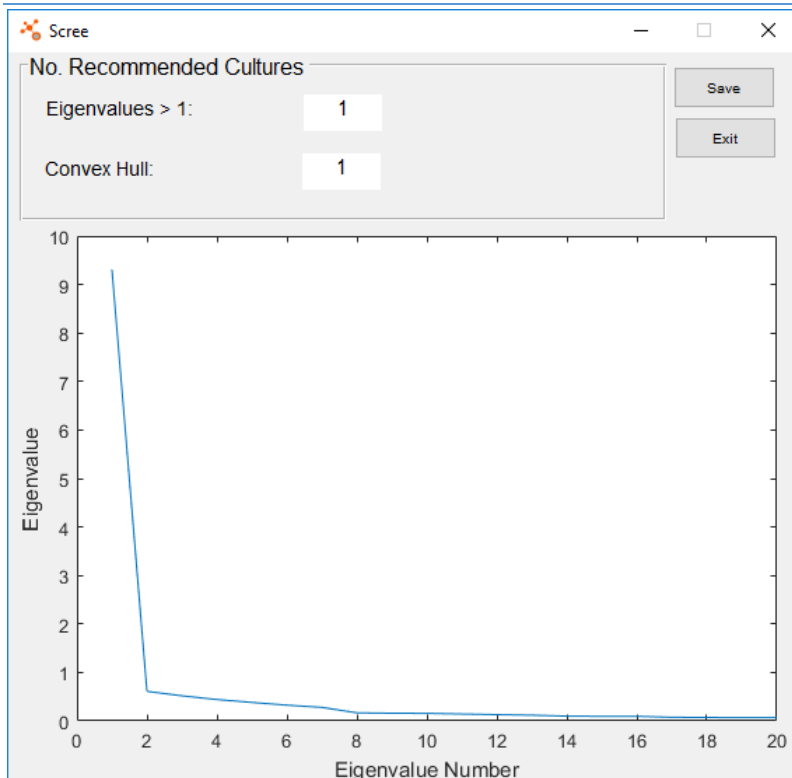
The Scree plot form utilizes an eigenvalue decomposition from Comrey's minimum residual factor analysis of the inter-rater correlation matrix to guess the number of clusters or cultures in the data. The rationale behind the use of this technique is that if there is one culture and all the raters have a similar pattern of answers then the correlation matrix should show a one factor structure. Multiple cultures indicates different preferences, e.g., when rating movies, some users like horror movies and hate romance movies, but other users have opposite preferences. In the main graph for the form, each eigenvector number (1st, 2nd, 3rd etc.) is plotted against its actual value. This plot can be used to infer the number of eigenvectors using the criteria in the No. Recommended Cultures criteria:

No. Recommended Cultures

EigenValues > 1

This heuristic states that eigenvalues are taken as significant if they are greater than 1. This is a common criterion in factor analysis. In the preceeding example, only the 1st eigenvalue is greater than 1, so the number of recommended cultures is 1.

Convex Hull
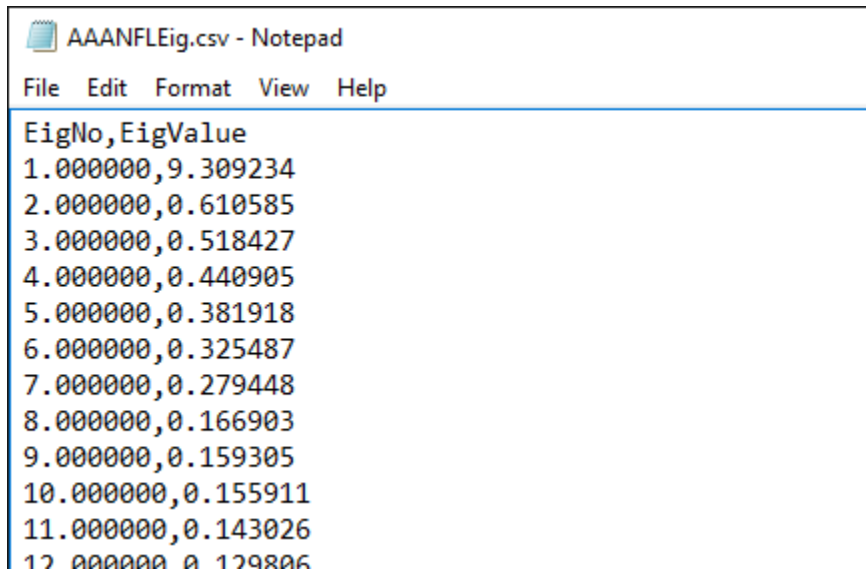
Here, the "convex hull" of the graph is used to determine the number of cultures. The convex hull is the point at which the graph of the eigenvalues goes from making an angle greater than 45° with the origin to making an angle less than 45° with the origin. The chosen eigenvalue number is the one before the transition from >45° to <45°. In the preceding example, the transition occurs at the 2nd eigenvalue, so the number of recommended cultures is 1.

Form Buttons

The main form has two buttons, which are situated at the top right hand side of the form. The functionality for each button is described below.

Save

Allows the saving of the eigenvalue information that is plotted in the graph into a csv file. Brings up a standard Microsoft Windows file dialog box. If a file name is chosen and OK is checked then the file will be saved. An example of the format used is given below.

```
AAANFLEig.csv - Notepad

File   Edit   Format   View   Help

EigNo,EigValue
1.000000,9.309234
2.000000,0.610585
3.000000,0.518427
4.000000,0.440905
5.000000,0.381918
6.000000,0.325487
7.000000,0.279448
8.000000,0.166903
9.000000,0.159305
10.000000,0.155911
11.000000,0.143026
12.000000,0.129806
```

Exit

Exits the form and returns to the main GUI form.

Advanced Features

The Advanced Features form gives a range of features to help customize the FlexCCT analysis. Features include the ability to specify item identifiability options, define confidence intervals, and set the number of cultures/clusters.

Variable Identifiability

If all the variables in some of the advanced FlexCCT models are fit simultaneously, then the model may be over-identified, i.e., multiple sets of parameters with different values could give the same solution criteria. To prevent this problem, variables can be restricted, so as to give a grounded model or a model where different types of parameters have different precedencies. Variable identification options are given for the answer key (**z**), the rater competencies (**d**), the additive biases (**b**$_A$) and the multiplicative biases (**b**$_M$). The options are given as follows:

Fix z:

This variable determines how to fix **z**. The options are given below. The default value is "Partial Fix".

- **No Fix:** Does not fix the variables. Allows the variables to be optimized freely.
- **Full Fix:** Fixes the **z** values to be the values from a model with no bias or item difficultly parameters. This essentially fits a sub-model, where **b**$_A$, **b**$_M$, and **β** parameters are then calculated as residual measures of fit.
- **Partial Fix:** Fixes a single value in **z** from the sub model given above. The value of **z** to be chosen is the one with the lowest average error. This grounds the model and prevents **z** values and bias values from working against each other (e.g., increase each value in **z** by 1 and decrease each value in **b**$_A$ by 1).

- **Fix Average = Average Mean:** Uses constrained optimization to fit the average value of **z** to the average of **z** calculated when taking a simple arithmetic mean. Again, this grounds the value of **z**.

Fix d:

This variable determines how to fix **d**. The options are given below. The default value is "Partial Fix".

- **No Fix:** Does not fix the variables. Allows the variables to be optimized freely.
- **Full Fix:** Fixes the **d** values to be the values from a model with no bias or item difficultly parameters. This essentially fits a sub-model, where **b**$_A$, **b**$_M$, and **β** parameters are residual measures of fit.
- **Partial Fix:** Fixes a single value in **d** from the sub model given above. The value of d_i to be chosen is the one closest to the average. As competency values are only important in a relative sense (e.g., via a competency weighted means), this prevents the competencies becoming arbitrarily too low or high.
- **Fix Average = 1:** Uses constrained optimization to fit the average value of **d** to 1, grounding the competencies.

Fix b(add):

Both additive biases and multiplicative biases are assigned to individual raters. Identifiability options can be set to give precedence to fitting one type of bias over the other. This option is available if on the Bias Type dropdown menu, either "Additive Bias" or "Add and Mult Bias" are selected.

- **No Fix:** Does not fix the variables. Allows the variables to be optimized freely.
- **Full Fix:** Fixes the **b**$_A$ values to be the values from a model with only additive bias fit. The multiplicative bias variables **b**$_M$ are residual measures of fit.
- **Partial Fix:** Fixes a single **b**$_A$ value from model given above. The value of b_{Ai} to be chosen is the one closest to the average. This prevents the additive biases from becoming arbitrarily too low or high.
- **Fix Average = 0:** Uses constrained optimization to fit the average value of **b**$_A$ to 0, grounding the additive bias around 0, i.e., no effective multiplicative bias.

Fix b(mult):

As above. This option is available if for the Bias Type dropdown menu either "Multiplicative Bias" or "Add and Mult Bias" are selected.

- **No Fix:** Does not fix the variables. Allows the variables to be optimized freely.
- **Full Fix:** Fixes the **b**$_M$ values to be the values from a model with only multiplicative bias fit. The additive bias variables **b**$_A$ are residual measures of fit.
- **Partial Fix:** Fixes a single **b**$_M$ value from model given above. The value of **b**$_M$ to be chosen is the one closest to the average. This prevents the additive biases from becoming arbitrarily too low or high.
- **Fix Average = 1:** Uses constrained optimization to fit the average value of **b**$_M$ to 1, grounding the multiplicative bias around 1, i.e. no effective multiplicative bias.

As the partial fix and full fix options for Fix b(add) and Fix b(mult) implicitly define a fitting hierarchy between **b**$_A$ and **b**$_M$, these options can only be chosen for either Fix b(add) and Fix b(mult) but not for both.

Confidence Intervals

The confidence intervals option allows $(1 - \alpha) \times 100\%$ confidence intervals to be defined for each set of parameters included in the CCT model. The confidence intervals are calculated using a bootstrapping or

jackknifing procedure. For a detailed overview of the implemented bootstrap procedures see the following link.

Singh, K., & Xie, M. (2008). Bootstrap: a statistical method. *Unpublished manuscript, Rutgers University, USA. Retrieved from http://www.stat.rutgers.edu/home/mxie/RCPapers/bootstrap.pdf* .

Confidence intervals can be added for variable sets **d**, **z**, **b**$_A$, **b**$_M$, and **β** by selecting the appropriate checkboxes.

- z – CIs for the solution answer key **z**.
- d **–** CIs for the rater competencies **d**.
- b(add) – CIs for the additive biases **b**$_A$.
- b(mult) – CIs for the multiplicative biases **b**$_M$.
- IE – CIs for the item easiness values **β**.

Alpha

Determines the value of α to create a $(1-\alpha)\times100\%$ confidence interval. For example, setting $\alpha = 0.05$ creates a 95% confidence interval.

No. Resamples

Bootstrapping takes samples witH replacement from the original data sample. This field defines the number of resamples. Jack-knifing works by taking each item out of the original dataset and then using the parameter estimates from these "take one out" data samples to form a confidence interval. If No. Resamples is greater or equal to the number of items in the sample (*n*) then all *n* take one out samples are processed. If No. Resamples is less than *n* then then "No. Resamples" value is used, with the items to be removed selected randomly.

Interval Type

The Interval Type dropdown menu determines the type of confidence interval that is to be created. The first three options create bootstrap confidence intervals. The last option creates a jackknife confidence interval. The confidence intervals for each parameter are defined below. Here, *N* is the number of resamples, α defines a $(1-\alpha)\times100\%$ confidence interval, $\hat{\theta}$ is the point estimate of the parameter, and $\{\theta_1^*, \cdots, \theta_N^*\}$ are the ordered parameter values taken from the *N* resamples.

- **Percentile Bootstrap:** The basic percentile bootstrap, where the confidence interval for parameter θ is calculated as $\left[\theta_{N\left(\frac{\alpha}{2}\right)}^*, \theta_{N\left(1-\frac{\alpha}{2}\right)}^*\right]$.
- **Centered Percentile Bootstrap**: The centered percentile bootstrap, where the confidence interval for parameter θ is calculated as $\left[2\hat{\theta} - \theta_{N\left(1-\frac{\alpha}{2}\right)}^*, 2\hat{\theta} - \theta_{N\left(\frac{\alpha}{2}\right)}^*\right]$.
- **Bootstrap-t:** The *t*-distribution bootstrap, which utilizes a *t*-distribution confidence interval where $t^* = \left(\theta^* - \hat{\theta}\right)/SE(\theta^*)$.
- **Jackknife:** First calculates the standard error for the confidence interval using a series of hold-one out samples and then uses the standard error to create a bias-corrected confidence interval.

Replace Parameters with CI Mean

If this option is checked, the parameters from the initial estimation are replaced by the midpoints of the confidence intervals. This ensures that the confidence intervals are symmetric about the reported parameter estimates.

Multiple Culture Options

FlexCCT can simultaneously split raters in cultures and estimate within culture rater and item parameters (i.e., **d**, **z**, **b**$_A$, **b**$_M$, and **β**). Raters are split to give the maximum within culture similarity. The number of cultures is specified by the users. The Scree Plot form can be used to give guidance on the number of clusters.

No Cultures

The number of cultures. Defaults to 1. In this case all raters are assigned to the one default cluster.

No. Clustering Runs

The number of clustering runs. FlexCCT utilizes a k-means like procedure and is not guaranteed to find a globally optimal solution. Multiple cluster runs increase the chance of finding globally optimal solutions. Defaults to 10.

Incorporate Bias into Clustering

If this option is checked, bias is accounted for when clustering raters. First, the CCT model is estimated with the assumption of an individual culture. Biases are calculated (depending on the bias settings on the main form) and bias corrected scores (subtracting additive bias and then dividing by multiplicative bias) are calculated for each user. The bias corrected scores are then used to cluster the raters into cultures.

Form Buttons

The main form has two buttons, which are situated at the top right hand side of the form. The functionality for each button is described below.

Save

Save the currently selected advanced settings.

Default

Return all of the advanced setting form fields to their default settings.

Validation

The Validation form allows users to specify a range of hold-k out validation options, validate a consensus solution against a ground truth, and report a range of different evaluation metrics. Validation is performed if at least one validation test is selected (either holdout or ground truth) and at least one criteria is selected.

Internal Holdout Validation

Internal holdout validation allows a user to repeatedly hold out a number (k) of ratings and calculate the model from the remaining data. The difference between the model parameters derived from the main data and the models parameters derived from the remaining data is used to then evaluate the stability of the model. As a rule, more strongly identified models, with more restrictions, perform better with holdout validation.

Parameters can be selected for internal holdout validation for variable sets **d**, **z**, **b**$_A$, **b**$_M$, and **β**, by selecting the appropriate checkboxes:

- z – Validation for the solution answer key **z**.
- d – Validation for the rater competencies **d**.
- b(add) – Validation for the additive biases **b**$_A$.
- b(mult) – Validation for the multiplicative biases **b**$_M$.
- IE – Validation for the item easiness values **β**.
- XPred – Validation for the overall ratings. In this case, the values of x_{ij} generated by the model for the holdout data are compared with the actual x_{ij} values. This allows for the calculation of a holdout prediction log-likelihood statistic and consensus adjusted prediction statistics for each holdout criterion.

No Raters
Gives the number of raters for the currently selected data file.

No Items
Gives the number of items to be rated for the currently selected data file.

No Ratings

Gives the number of ratings present in the currently selected data file.

Holdout Size

The number of ratings to holdout for each run of the holdout samples. If this is one then LOOCV (leave one-out cross validation is performed. Note: For a one-cluster solution, one can implement *n*-fold cross validation by deciding *n* and then setting the holdout size to be the number of ratings divided by *n*.

Maximum Runs

By default the number of runs is $NoRuns = \lfloor NoRatings/Holdoutsize \rfloor$. However, large datasets and small values of n, can lead to long run times. Thus, the maximum number of runs can be set. Here, ratings are placed in random order and holdout sets of n items are taken until the maximum of Maximum Runs and $NoRuns$ is reached.

External Holdout Validation

External holdout validation exists for the answer key **z**. If "Ground Truth on First Row" is selected on the main screen, then a ground truth answer key, **z*** will be available and the "Compare to Ground Truth" checkbox will be enabled.

Compare To Ground Truth

If this option is selected then the ground truth **z*** is compared to the consensus answer key **z**.

Criteria

The checkboxes in the criteria frame specify which measures are reported for both the internal and external holdout validation procedures. At least one of the criteria must be selected in order for the validation procedures to be run. Let **p*** be a vector of length *n* of desired parameter values (the solution of the full data for internal validation or the ground truth for ground truth validation and **p** be the set of comparison values (the holdout values for internal validation and the consensus solution for ground truth validation). The individual criteria selections are described below.

- MSE –Mean squared error, where $MSE = \frac{1}{n}\sum_{i=1}^{n}(p_i^* - p_i)^2$.

- MSPE **–** Mean squared percentage error, where $MSPE = \frac{100}{n}\sum_{i=1}^{n}\left(\frac{p_i^* - p_i}{p_i^*}\right)^2$

- MAE – Mean absolute error, where $MAE = \frac{1}{n}\sum_{i=1}^{n}|p_i^* - p_i|$

- MAPE – Mean absolute percentage error, where $MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{p_i^* - p_i}{p_i^*}\right|$

- Corr – Pearson correlation, where $Corr = \frac{\sqrt{\sum_{i=1}^{n}(p_i^* - \bar{p}_i^*)(p_i - \bar{p}_i)}}{\sqrt{\sum_{i=1}^{n}(p_i^* - \bar{p}_i^*)^2}\sqrt{\sum_{i=1}^{n}(p_i - \bar{p}_i)^2}}$. Note that here, the closer the

    correlation to 1, the stronger the result.

CCT Results Output

The CCT Results Output form gives a listing of results. The columns returned are based upon the previously described settings.

	Culture	LL;LLPred;RB;RCCT	zMSE;XPMSE	d	dCILower	dCIUpp
1	1	-895.9256	4.1654e-04	0.8527	0.7724	0.9
2	NaN	-1.0536e+03	1.6288	0.4710	0.4135	0.5
3	NaN	0.5448	NaN	0.4149	0.3612	0.4
4	NaN	0.6665	NaN	0.3158	0.2701	0.3
5	NaN	NaN	NaN	1.3970	1.2793	1.5
6	NaN	NaN	NaN	0.4511	0.3885	0.5
7	NaN	NaN	NaN	0.2375	0.2032	0.2
8	NaN	NaN	NaN	0.3635	0.3220	0.4
9	NaN	NaN	NaN	0.2800	0.2425	0.3
10	NaN	NaN	NaN	0.1905	0.1522	0.2
11	NaN	NaN	NaN	0.8264	0.7495	0.9
12	NaN	NaN	NaN	0.7503	0.6814	0.8
13	NaN	NaN	NaN	0.2852	0.2280	0.3
14	NaN	NaN	NaN	0.2268	0.1879	0.2
15	NaN	NaN	NaN	0.2321	0.1951	0.2
16	NaN	NaN	NaN	0.6077	0.5276	0.6
17	NaN	NaN	NaN	NaN	NaN	
18	NaN	NaN	NaN	NaN	NaN	
19	NaN	NaN	NaN	NaN	NaN	
20	NaN	NaN	NaN	NaN	NaN	
21	NaN	NaN	NaN	NaN	NaN	
22	NaN	NaN	NaN	NaN	NaN	
23	NaN	NaN	NaN	NaN	NaN	
24	NaN	NaN	NaN	NaN	NaN	
25	NaN	NaN	NaN	NaN	NaN	
26	NaN	NaN	NaN	NaN	NaN	
27	NaN	NaN	NaN	NaN	NaN	
28	NaN	NaN	NaN	NaN	NaN	

Buttons: Save, Save Resid., Plot Resid., Scatter Plot, Plot CI, Exit

All the results are placed in an exportable numeric matrix. Separate results are given for each separate culture. The columns for each culture are described below:

- LL;LLPred;RB;RCCT: This column gives the overall measures of fit. The first value (LL) is the optimal log-likelihood value. If XPred holdout validation is implemented, the next value is the holdout prediction log-likelihood. The next value (RB) is the basic measure of solution reliability described in France and Batchelder (2015) and from Janson and Olson (2001). The next value (RCCT) is a consensus adjusted reliability metric, as given in France and Batchelder (2015).
- Validation Results: This column gives a value for each combination of validation test and results metric. The validation tests are z, d, badd, bmult, IE, XPred, and zGr (external comparison of z with the ground truth). The results metrics are the previously described MSE, MSPE, MAE, MAPE, and Corr. For the XPred test, consensus versions of these metrics are used.
- d: The values of **d**, the rater competencies.
- dCILower: The lower bound of the confidence interval for **d**.
- dCIUpper: The upper bound of the confidence interval for **d**.
- badd: The rater additive biases $\mathbf{b}_A$.
- baddCILower: The lower bound of the confidence interval for $\mathbf{b}_A$.
- baddCIUpper: The upper bound of the confidence interval for $\mathbf{b}_A$.
- bmult: The rater multiplicative biases $\mathbf{b}_M$.
- bmultCILower: The lower bound of the confidence interval for $\mathbf{b}_M$.

- bmultCIUpper: The upper bound of the confidence interval for $\mathbf{b}_M$.
- z: The rater multiplicative biases $\mathbf{z}$.
- zCILower: The lower bound of the confidence interval for $\mathbf{z}$
- zCIUpper: The upper bound of the confidence interval for $\mathbf{z}$
- IE: The item easiness values $\boldsymbol{\beta}$.
- IECILower: The lower bound of the confidence interval for $\boldsymbol{\beta}$.
- IECIUpper: The upper bound of the confidence interval for $\boldsymbol{\beta}$.
- LLPartial(LB): The value for the partial log-likelihood values of $\lfloor \mathbf{z} \rfloor$, i.e., the integer lower bound for $\mathbf{z}$.
- LLPartial: The value for the partial log-likelihood values of $\mathbf{z}$.
- LLPartial(UB): The value for the partial log-likelihood values of $\lceil \mathbf{z} \rceil$, i.e., the integer upper bound for $\mathbf{z}$.

The rationale behind the partial log-likelihoods is to provide results for integer values of $\mathbf{z}$, in applications in which integer results are important.

Form Buttons
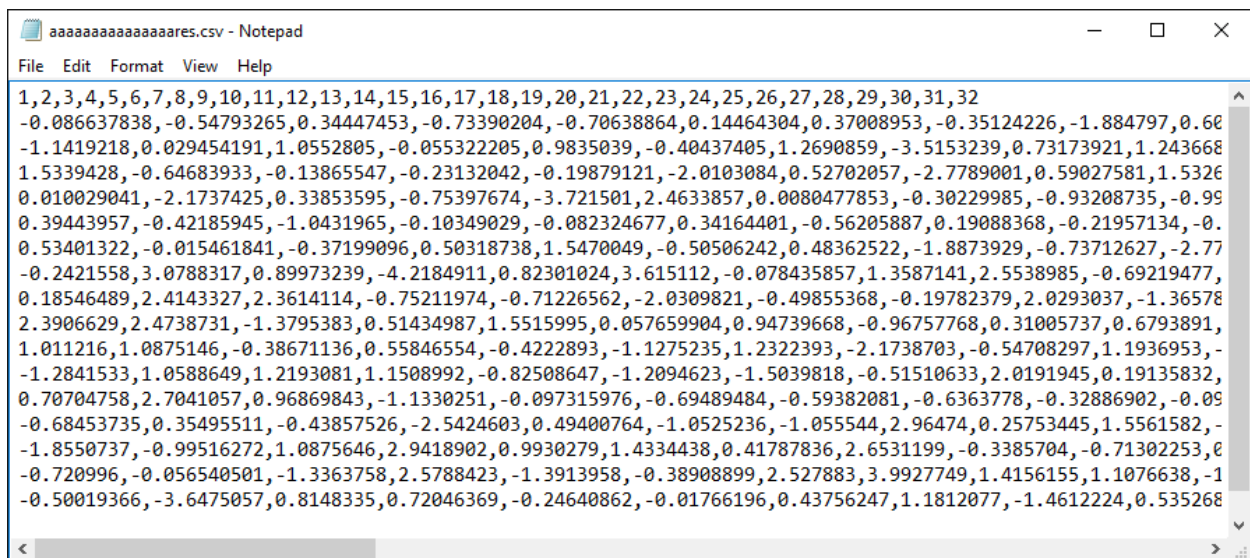
The results form has six buttons, which are situated at the top right hand side of the form. The functionality for each button is described below.

Save

Allows a user save the results in the results matrix into a .csv data file using the standard Microsoft Windows file dialog box.

Save Resid.

Allows a user save the model residuals into a .csv data file. For a rater i and an item k, the residual is calculated as $\varepsilon_{ik} = x_{ik} - b_{Mi}z_k + b_{Ai}$. For n users and m items, an $n \times m$ matrix of residuals is plotted. Each column is labeled with the number of the item. An example residuals file, for 16 users and 32 items, is given below.

Plot Resid.

Opens the "Residual Plots" form, which gives a visual analysis of the model residuals.

Scatter Plot
Opens the "Scatter Plots" form, which allows users to create scatterplots using any combination of rater or item parameter sets.

Plot CI
Opens the "Residual Plots" form, which gives a visual analysis of any confidence intervals specified in the advanced form. This option is only available when confidence interval have been created for at least one parameter set.

Exit
Exits from the CCT Results Output form.

Residual Plots

The Residual Plots form gives graphs of model residuals, allowing users to examine residual patterns and check for error heteroscedasticity. The residuals are plotted against the answer key scores contained in **z**.
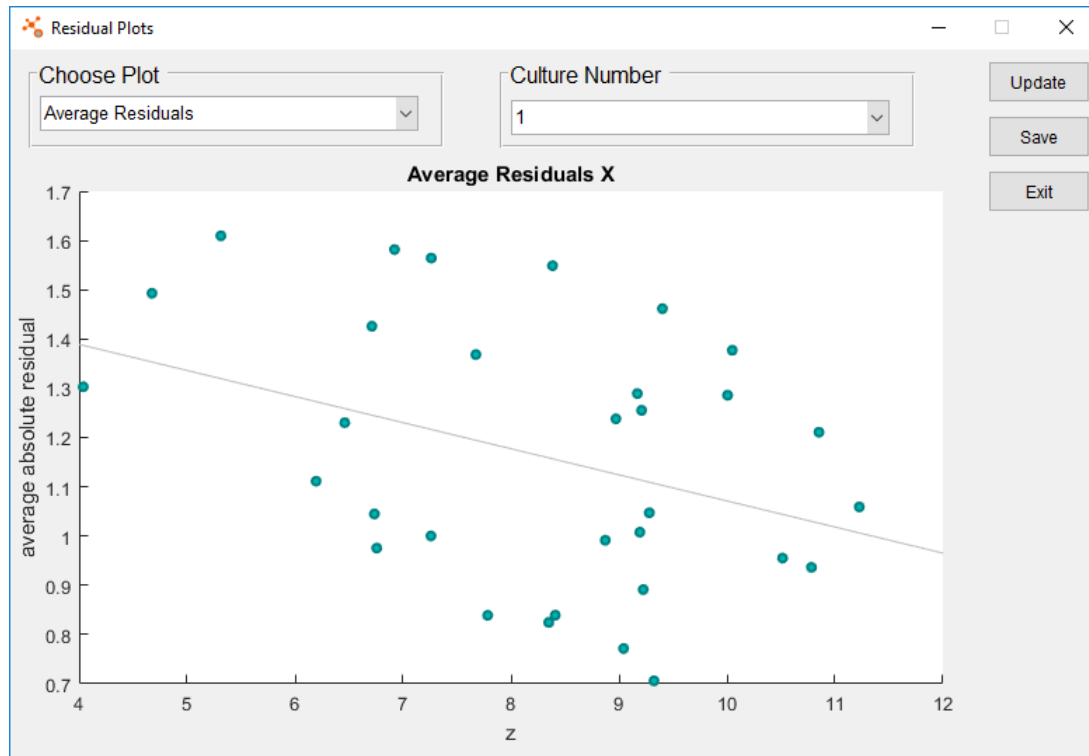
Choose Plot

The dropdown specifies the current plot type. The plot types are:

Average Residuals: Plots the answer key score z_k against the average absolute residual score $\sum_{i=1}^{n}|\varepsilon_{ik}|/n$ for each answer key item. Plots the optimal least squares fit line. If the line has a steep slope then this indicates error heterogeneity and data transformation (e.g., log transformation) may be recommended for the data.

All Residuals: Plots the answer key score against the residual value ε_{ik} for each combination of rater and item. If the spread of the residuals increases with z_k then this indicates error heterogeneity and data transformation may be recommended for the data. If the level of the residuals changes then this indicates that there may be some error autocorrelation.

Culture Number

Residuals are plotted for a single culture. For situations where there are multiple cultures, the Culture Number dropdown selects the currently displayed culture.

Form Buttons

The results form has three buttons, which are situated at the top right hand side of the form. The functionality for each button is described below.

Update

Updates the plot to reflect the current settings.

Save

Allows the saving of the current plot in a graphics (jpg, png, gif, or bmp) file. Brings up a standard Microsoft Windows file dialog box. If a file name is chosen and OK is checked then the file will be saved.

Exit

Exits the form.

Custom Scatterplots

The Custom Scatteplots form displays 2D scatterplots for any combination of available rater ($\mathbf{d}$, $\mathbf{b}_A$, $\mathbf{b}_M$) or item ($\mathbf{z}$, $\boldsymbol{\beta}$) parameter sets.
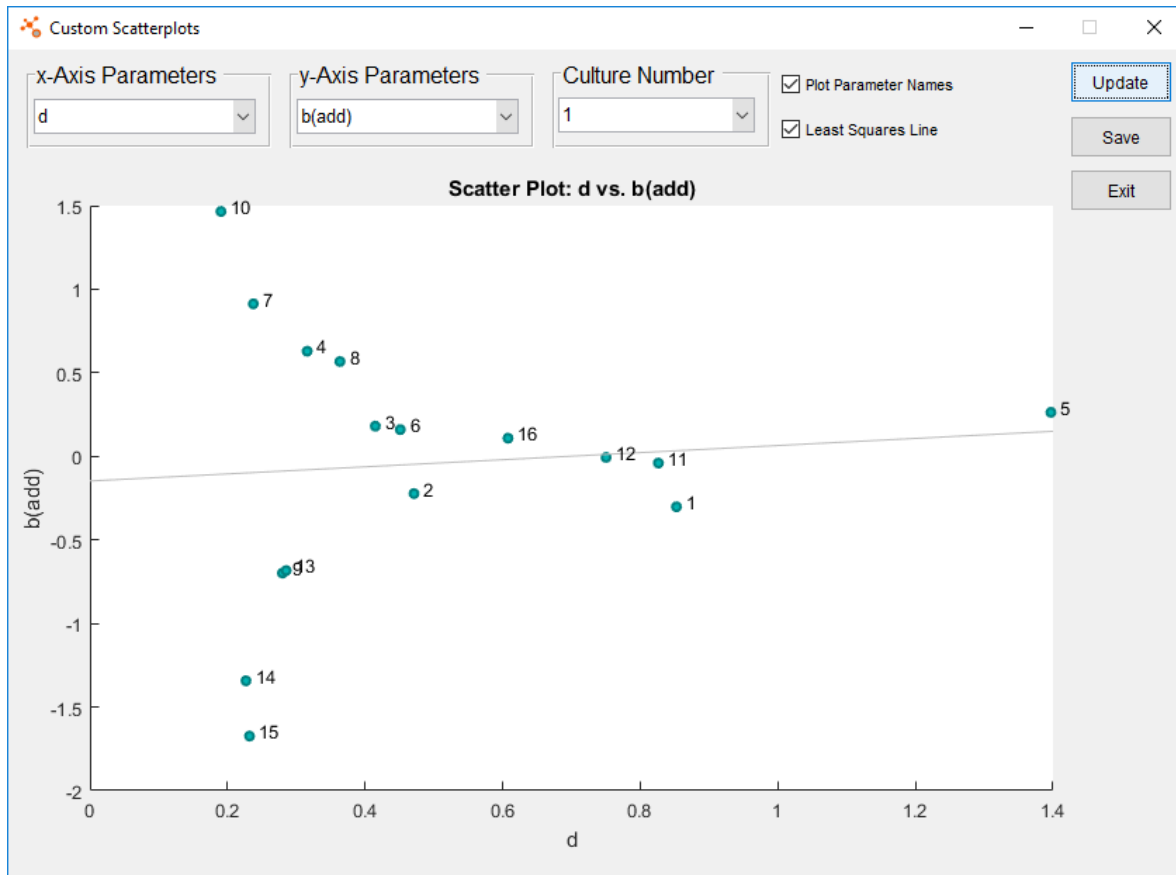
x-Axis Parameters

Select the parameter set for the x-axis. This option automatically controls the options in the "y-Axis Parameters" dropdown. If a rater parameter set ($\mathbf{d}$, $\mathbf{b}_A$, $\mathbf{b}_M$) is chosen for the x-axis then the y-axis dropdown will be populated with the rater parameter sets. Likewise will occur for an item parameter set selection.

y-Axis Parameters

Select the parameter set for the y-axis. This is restricted to be of the same type as the parameter set for the x-axis.

Culture Number

This selects the number of the culture/cluster. For single cluster solutions, there is one option, which is 1. For multi-cluster solutions, this selects the solution derived from the raters in the selected cluster (from 1 to the number of clusters). For each of the rater parameter sets ($\mathbf{d}$, $\mathbf{b}_A$, $\mathbf{b}_M$) the number of plotted points is the number of raters in the cluster. The number of items remains unchanged.

Checkboxes

Plot Parameter Names

Depending on the parameter set selected, plots the names of the raters or items. If names were not uploaded with the data, then the numbers (in order) of the raters/items are used.

Least Squares Line

Plots a least squares regression line for the linear relationship between the two displayed parameter sets.

Form Buttons

The custom scatterplots form has three buttons, which are situated at the top right hand side of the form. The functionality for each button is described below.

Update

Updates the plot to reflect the current settings.

Save

Allows the saving of the current plot in a graphics (jpg, png, gif, or bmp) file. Brings up a standard Microsoft Windows file dialog box. If a file name is chosen and OK is checked then the file will be saved.

Exit

Exits the form.

Confidence Interval Plots

The Confidence Interval Plots form displays confidence interval plots for any parameter sets for which confidence intervals were set on the advanced form. Two types of plot are available, a bar plot, which displays each confidence interval as a vertical line, and a shaded line plot, which displays confidence intervals using line plot across the different parameters, with shaded regions representing the confidence intervals.
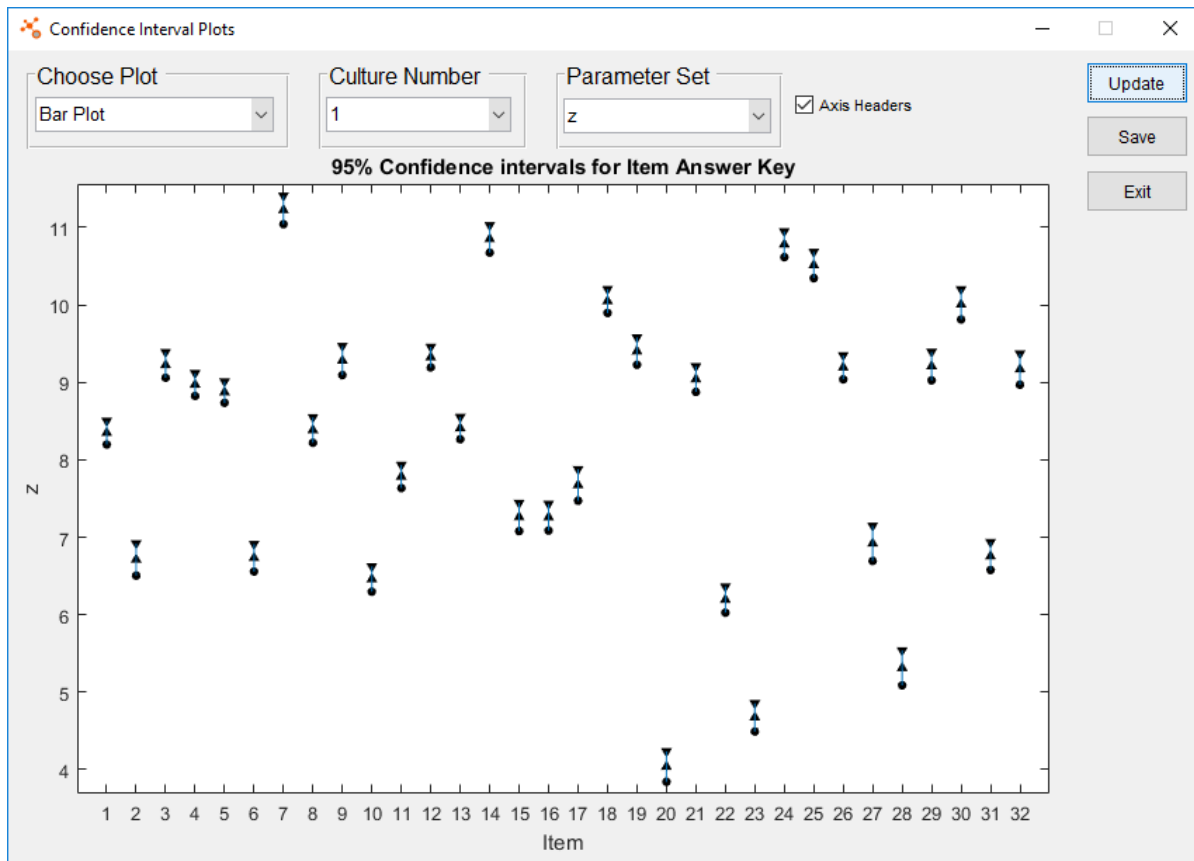
Choose Plot

Selects the type of plot. Consider a 100%(1-) confidence interval [L,P,U], where L is the lower bound of the confidence interval, P is the parameter estimate, and U is the upper bound. The confidence interval is not necessarily symmetric. For example, unless the "Replace Parameters with CI Mean" option is checked, a bootstrap confidence interval has the parameter estimate on the full dataset as its midpoint, which may not be at the center of the interval derived from bootstrapping the data.

- Bar Plot: For each parameter, a vertical line is drawn between L and U. A circle is displayed at L, an upwards arrow at P, and a downwards arrow at U.
- Shaded Line Plot: Three lines, for L, P, and U, are plotted across the parameter values in the order of the raters or items in the x-axis. The area between L and U is shaded. This graph only really makes sense if the raters or items are ordered in some way (e.g., students ranked from 1-50).
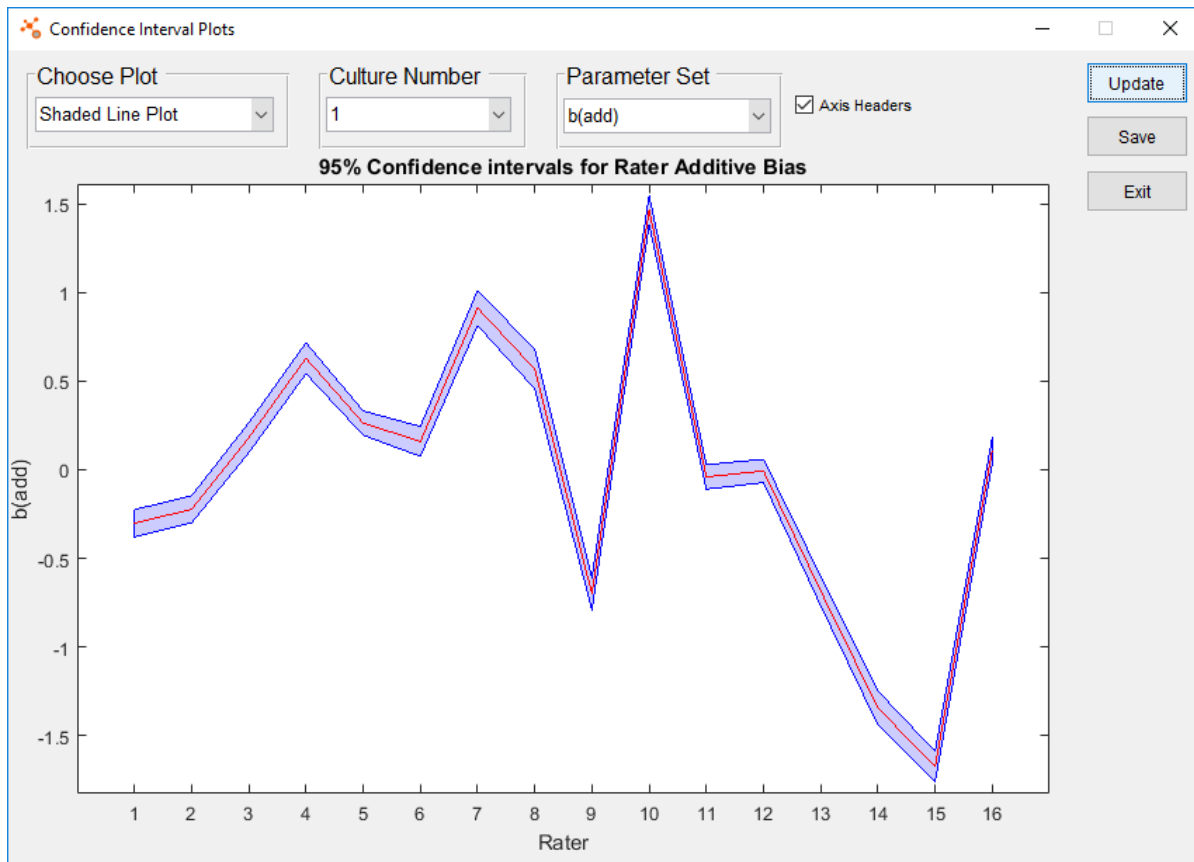
25

Culture Number

This selects the number of the culture/cluster. For single cluster solutions, there is one option, which is 1. For multi-cluster solutions, this selects the solution derived from the raters in the selected cluster (from 1 to the number of clusters). For each of the rater parameter sets (d, b_A, b_M) the number of plotted points is the number of raters in the cluster. The number of items remains unchanged.

Parameter Set

Select the parameter set for the confidence intervals.

Axis Headers

Depending on the parameter set selected, plots the name of the raters or items on the x-axis. If names were not uploaded with the data, then the numbers (in order) of the raters/items are used.

Form Buttons

The results form has three buttons, which are situated at the top right hand side of the form. The functionality for each button is described below.

Update

Updates the plot to reflect the current settings.

Save

Allows the saving of the current plot in a graphics (jpg, png, gif, or bmp) file. Brings up a standard Microsoft Windows file dialog box. If a file name is chosen and OK is checked then the file will be saved.
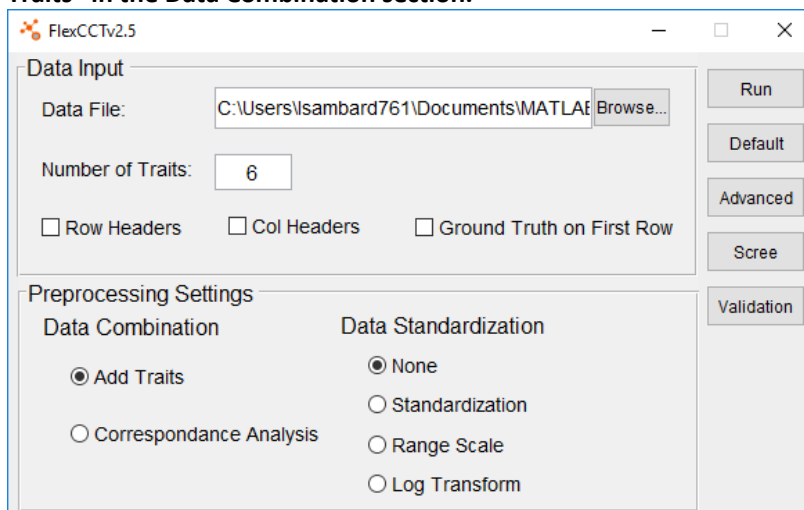
Exit

Exits the form.

Example: Essay Grading Competency

Overview

This example gives a typical educational evaluation example for the use of CCT. A series of essays were graded by two expert graders and 12 peer graders, giving a total of 14 evaluators. Each grader rated 50 different essays. These essays were written by middle school students on the subject of laughter. Each essay was graded using a fixed grading rubric that contained six different attributes. These attributes were "Ideas and Content", "Organization", "Voice", "Word Choice", "Sentence Fluency", and "Conventions". Each essay was graded on each attribute from 1 to 6 by each rater. The attributes were grouped using the format given on p2 in the file GradingSixTraits.csv, which is included in the install directory.

Instructions.

From the main screen select the file, type in 6 into the Number of Traits textbox and then select "Add Traits" in the Data Combination section.

Given that all raters used the same grading rubric, it would be a sensible to assume that the raters belong to a single "rating" culture. However, this is not necessarily always true. Different groups of raters could interpret the rubric in different ways. Thus, it is a sensible precaution to ensure that the number of cultures is one.

Press the Scree button
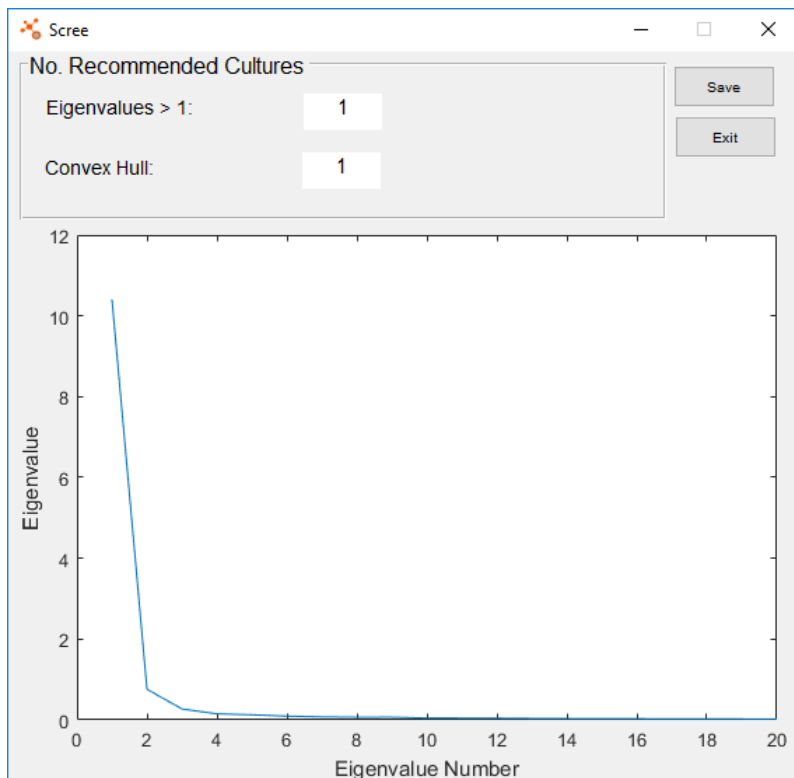
The scree plot form gives a plot of eigenvalue number versus eigenvalue for a principal axes factor analysis of the inter-rater correlations. Two common criteria for choosing the number of factors in factor analysis are 1) the number of eigenvalues less than one and ii) the convex hull, which is the eigenvalue number where the eigenvalue line plot transitions from making an angle >45° to the origin to making an angle <45° to the origin. In both cases here the number of cultures is equal to 1.

The input options on the main form are given below. Choose the basic "ML Model" with "Additive Bias". Keep with the default advanced settings.

Click the run button. The results are given below.

	Culture	LL;RB;RCCT	d	badd	bmult	z
1	1	-1.8180e+03	0.1100	0.4855	1	6.3630
2	NaN	0.7232	0.2042	1.2455	1	7.8560
3	NaN	0.8114	0.0854	-3.2945	1	12.5954
4	NaN	NaN	0.0809	-3.6545	1	9.4021
5	NaN	NaN	0.1505	1.3255	1	15.7545
6	NaN	NaN	0.1177	-0.2945	1	11.1835
7	NaN	NaN	0.0662	0.8055	1	11.1092
8	NaN	NaN	0.1638	0.8455	1	12.7256
9	NaN	NaN	0.0356	2.4855	1	14.2577
10	NaN	NaN	0.0739	-0.1945	1	14.7878
11	NaN	NaN	0.1471	-2.4745	1	14.6578
12	NaN	NaN	0.0400	-1.5345	1	12.9806
13	NaN	NaN	0.1212	1.1255	1	15.3803
14	NaN	NaN	0.0876	-4.6945	1	16.8495
15	NaN	NaN	NaN	NaN	NaN	14.5099
16	NaN	NaN	NaN	NaN	NaN	16.3520
17	NaN	NaN	NaN	NaN	NaN	17.6347
18	NaN	NaN	NaN	NaN	NaN	16.9553
19	NaN	NaN	NaN	NaN	NaN	15.9634
20	NaN	NaN	NaN	NaN	NaN	18.8624
21	NaN	NaN	NaN	NaN	NaN	18.1215
22	NaN	NaN	NaN	NaN	NaN	20.4483
23	NaN	NaN	NaN	NaN	NaN	21.3027
24	NaN	NaN	NaN	NaN	NaN	19.7874
25	NaN	NaN	NaN	NaN	NaN	22.9526
26	NaN	NaN	NaN	NaN	NaN	21.9203
27	NaN	NaN	NaN	NaN	NaN	21.8127
28	NaN	NaN	NaN	NaN	NaN	23.7971

Buttons: Save, Save Resid., Plot Resid., Scatter Plot, Plot CI, Exit

Once can see a range of competencies (d), ranging from 0.0356 to 0.2042. The first two raters are the expert raters. Expert rater 2 has the highest competency of all the rates, but expert rater 1 has a more than average competency. Note, that the results have high reliability (0.7232) and even higher competency adjusted reliability (0.8114).

To check for error heterogeneity, press the Plot Resid. Button. The results are given below.
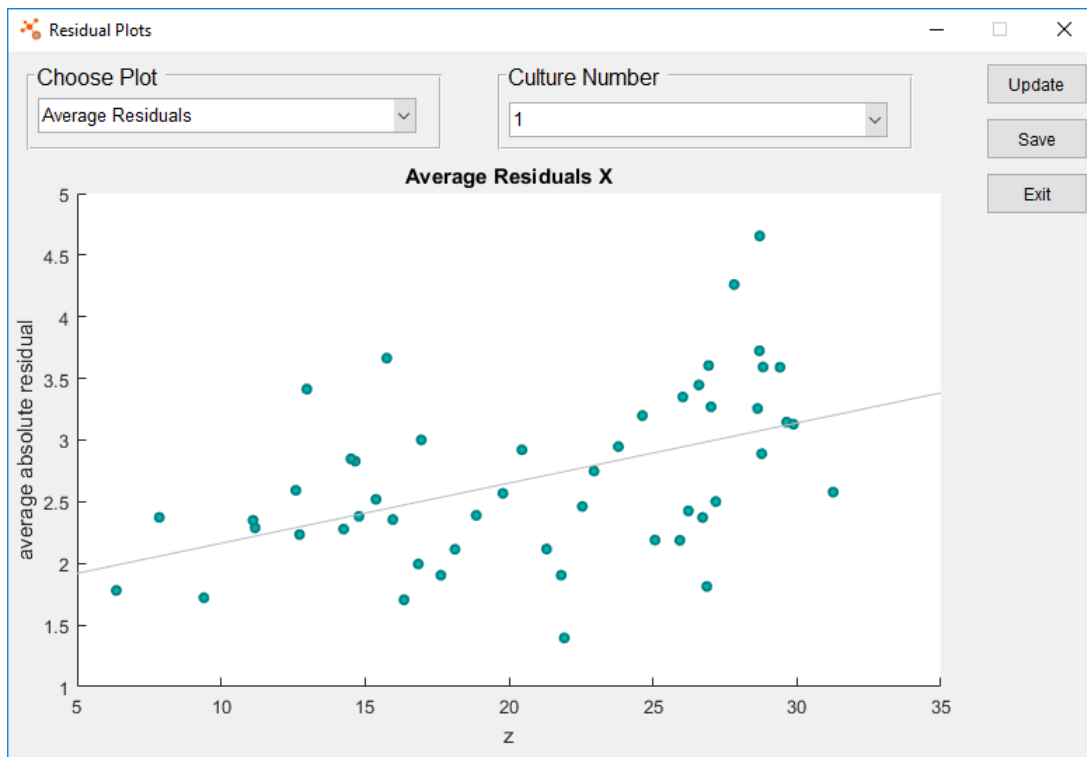
The plot shows the average magnitude of residuals across the values of z. There is a slight upward trend in the residuals, but no clear trend indicating heterogeneity.

To get a further review of the residuals, click the All Residuals dropdown and then click the Update button.
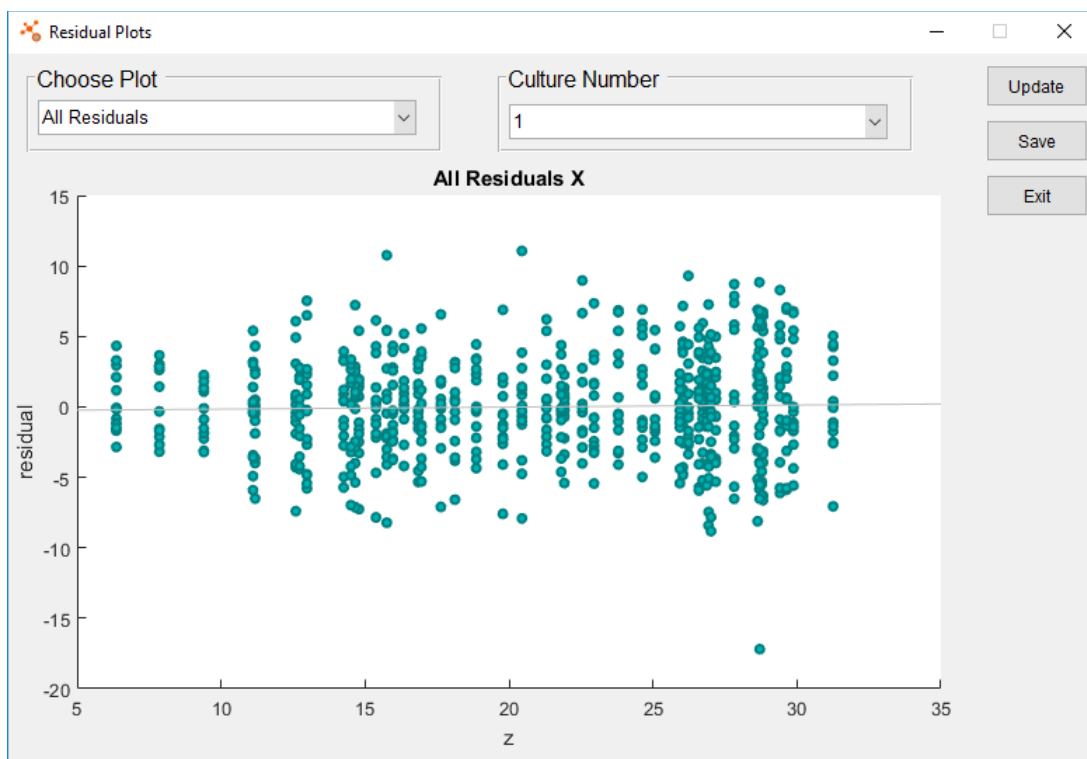
This further shows that relative to the range of residuals, any trend is almost non-existent, giving evidence of a lack of error autocorrelation.

To examine the interplay of model parameters, from the Results screen, select the "Scatter Plot" button. In the selected model, there are two sets of rater parameters (**d**, **b$_A$**) in the model (**b$_A$** is set to default of **1**) and one set of item parameters (**z**) in the model (**β** is set to default of **1**).
Select the d option for the x-axis and the b(add) option for the y-axis. Select the checkboxes for both the parameter names and the least squares line.
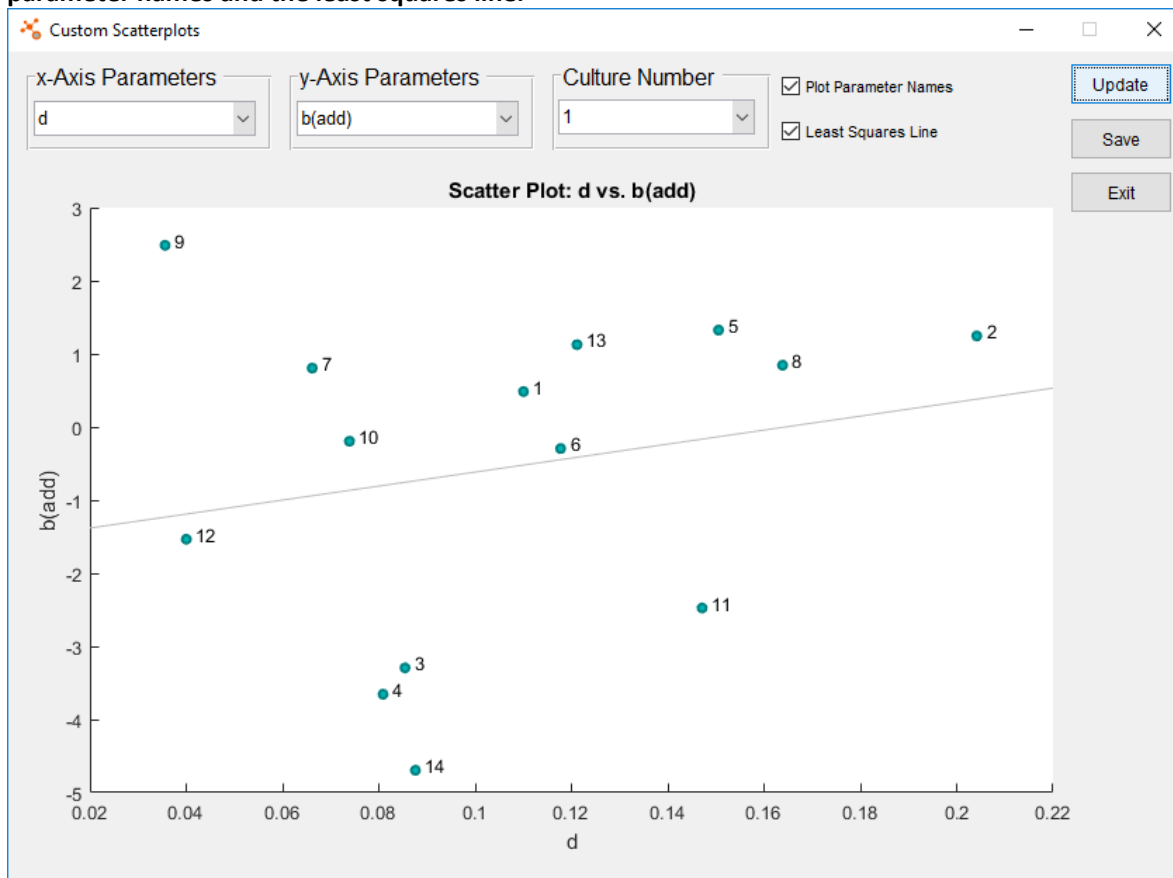
The regression line is almost linear, indicating no or little linear relationship between competency and additive bias. Raters 1 and 2 are expert raters, while raters 3-12 are student raters. Rater 2 has the highest competency, but rater 1 is somewhat in the middle of the field.

Instructions for Advanced Model
Now select the multiplicative item easiness ("IE Multiply") model with both additive and multiplicative bias in the Model Settings section of the main form.

The model is now more flexible and fits item easiness parameters β, additive biases b_A, and multiplicative biases b_M.

Click the "Advanced" button, configure the form as follows, and click the "Save" button.

The Variable Identifiability section is used to ground parameters and prevent parameters being traded off against one another. For example, the partial fixes for **z** sets a single value of **z** to be its value for a restricted model for no bias. The partial fix for **d** does likewise. We choose jackknife hold-one out sampling for the confidence intervals (hence the number of resamples is only an upper bound). The confidence intervals provide a measure of confidence for the underlying parameters. For rater parameters (**d**, b_A, and b_M), the confidence intervals give population estimates of the parameters, under the assumption that the items are sampled randomly from an underlying population of items. For item parameters (β, **z**), the assumption that the raters are sampled randomly from an underlying population of raters. The value of Alpha (α), defines a $100\%(1-\alpha)$ confidence interval.

We wish to include some validation procedures to check the stability of the parameters.
From the main form, select the validation button. Select the Internal Validation checkbooks for all available parameters and select the MSE (mean squared error) criterion. Select the holdout size to be 1, which will give 700 holdout validation test datasets, each with one item.

Validation

Internal Holdout Validation

☑ z   ☑ d   ☑ IE   ☑ b(add)   ☑ b(mult)   ☐ XPred

No. Raters: 14   No. Items: 50

No. Ratings: 700

Holdout Size: 1

Maximum Runs: 100000

Save
Default

External Holdout Validation

☐ Compare to Ground Truth

Criteria

☑ MSE   ☐ MSPE   ☐ MAE   ☐ MAPE   ☐ Correl

Click the Run Button

CCT Results Output

	Culture	LL;RB;RCCT	zMSE;dMSE;baddMSE;bmultMSE;IEMSE	d
1	1	-1.7308e+03	0.0179	0.1191
2	NaN	0.7232	0.0114	0.2768
3	NaN	0.9094	0.0133	0.0702
4	NaN	NaN	2.9238e-05	0.0656
5	NaN	NaN	1.8110	0.1211
6	NaN	NaN	NaN	0.1344
7	NaN	NaN	NaN	0.0719
8	NaN	NaN	NaN	0.1529
9	NaN	NaN	NaN	0.0340
10	NaN	NaN	NaN	0.0739
11	NaN	NaN	NaN	0.2319
12	NaN	NaN	NaN	0.0427
13	NaN	NaN	NaN	0.1159
14	NaN	NaN	NaN	0.0831
15	NaN	NaN	NaN	NaN
16	NaN	NaN	NaN	NaN
17	NaN	NaN	NaN	NaN
18	NaN	NaN	NaN	NaN
19	NaN	NaN	NaN	NaN
20	NaN	NaN	NaN	NaN
21	NaN	NaN	NaN	NaN
22	NaN	NaN	NaN	NaN
23	NaN	NaN	NaN	NaN
24	NaN	NaN	NaN	NaN
25	NaN	NaN	NaN	NaN
26	NaN	NaN	NaN	NaN
27	NaN	NaN	NaN	NaN
28	NaN	NaN	NaN	NaN

Save
Save Resid.
Plot Resid.
Scatter Plot
Plot CI
Exit

The consensus adjusted reliability in this model is 0.9094, which is much higher than for the basic model. Validating the model parameters gives MSE values of 0.0179 for z, 0.0114 for d, 0.0133 for b_A, <0.0001 for b_M, and 1.8110 for β. This suggests that the model parameters are relatively stable, except perhaps for those for β. For each set of parameters with a specified confidence interval, the lower bound and upper bound for the confidence interval are given after the actual value. To save all of the results and perhaps plot/do initial analyses, the Save button allows an output .csv file to be exported using a standard Microsoft file dialog box.

As before, the model residuals can be analyzed using the "Plot Residuals" option. Given that all possible parameters are estimated, more plots are available on the "Custom Scatterplots" screen.

Push the "Scatter Plot" button. Select z under x-Axis Parameters and IE under y-Axis Parameters. Select both the checkbox options and press "Update".
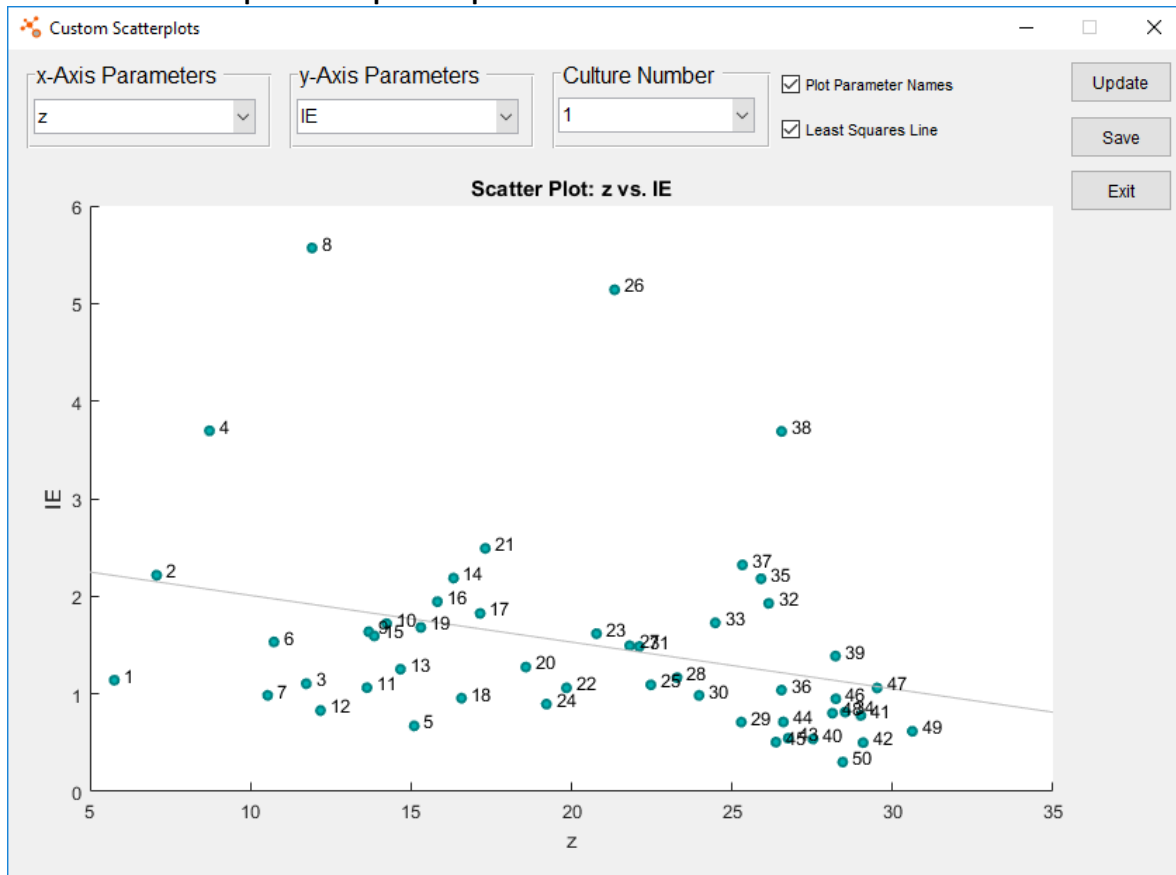
The results show a slight decrease in easiness as the quality (z score) of the essay increases. This makes intuitive sense, as a poor essay can be easily critiqued on major faults, while with a strong essay, the process of finding faults is more difficult.

To examine the jackknife confidence intervals for the individual variables, from the Results screen, press the "Plot CI" button and then select "Bar Plot" and the "z" parameter set. Press "Update".
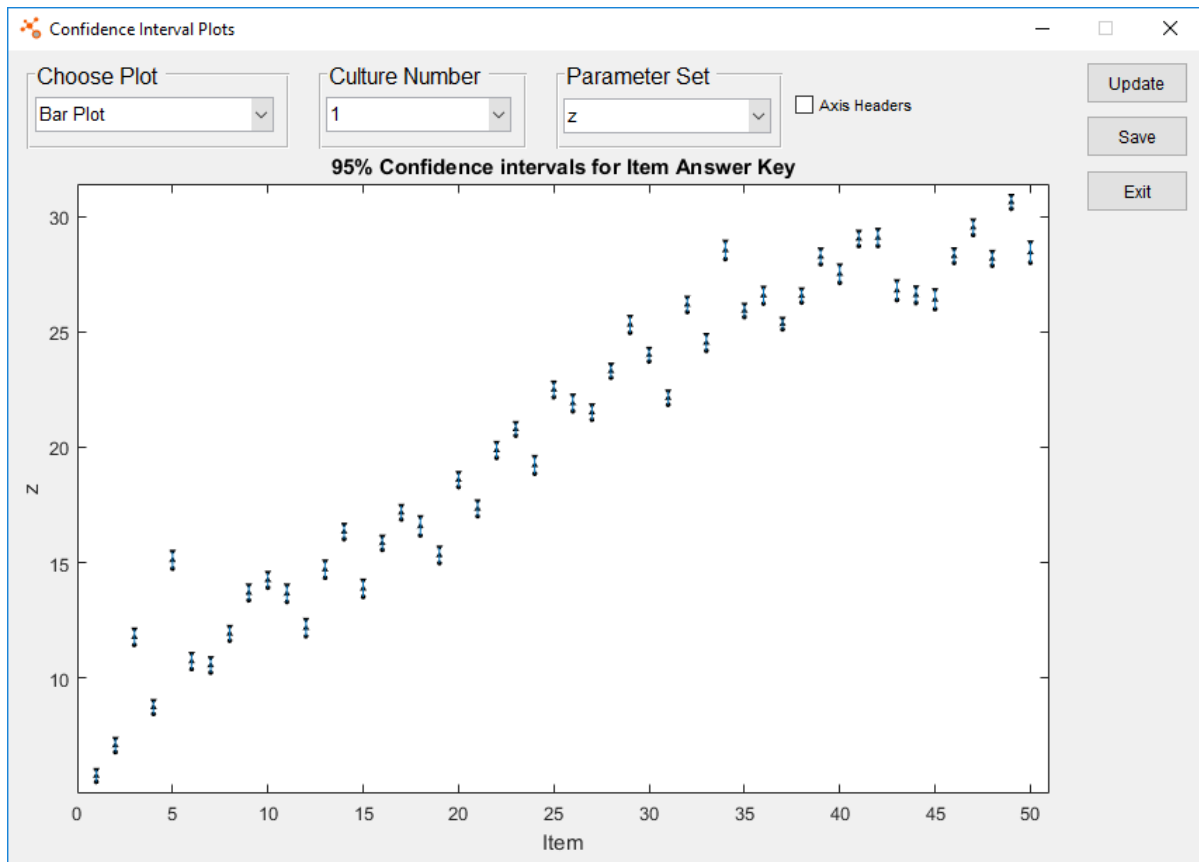
The essays are ordered roughly by the quality of the essay. The confidence intervals are relatively narrow, suggesting a high degree of parameter stability in the face of substantial missing data.