IRTPRO™

Contents

IRT scoring

1. Introduction

Unlike classical test theory, IRT does not in general base the estimate of the respondent's ability (or other attribute) on the number-correct (NC) or summed score. To distinguish IRT scores from their classical counterparts, we refer to them as "scale" scores. There are two instances under which the IRT scale scores may be one-to-one related (in a nonlinear fashion) to summed scores. First, when the one-parameter logistic (or in general, Rasch) model is used, the summed scores are sufficient statistics for the latent ability variable. Second, when the scale scores are based on summed-score posteriors for any IRT model, the summed scores can be directly translated into scale scores.

The main advantages of scale scores are that they:

- Remain comparable when items are added to or deleted from the tests.
- Weight the individual items optimally according to their discriminating powers.
- Have standard errors that are more accurate.
- Provide more flexible and robust adjustments for guessing than the classical corrections.
- Are on the same continuum as the item locations.

There are three types of IRT scale score estimation methods that IRTPRO supports:

- Bayes estimation (EAP)
- Summed Score EAP (SSEAP)

- Bayes modal estimation (MAP)

The three types of IRT scale score estimation methods are discussed in the sections to follow.

1.1 Bayes estimation (EAP)

The Bayes estimate is the mean of the posterior distribution of θ, given the observed response pattern x_i (Bock & Mislevy, 1982). It can be approximated as accurately as required by the Gaussian quadrature (see the section on MML estimation):

$$\overline{\theta_i} \cong \frac{\sum_{k=1}^{q} X_k P(\mathbf{x}_i \mid X_k) A(X_k)}{\sum_{k=1}^{q} P(\mathbf{x}_i \mid X_k) A(X_k)}.$$

This function of the response pattern $\mathbf{x}_i$ has also been called the expected a posteriori (EAP) estimator. A measure of its precision is the posterior standard deviation (PSD) approximated by

$$PSD(\overline{\theta_i}) \cong \frac{\sum_{k=1}^{q} (X_k - \overline{\theta_i})^2 P(\mathbf{x}_i \mid X_k) A(X_k)}{\sum_{k=1}^{q} P(\mathbf{x}_i \mid X_k) A(X_k)}.$$

The weights, $A(X_k)$, in these formulas depend on the assumed distribution of θ. Theoretical weights, empirical weights, $A^*(X_k)$, or subjective weights are possibilities.

The EAP estimator exists for any answer pattern and has a smaller average error in the population than any other estimator, including the ML estimator. It is in general biased toward the population mean, but the bias is small within $\pm 3\sigma$ of the mean when the PSD is small (*e.g.*, less than 0.2σ). Although the sample mean of EAP estimates is an unbiased estimator of the mean of the latent population, the sample standard deviation is in general smaller than that of the latent population. This is not a serious problem if all the respondents are measured within the same PSD. However, it could be a problem if respondents are compared using alternative test forms that have much different PSDs. The same problem occurs, of course, when number-right scores from alternative test forms with differing reliabilities are used to compare respondents. Tests administrators should avoid making comparisons between respondents who have taken alternative forms that differed appreciably in their psychometric properties. A further implication is that, if EAP estimates are used in computerized adaptive testing, the trials should not terminate after a fixed number of items, but should continue until a prespecified PSD is reached.

1.2 Summed Score EAP (SSEAP)

IRT models also imply posteriors for the summed scores, even if the IRT model used is not among the Rasch family of models. Without loss of generality, consider the dichotomous case first. For any IRT

model with dichotomous item scores $(u_i = 0,1)$, the likelihood for any summed score $x = \sum u_i$ is

$$L_x(\theta) = \sum_{\sum u_i = x} L(\mathrm{u} \mid \theta)$$

where

$$L(\mathrm{u} \mid \theta) = \prod_i T(u_i \mid \theta)$$

and $T(u_i \mid \theta)$ is the traceline for response u to item i. The first summation is over all such response patterns that the summed score equals x. The probability of each score is

$$P_x = \int L_x(\theta) g(\theta)$$

and the expected θ associated with each summed score x is

$$E(\theta \mid x) = \frac{\int \theta L_x(\theta) g(\theta)}{P_x}$$

with posterior standard deviation given by

$$PSD\left(\theta \mid x = \sum u_i\right) = \left(\frac{\int \left[\theta - E(\theta \mid x)\right]^2 L_x(\theta) g(\theta)}{P_x} \right)^{\frac{1}{2}}.$$

1.3 Bayes modal estimation (MAP)

Similar to the Bayes estimator, but with a somewhat larger average error, is the Bayes modal or so-called maximum a posteriori (MAP) estimator. It is the value of θ that maximizes

$$P(\theta \mid x_i) = \sum_{j=1}^{n} \{x_{ij} \log_e P_i(\theta) + (1 - x_{ij}) \log_e [1 - P_i(\theta)]\} + \log_e g(\theta),$$

where $g(\theta)$ is the density function of a continuous population distribution of θ. The likelihood equation is

$$\sum_{j=1}^{n} \frac{x_{ij} - P_j(\theta)}{P_j(\theta)[1 - P_j(\theta)]} \cdot \frac{\partial P_j(\theta)}{\partial \theta} + \frac{\partial \log_e g(\theta)}{\partial \theta} = 0.$$

Analogous to the maximum likelihood estimate, the MAP estimate is calculated by Fisher scoring, employing the *posterior information*,

$$J(\theta) = I(\theta) + \frac{\partial^2 \log_e g(\theta)}{\partial \theta^2},$$

where the right-most term is the second derivative of the population log density of θ.

In the case of the 2PL model and a normal distribution of θ with variance σ^2, the posterior information is

$$I(\theta) = \sum_{j=1}^{n} a_j^2 P_j(\theta)[1 - P_j(\theta)] + \frac{1}{\sigma^2}.$$

The PSD of the MAP estimate, θ, is approximated by

$$\text{PSD}(\theta) = \sqrt{1 / I(\hat{\theta})}.$$

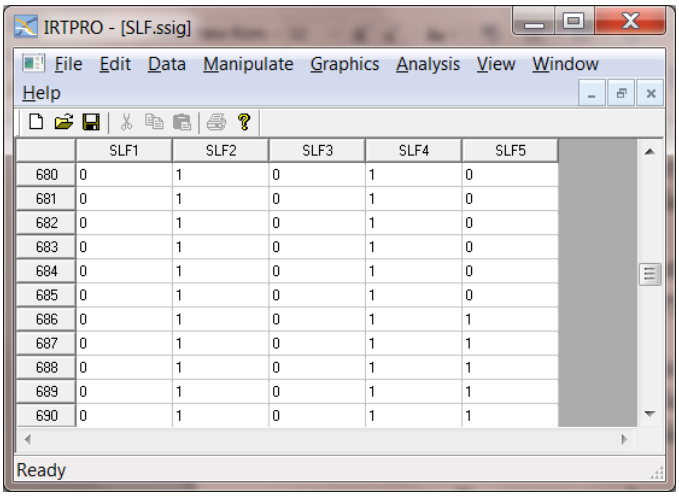
Like the EAP estimator, the MAP estimator exists for all response patterns, but is generally biased toward the population mean.

2. Scoring using a social life feelings (SLF) dataset

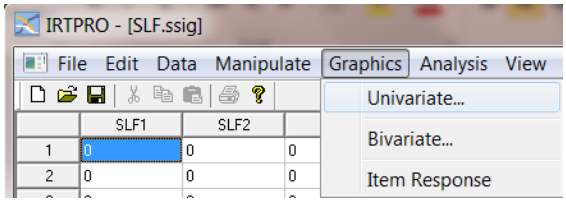
The dataset used in this section is taken from an extensive study of social life feelings reported in Schuessler (1982) and Krebs and Schuessler (1987). The aim was to establish scales for use in social research. According to Bartholomew (1998), the aim of the study was to establish scales for use in social research that were comparable in quality with those used in ability testing. For illustration purposes, the data used in this section is from the German sample consisting of the following five items:

- Anyone can raise his standard of living if he is willing to work at it (SLS1).
- Our country has too many poor people who can do little to raise their standard of living (SLS2).
- Individuals are poor because of the lack of effort on their part (SLS3).
- Poor people could improve their lot if they tried (SLS4).
- Most people have a good deal of freedom in deciding how to live (SLS5).

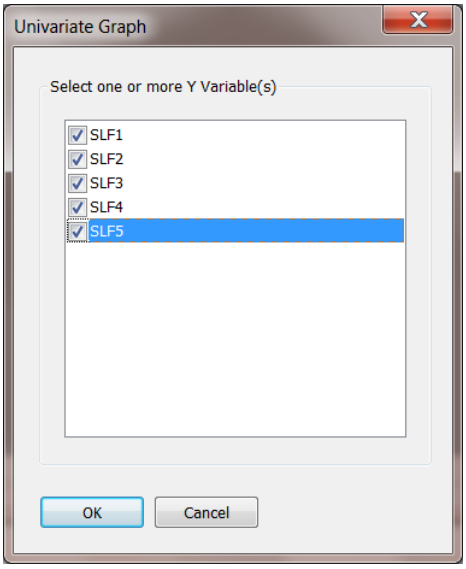
Responses are based on a sample size of 1490 individuals. The spreadsheet below displays item values for cases 680 to 690. The name of the dataset is **SLF.ssig** and is stored in the folder **IRTPRO Examples\By Dataset\Social Life Feelings**



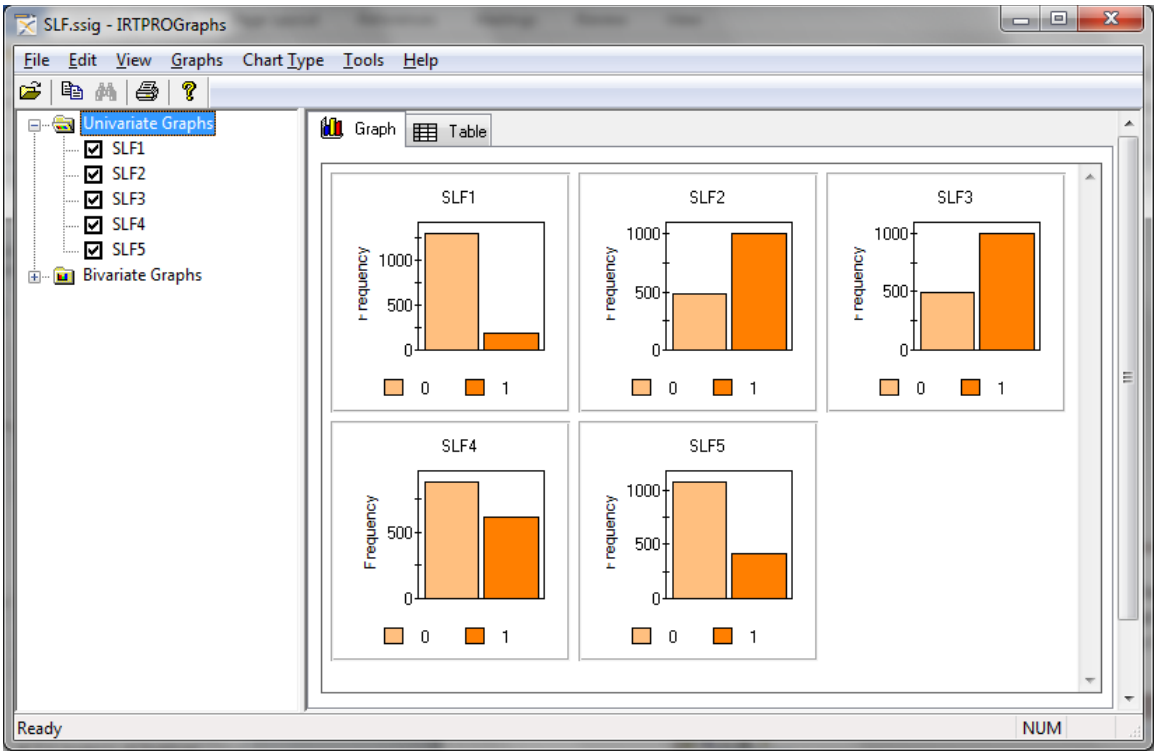
In order to display the frequency distribution of the five items visually, the **Graphics, Univariate…** option is selected from the main menu bar.



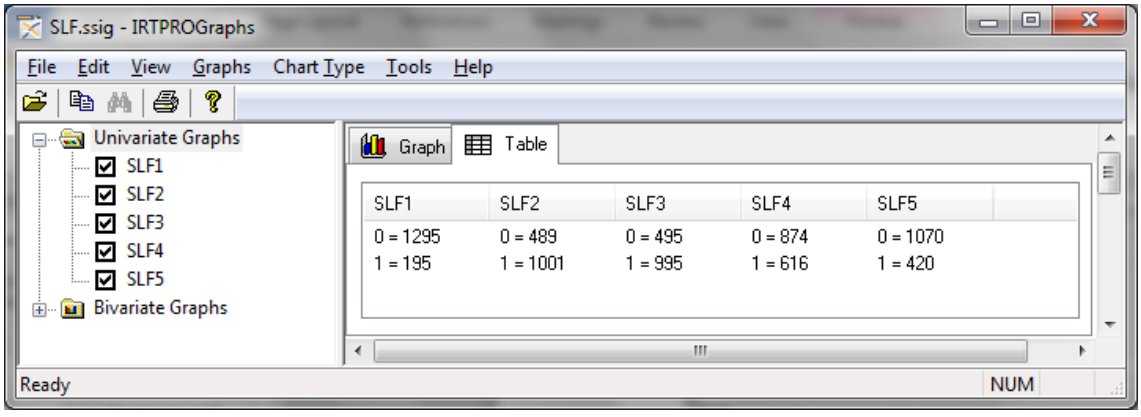
By making this selection, a **Univariate Graph** dialog is displayed.



After selecting the items (see above), click **OK** to obtain a bar chart presentation for each item. From this display, it can be concluded, for example, that a relatively small proportion of individuals have selected the category corresponding to "1" for the item SLF1.



The frequency counts for each item can be displayed by clicking the **Table** icon in the **Graph** window. From this display it follows that all items are binary. We further conclude that there are no missing values present since the total count of the datavalues 0 and 1 equals 1490 for each item.

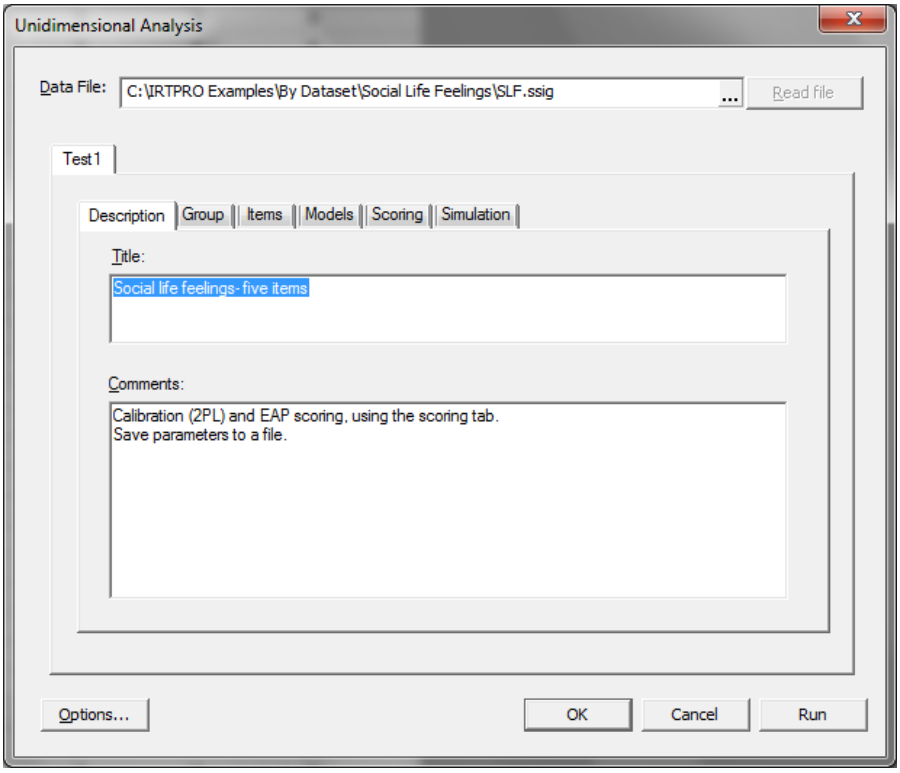


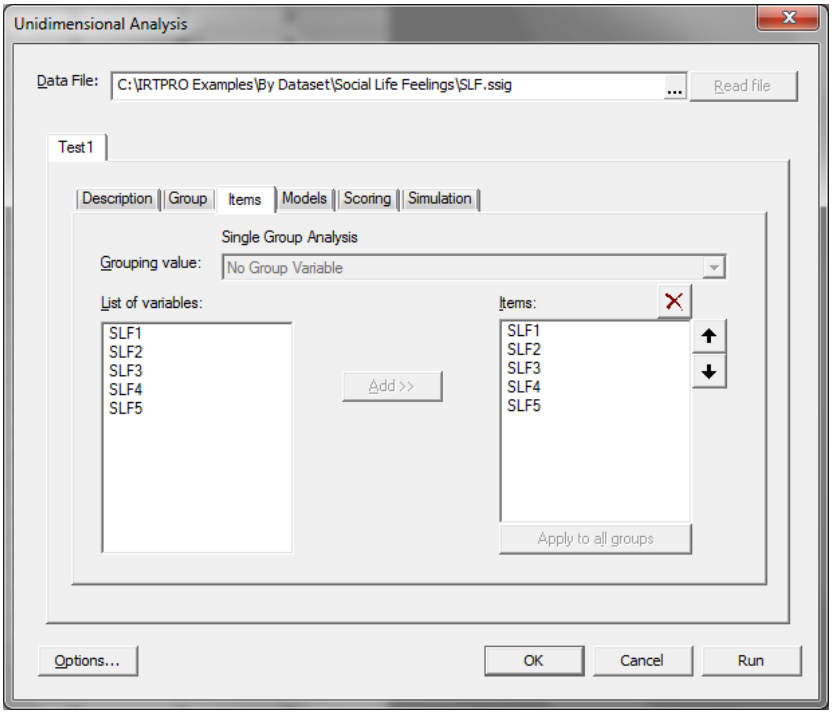
In Sections 8.2.1 and 8.2.2, examples that illustrate item scoring are presented. In the example presented in Section 8.2.1, 2PL models are fitted to the five items, labeled SLS1 to SLS5 respectively and EAP scores are computed. Use is made of the **Advanced Options, Miscellaneous; Save** selection to specify that the estimated parameters must be saved in a file with extension **-prm.txt**. The example in Section 8.2.2 demonstrates item scoring using a **-prm.txt** parameter file obtained from a previously executed IRT (calibration) run.

2.1 Calibration and Scoring

In this section, 2PL models are fitted to the five items, labeled SLS1 to SLS5 respectively, and the

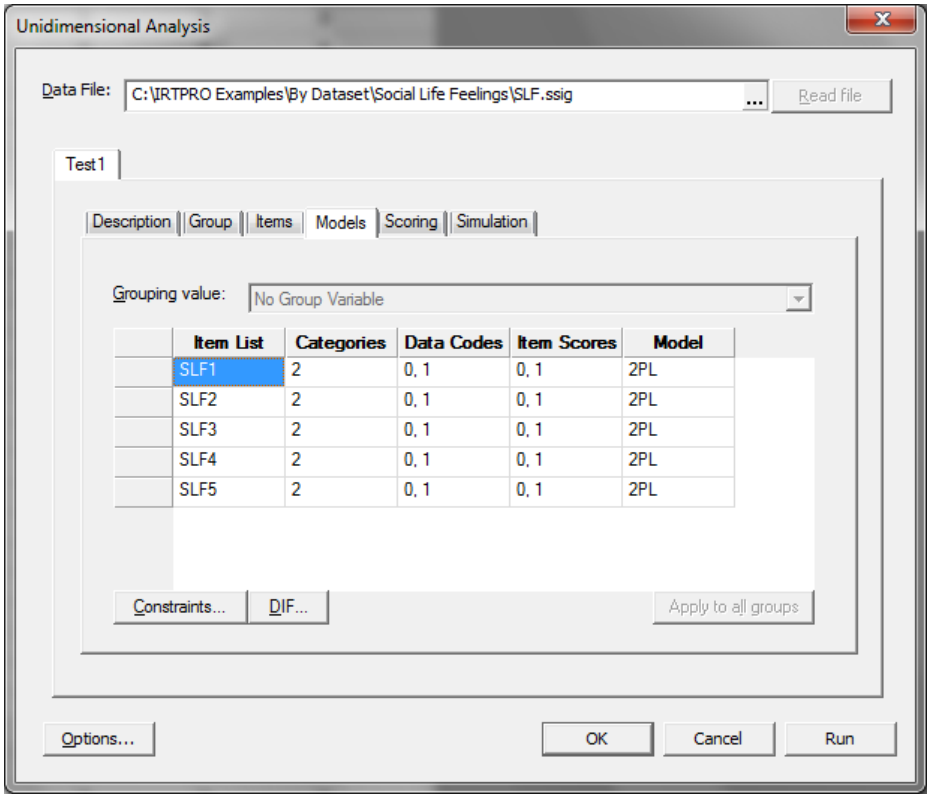
estimated parameters are saved to a text file with extension **-prm.txt**. EAP scores (See Section 8.1.1) are also computed.



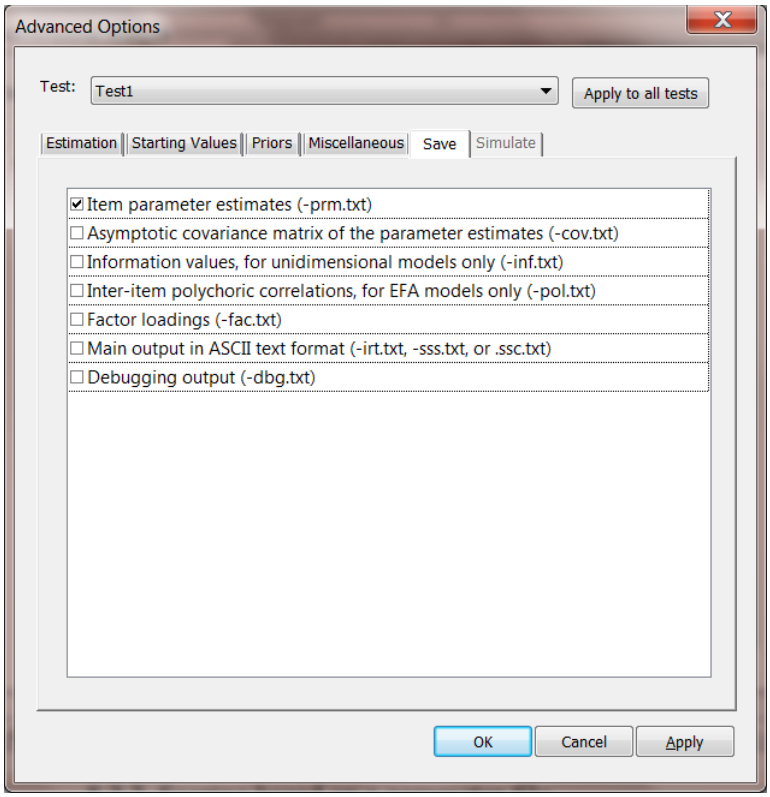


From the main menu bar, select the **Analysis, Unidimensional IRT…** option to obtain the **Unidimensional Analysis** window shown above. Use the **Description** tab to enter a title and comments. Since the dataset **SLF.ssig** is based on a single group (Germany, 1490 individuals), the **Group** tab is skipped and we proceed to the **Items** tab to select all five items.

The **Models** dialog displays the model-type to be fitted to each item. Since all items are binary, the default model is 2PL.



To ensure that the estimated parameters are save to a **–prm.txt** file, click the **Options…** button (bottom right-hand corner of the previous display). This action invokes the **Advance Options** window. Click the **Save** tab and make sure that **Item parameters estimates** (**–prm.txt)** is selected.

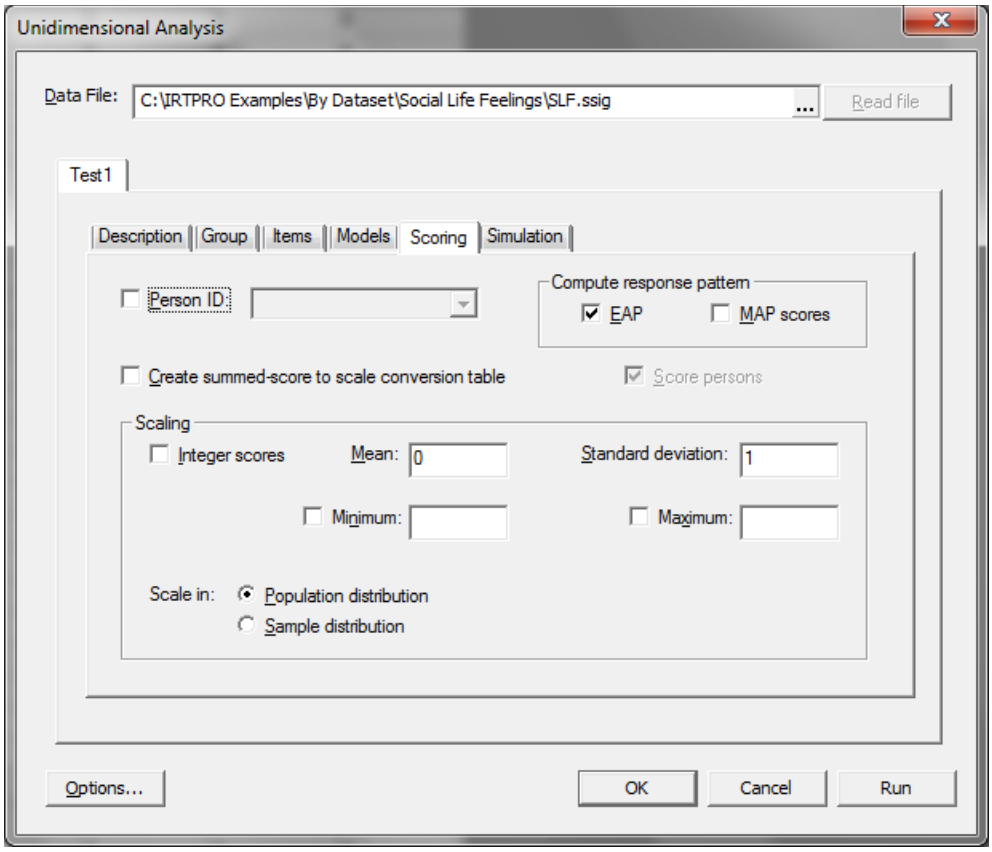


Click **OK** to return to the **Unidimensional Analysis** window and click on the **Scoring** tab to display the **Scoring** dialog.

Using this dialog, make the following selections:

- Scoring method: EAP
- **Scaling**: **Mean** = 0; **Standard deviation** = 1 (the defaults)
- **Scale in**: Population distribution

When done, click the **Run** button to start calibration and scoring.



If the analysis completes successfully, two output files are created with extensions:

- –**irt.htm** (calibration), and
- –**ssc.htm** (Scoring).

The Window menu (below) shows the selection of the IRT analysis (calibration) output.

Project:	Social life feelings- five items
Description:	Calibration (2PL) and EAP scoring, using the scoring tab. Save parameters to a file.
Date:	17 December 2014
Time:	10:26 PM

Table of Contents

Portions of the output are given next. The first table gives the parameter estimates and standard error estimates for the five items.

2PL Model Item Parameter Estimates for Group 1, logit: aθ + c or a(θ – b) (Back to TOC)

Item	Label	a		s.e.	c		s.e.	b	s.e.
1	SLF1	2	1.20	0.15	1	-2.35	0.13	1.97	0.18
2	SLF2	4	0.71	0.09	3	0.80	0.06	-1.11	0.14
3	SLF3	6	1.53	0.17	5	0.99	0.09	-0.65	0.06
4	SLF4	8	2.55	0.39	7	-0.67	0.12	0.26	0.04
5	SLF5	10	0.92	0.10	9	-1.10	0.07	1.19	0.12

Likelihood based statistics and fit statistics are given in the output shown below. The statistic: – 2 log likelihood (also called the deviance statistic) is used to compare nested models. Both the AIC and BIC statistics are used as a model selection tool.

Likelihood-based Values and Goodness of Fit Statistics (Back to TOC)

Statistics based on the loglikelihood	
-2loglikelihood:	8258.37
Akaike Information Criterion (AIC):	8278.37
Bayesian Information Criterion (BIC):	8331.43

The RMSEA value of 0.02 indicates a relatively good fit using the 2PL model.

Statistics based on the full item x item x ... classification			
G^2	Degrees of freedom	Probability	RMSEA
39.09	21	0.0095	0.02
X^2	Degrees of freedom	Probability	RMSEA
38.92	21	0.0100	0.02

Next, we select the output file generated for the scoring part of the analysis (**-ssc.htm**). Selected output is shown below. The first portion is a table containing the parameter estimates obtained in the calibration phase.

2PL Model Item Parameter Estimates for Group 1, logit: aθ + c or a(θ – b) (Back to TOC)

Item	Label	a	c	b
1	SLF1	1.20	-2.35	1.97
2	SLF2	0.71	0.80	-1.11
3	SLF3	1.53	0.99	-0.65
4	SLF4	2.55	-0.67	0.26
5	SLF5	0.92	-1.10	1.19

The next portion of the output shows that the item scores are saved to the text file **SLF.Test-sco.txt**. Text files can be opened with any text editor such as Notepad.

Summary of the Data and Control Parameters (Back to TOC)

Sample Size	1490
Number of Items	5
Number of Dimensions	1

Scoring Control Values

Response pattern EAPs are computed

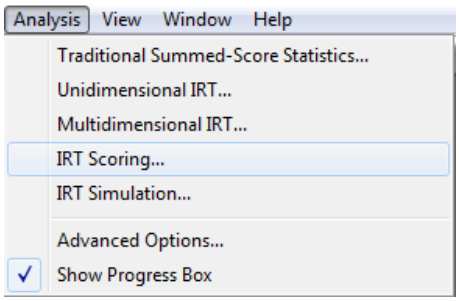
Output Files

HTML results and control parameters:	SLF.Test1-ssc.htm
Text scaled score file:	SLF.Test1-sco.txt

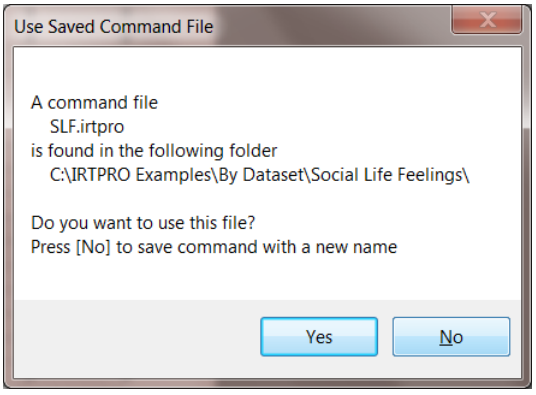
2.2 Scoring based on a parameter file

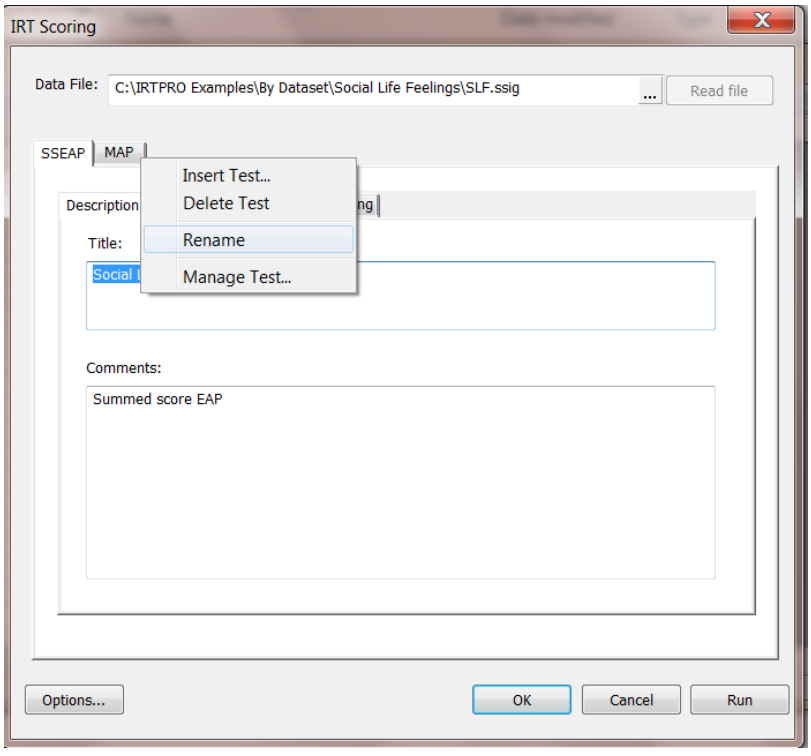
In this section, the summed-score EAP (SSEAP) and maximum a posteriori (MAP) scoring methods are considered. Use is made of the IRTPRO dataset **SLF.ssig** and the parameter estimates, obtained as described in the previous section, are read from a **–prm.txt** parameter file. Scoring is accomplished by selecting the **Analysis, IRT Scoring…** option from the main menu bar.

Start by opening the IRTPRO data file **SLF.ssig** located in the folder **IRTPRO Examples\By Dataset\Social Life Feelings\.** If this file is still open from a previous session, close it first and then re-open it, otherwise the **IRT Scoring…** option might be disabled.



By selecting **IRT Scoring…**, a **Use Save Command File** message box is displayed. Since we do not want to overwrite the existing command file (generated in Section 8.2.1), the **No** button is clicked.

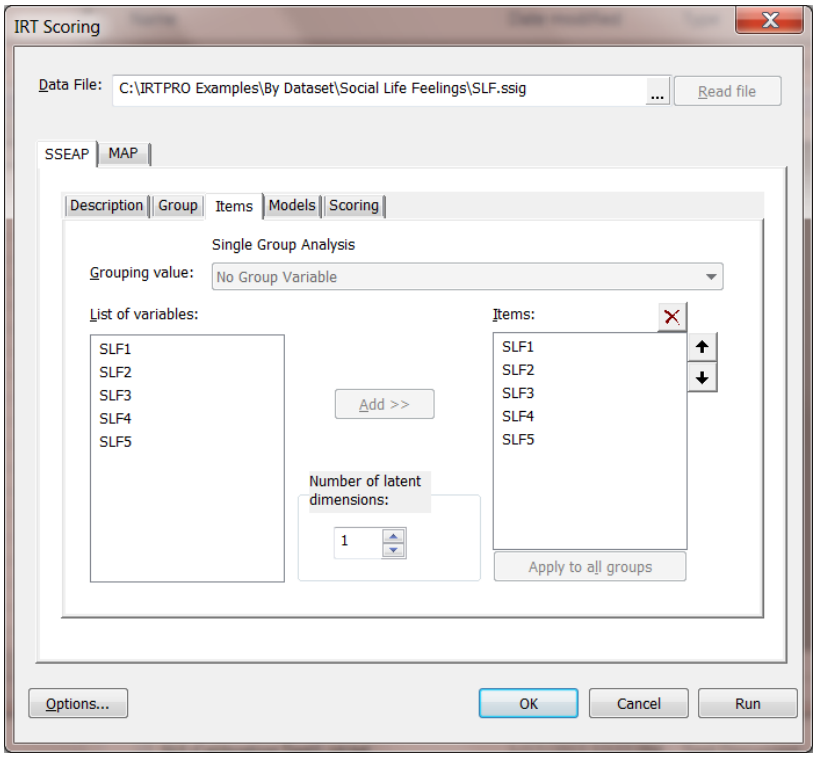




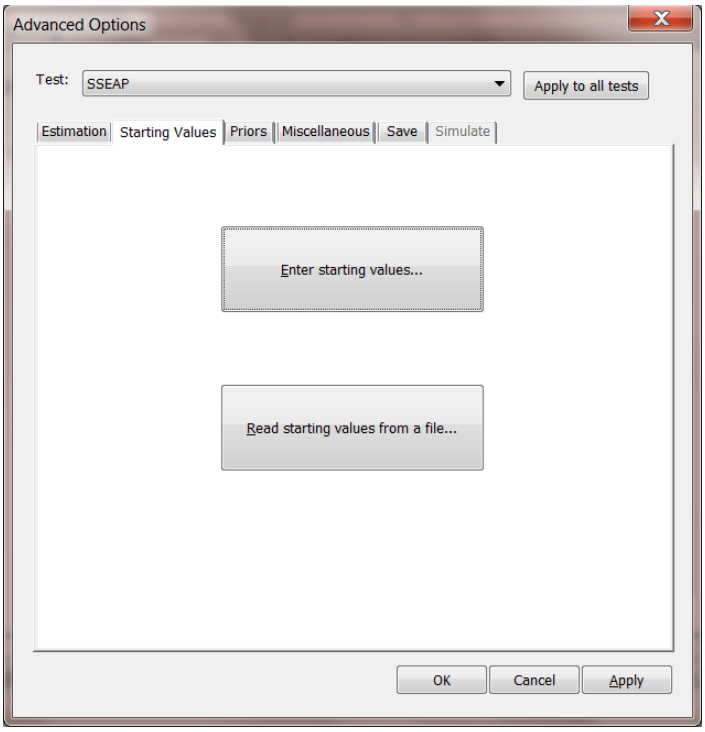
Use the **Insert Test…** and **Rename** options (obtained by right-clicking next to an existing test tab to

insert a new test or on a tab to rename a test) to insert a second test and to rename the **Test1** and **Test2** tabs to **SSEAP** and **MAP** respectively as shown above.

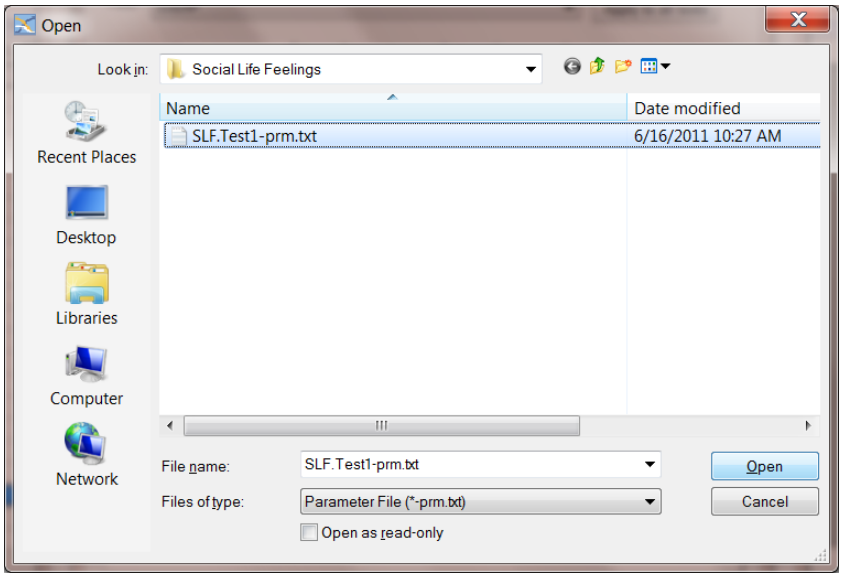
Starting with the **SSEAP** tab, enter a title and (optionally) comments as illustrated. Proceed to the **Items** tab and select the items SLF1 to SLF5.



Click the **Options…** button (lower right-hand corner in display above) to activate the **Advanced Options** window and click the **Starting Values** tab to obtain the dialog shown below.

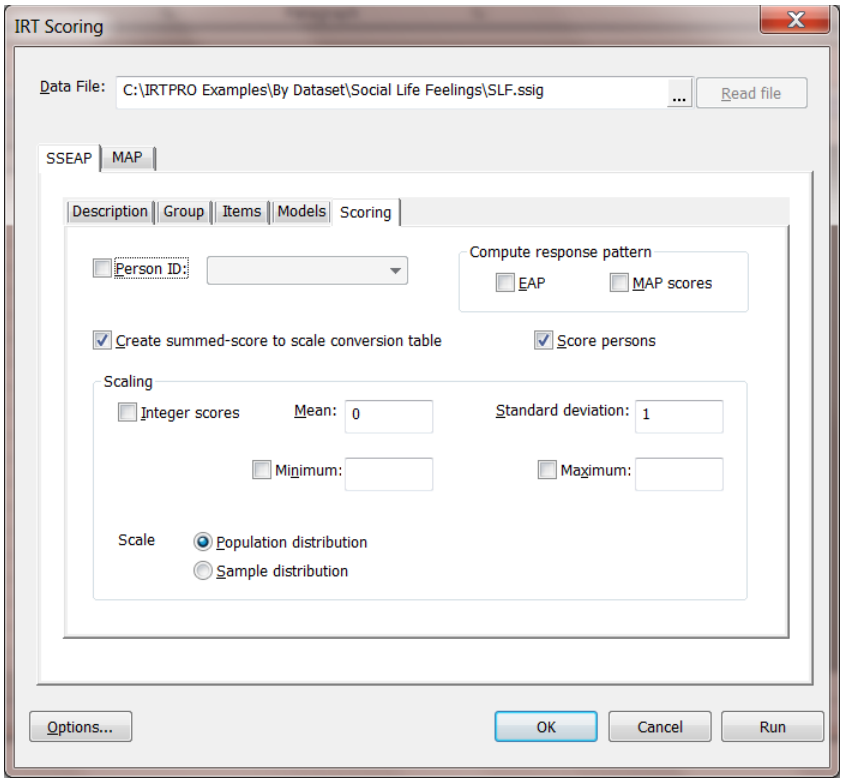


Next click the **Read starting values from a file…** button to display the **Open** dialog, then select **SLF.Test1-prm.txt**. Click the **Open** button to return to the IRT scoring menu.



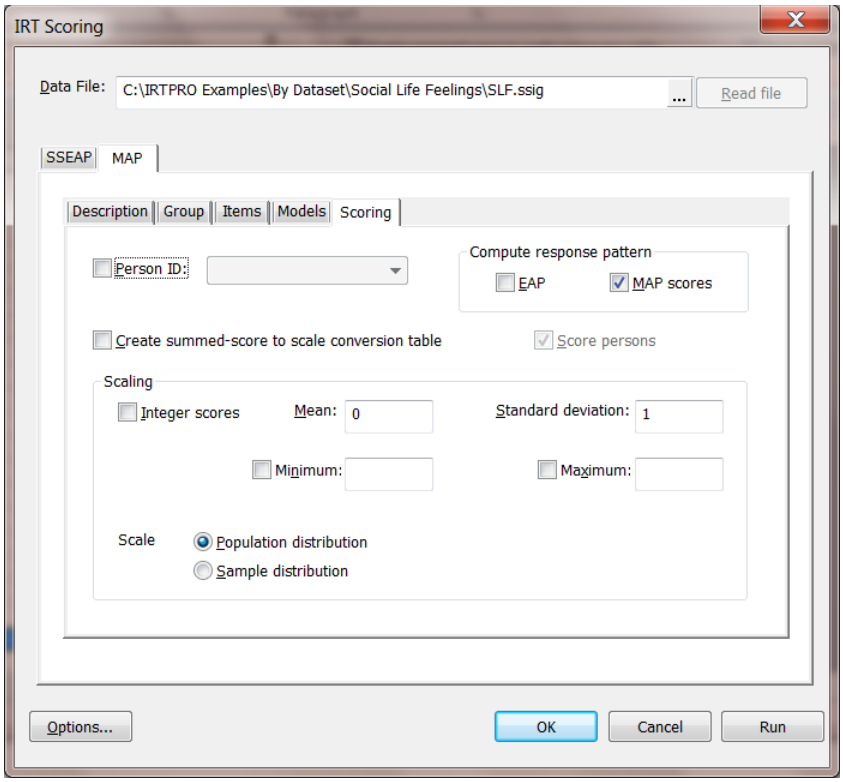
Click the **Scoring** tab and make the following selections:

- Check the **Create summed-score to scale conversion table** option.
- Check the **Score persons** option.
- Select **Scale in: Population distribution**.

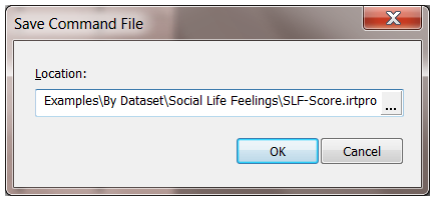


Finally, select the MAP test tab and repeat all the steps (**Description**, **Items**, **Starting values**) described above. However, the **Scoring** dialog should now contain the following selections:

- Check the **MAP scores** option.
- Select **Scale in: Population distribution**.



The scoring procedure is started by clicking the **Run** button. At this stage, the user will have the opportunity to save the command file under a new name. In this case, the default name **SLF.irtpro** is changed to **SLF-Score.irtpro** to ensure that the command file generated in Section 8.2.1 is not overwritten. Click the **OK** button to start the analysis.



Selections of the output for the SSEAP scoring procedure are shown below:

Project:	Social Life Feelings
Description:	Summed score EAP
Date:	16 June 2011
Time:	11:03 AM

Table of Contents

2PL Model Item Parameter Estimates for Group 1, logit: aθ + c or a(θ – b) **(Back to TOC)**

Item	Label	a	c	b
1	SLF1	1.20	-2.35	1.97
2	SLF2	0.71	0.80	-1.11
3	SLF3	1.53	0.99	-0.65
4	SLF4	2.55	-0.67	0.26
5	SLF5	0.92	-1.10	1.19

Summed Score to Scale Score Conversion Table **(Back to TOC)**

Summed Score	EAP[θ|x]	SD[θ|x]	Modeled Proportion
0	-1.191	0.717	0.1087594
1	-0.679	0.682	0.2286010
2	-0.110	0.652	0.2617982
3	0.511	0.649	0.2268359
4	1.022	0.653	0.1359492
5	1.544	0.701	0.0380563

Marginal reliability of the scaled scores for summed scores = 0.55444

Scoring Control Values

Scale scores for summed scores are tabulated and computed	
Summed score equivalence threshold:	0.000010

Note that the file containing the scores is saved as **SL-Score.SSEAP-sco.txt** and can be opened with any text editor.

Output Files

HTML results and control parameters:	SLF-Score.SSEAP-ssc.htm
Text scaled score file:	SLF-Score.SSEAP-sco.txt

The results shown next were obtained for the MAP scoring procedure. The parameter estimates are those obtained from the parameter file created as described in Section 8.2.1.

Project:	Social Life Feelings
Description:	Response Pattern MAP
Date:	16 June 2011
Time:	11:03 AM

Table of Contents

2PL Model Item Parameter Estimates for Group 1, logit: aθ + c or a(θ – b) **(Back to TOC)**

Item	Label	a	c	b
1	SLF1	1.20	-2.35	1.97
2	SLF2	0.71	0.80	-1.11
3	SLF3	1.53	0.99	-0.65
4	SLF4	2.55	-0.67	0.26
5	SLF5	0.92	-1.10	1.19

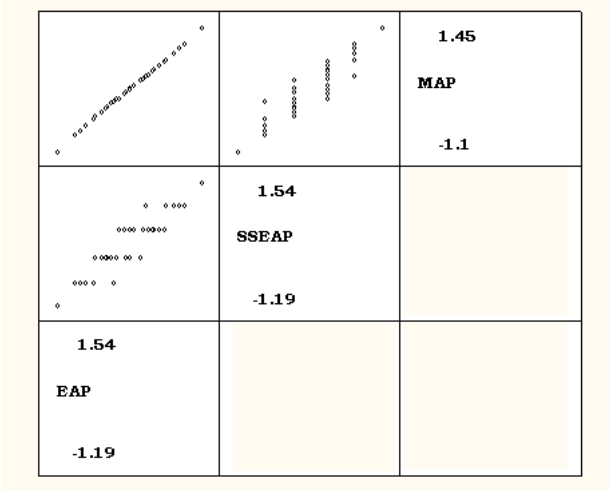
Scoring Control Values

Response pattern MAPs are computed

Output Files

HTML results and control parameters:	SLF-Score.MAP-ssc.htm
Text scaled score file:	SLF-Score.MAP-sco.txt

A matrix plot of the three sets of scores reveals an almost perfect correlation between the scores obtained with EAP and MAP. A plot of the SSEAP scores against the EAP and against the MAP scores shows strong positive correlation.



Conclusions drawn from the matrix scatter plot, are substantiated by calculating the sample statistics of the three sets of scores, the results being reported below:

Descriptive Statistics for three scoring methods

Correlation Matrix

	EAP	SSEAP	MAP
EAP	1.000		
SSEAP	0.962	1.000	
MAP	1.000	0.964	1.000

Means

EAP	SSEAP	MAP
0.000	-0.001	0.016

Standard Deviations

EAP	SSEAP	MAP
0.773	0.746	0.704

The correlations between the three scoring methods are shown below, followed by the means and standard errors of EAP, SSEAP and MAP. There is almost perfect correlation between EAP and MAP (1.000 to three decimal places).

In applied testing situations, the larger standard errors (about 10%) associated with the SSEAPs may be considered a reasonable penalty offset by the ease of summed score based IRT scoring.