Program IRT-CLASS

For Item Response Theory Classification Consistency and Accuracy
(March 2006; Revised April 2008)

Won-Chan Lee
CASMA
University of Iowa
Email: won-chan-lee@uiowa.edu

Michael J. Kolen
University of Iowa
Email: michael-kolen@uiowa.edu

Table of Contents

Page

Program IRT-CLASS

For Item Response Theory Classification Consistency and Accuracy

IRT-CLASS is a PC console version of a FORTRAN computer program that computes classification consistency and accuracy indices. IRT-CLASS is intended to be used for tests that are scaled using dichotomous, polytomous, or mixtures of different IRT models. The following IRT models are implemented in IRT-CLASS: three-parameter logistic, normal ogive graded response (Samejima, 1997), logistic graded response (Samejima, 1997), generalized partial credit (Muraki, 1997), and nominal response (Bock, 1997) models. It is assumed that each item is scored in two or more score categories using an ***integer*** score for each category. The total raw scores over all items and scale scores converted from the total raw scores are assumed to be the scores of interest. The terms, "raw scores" and "scale scores", are used here in the sense that they are expected scores based on the IRT theta distribution and item parameters.

IRT-CLASS is intended to be used to compute classification consistency and accuracy indices discussed in Lee (2008). The output of the program includes the consistency index (ϕ), the kappa coefficient, accuracy index (γ), and false positive and false negative error rates.

Input Files

IRT-CLASS requires a control card file, in which the estimation method and names of an output file and four input files are specified:

- `Estimation method` type of estimation method. Use **d** or **D** to use theta quadrature points and weights; use **p** or **P** to use individual theta estimates (see Lee, 2008 for details).
- `Output file` name of an output file.
- `Item file` name of the file that contains item scores and item parameter estimates.
- `Theta file` name of the file that contains either theta quadrature points and weights if the estimation method is specified as **d** or **D**; or individual theta estimates if the estimation method is specified as **p** or **P**.
- `Conversion file` name of the file that contains raw-to-scale score conversions.
- `Cut score file` name of the file that contains observed cut scores for computing classification consistency indices and true cut scores for computing classification accuracy indices.

An example of a control card file is provided below.

Control card file

```
d
example.out
example.itm
example.tht
example.cnv
example.cut
```

The first line of the control card file specifies that the distributional approach employing quadrature points and weights be used to estimate classification indices. As indicated in the example, the five file names must be provided in the following order: an output file, an item file, a theta file, a conversion file, and a cut score file.

Examples of each of the four input files are provided below. Note that any file names, with any extension, can be used. The example is from the analysis of a test, in which the first four items are calibrated using the three-parameter logistic model, and the last four items are calibrated using the Muraki's generalized partial credit model. Examples of item parameter files using different IRT models are provided later in this manual.

Item parameter file

```
Item Scores and Parameters
1  DI  2
 0  1
 1.7   1.25806  -0.11009   0.25271
2  DI  2
 0  1
 1.7   2.12025  -0.20963   0.42051
3  DI  2
 0  1
 1.7   1.51374   0.32304   0.13758
4  DI  2
 0  1
 1.7   1.71683   1.35308   0.23268
5  MU  3
 0  1  2
 1.7   0.45745  -0.53281   0.00000  -0.10801   0.10801
6  MU  4
 0  1  2  3
 1.7   0.38173   0.66010   0.00000  -0.95371   1.53095  -0.57723
7  MU  3
 0  1  2
 1.7   0.53245  -0.42415   0.00000  -0.83639   0.83639
8  MU  4
 0  1  2  3
 1.7   0.81579   2.02334   0.00000  -0.82618   0.55608   0.27010
```

The first line of the item parameter file contains a title. Parameters for each item are then provided in blocks of three lines. The first line for an item contains the item number, followed by a two-letter code for the IRT model used for that item, followed by the number of categories for that item. A space or tab character must be present between entries. Note that "DI" is used to indicate that the item is calibrated using a dichotomous IRT model, and in this example, the three-parameter logistic IRT model. The last four items are associated with "MU", which indicates that those items are calibrated using the generalized partial credit model**.** Note that "MU" should be used for a partial credit model, in which the discrimination parameter is all equal to 1.

The second line for an item contains the score for each category, with a space or tab in between each entry. Note that item scores must be integers and provided in an increasing order. The third line contains the item parameters. For the item parameters for the three-parameter logistic model, the first value is the scaling constant, *D*, which is typically either 1.7 or 1. The other item parameters are the discrimination *a*, difficulty *b*, and lower asymptote *c*. For the generalized partial credit model, the parameters for each item are the scaling constant *D*, discrimination parameter *a*, item difficulty (also called location) *b*, and category parameters for each category *d*.

Note that there is an alternate parameterization for this model. Consider an item having 3 categories ($K = 3$). Then, the parameters for the item can be parameterized in terms of difficulty and category parameters as: b, d_1, d_2, d_3. Alternatively, parameterization can be done in terms of item step parameters as: b_1, b_2, b_3. Typically, d_1 and b_1 are set to zero. In order to transform the step parameters to the difficulty and category parameters, the following formula can be used: $b = \left(\sum_{k=2}^{K} b_k\right) \Big/ (K-1)$, and $d_k = b - b_k$ (for $k = 2,...,K$).

Theta file

```
Theta quadrature points and weights
 -4.00 .0001
 -3.11 .0028
 -2.22 .0302
 -1.33 .1420
 -0.44 .3149
  0.44 .3158
  1.33 .1542
  2.22 .0360
  3.11 .0039
  4.00 .0002
```

The first line is a title. Since the estimation method was specified as **d** (i.e., distributional approach), each of the following records has two columns containing a theta quadrature point and a corresponding weight, separated by a space or tab. The posterior theta points and weights provided in the Phase 2 output from a PARSCALE run can be used.

If the estimation method is **p** (i.e., individual approach), the theta file requires only a single column that provides individual theta estimates. In this case, the number of rows in the theta file will be equal to the number of examinees in the sample, except for the title line.

Raw-to-scale score conversion file

```
Raw-to-scale score equivalents
   0  100
   1  100
   2  120
   3  120
   4  140
   5  160
   6  180
   7  200
   8  220
   9  240
  10  260
  11  280
  12  300
  13  300
  14  300
```

This file contains a title line followed by raw-to-scale score conversions. The conversion table must include columns for raw and scale scores. Note that records must be present for the minimum and maximum possible raw-score points consistent with the item scores and for all integer values between them with an increment of one. Even if the user is only interested in results for raw scores, the conversion table must be present. When a conversion table is not available and/or the metric of interest is the summed raw score, an identity conversion can be used. That is, the second column in the conversion file is the same as the first column for the raw scores. The results for the raw and scale scores will be identical.

Cut score file

```
Cutoff Scores
2
raw
3 10
120 260
3 10
120 260
```

The first line is a title. The second record indicates the number of cutoff scores. Note that the number of cut scores for the raw and scale scores is assumed to be the same. The

third line asks if the cut scores are expressed in terms of raw scores or theta values. Either **raw** (**Raw** or **RAW**) or **theta** (**Theta** or **THETA**) should be specified depending upon the cut-score metric. When the cut scores are specified in the metric of theta, they are transformed to the raw-score metric using test characteristic curves.

The fourth record specifies "observed" cut scores in the metric of either theta or the raw score, with a space or tab in between each entry, and the metric is determined by the choice made in the third line. The term "observed", as opposed to "true" as discussed later, means that these cut scores are ones that are used to make actual classification decisions for examinees. Note that the cut scores are inclusive—e.g., a raw score of 3 in this example is classified into the second level. The fifth record is for cut scores in the metric of scale scores. If the raw-to-scale score conversion is a one-to-one function and the scale-score cutoffs are a direct transformation of the corresponding raw-score cutoffs based on the conversion, then the classification consistency and accuracy for raw and scale scores will be identical. They differ only when several raw-score points including a raw-score cutoff are converted to a single scale-score value.

The last two lines specify "true" cut scores for the raw (or theta) and scale scores. True cut scores are expressed in the metric of the true score or true theta, and used for computing classification accuracy indices. In most cases, the observed and true cut scores are the same as specified in this example. It is not uncommon, however, that true cut scores can differ from the actual observed cut scores; for example, a testing program could have a set of true cut scores based on some national data, and determine a set of observed cut scores by making some adjustments such as rounding. Note that IRT-CLASS does ***not*** require that the cut scores in any score metric be integer values.

Program Operation

To execute IRT-CLASS, the user double-clicks IRT-CLASS.exe. A dos prompt will ask the user to type a control file name. The user then types the name of the control card file and presses return. The control file must reside in the same folder as IRT-CLASS. All the input files must be in text-only format and reside in the same folder as IRT-CLASS.exe. When execution is complete without any errors, "Successful execution!" is printed on the screen. The output file will be created in the same folder as IRT-CLASS.exe.

Output

The output file produced by IRT-CLASS can be opened and viewed by any text editing program or word processor. If the run proceeds normally, the output file contains the IRT-CLASS header, characteristics of the data, marginal classification statistics, conditional classification statistics, fitted frequency distribution of the marginal observed

scores, and reliability and standard errors of measurement. Below is the first part of the sample output including the header and characteristics of the data.

```
************************************************************************
***     IRT-CLASS: IRT Classification Consistency and Accuracy       ***
***                                                                  ***
***                          Version 2.0                             ***
***                                                                  ***
***               Won-Chan Lee and Michael J. Kolen                  ***
***                                                                  ***
***                       University of Iowa                         ***
***                                                                  ***
***                           April 2008                             ***
***                                                                  ***
***                       All Rights Reserved                        ***
************************************************************************

******************** Characteristics of the Data *********************
Number of items:   8
Number of cut scores:  2
Number of raw-score points:  15
Number of theta points:    10
```

The results for raw scores then follow. The results for raw scores consist of two blocks of output: consistency and accuracy. The results for classification consistency are provided first.

```
********************** Results For Raw Scores **********************

               ************** Consistency ***************

Observed raw cut scores
     3.00000    10.00000


Overall classification consistency

   0.06970   0.08266   0.00023
   0.08266   0.54683   0.06592
   0.00023   0.06592   0.08585
   -------   -------   -------
   0.15259   0.69541   0.15200

classification consistency (phi) = 0.70238
probability of misclassification = 0.29762
chance probability               = 0.52999
kappa                            = 0.36678


Less than or greater than cutoff  1
```

```
   0.06970   0.08289
   0.08289   0.76452
   -------   -------
   0.15259   0.84741

classification consistency (phi) = 0.83422
probability of misclassification = 0.16578
chance probability               = 0.74139
kappa                            = 0.35897


Less than or greater than cutoff  2

   0.78185   0.06615
   0.06615   0.08585
   -------   -------
   0.84800   0.15200

classification consistency (phi) = 0.86770
probability of misclassification = 0.13230
chance probability               = 0.74221
kappa                            = 0.48679
```

The results under the heading "Overall classification consistency" contain the marginal classification consistency over the theta distribution when all cut scores are applied together. The number of rows and columns (the fourth row is the sum of each column) should match the number of levels or classification categories. For the current example, there are three levels and thus the marginal classification table contains a 3x3 matrix. Each cell (*i*th row and *j*th column) gives the probability of a randomly selected examinee's (i.e., marginal probability) being classified into *i*th and *j*th category on two hypothetical replications of the test. Results for classification consistency (phi), probability of misclassification (1-phi), chance probability, and kappa coefficient are also provided. The results under the heading "Less than or greater than cutoff 1" are based on a binary classification decision using the first cut score as if there is only one cut score. Likewise, the results from applying the second cut score to a binary classification decision are provided under the heading "Less than or greater than cutoff 2." Binary classification result is produced for each of the cut scores.

Results for classification accuracy are provided next.

```
               ********** Accuracy **********

Observed raw cut scores
     3.00000    10.00000
True raw cut scores
     3.00000    10.00000
```

```
row = true classification
col = obs classification

Overall classification accuracy

   0.10297   0.07211   0.00001   |   0.17508
   0.04962   0.62187   0.11333   |   0.78482
   0.00000   0.00143   0.03866   |   0.04010
   -------   -------   -------
   0.15259   0.69541   0.15200

classification accuracy (gamma)  = 0.76349
false negative error rate        = 0.05106
false positive error rate        = 0.18545


Less than or greater than cutoff  1

   0.10297   0.07212   |   0.17508
   0.04962   0.77530   |   0.82492
   -------   -------
   0.15259   0.84741

classification accuracy (gamma)  = 0.87826
false negative error rate        = 0.04962
false positive error rate        = 0.07212


Less than or greater than cutoff  2

   0.84657   0.11334   |   0.95990
   0.00143   0.03866   |   0.04010
   -------   -------
   0.84800   0.15200

classification accuracy (gamma)  = 0.88523
false negative error rate        = 0.00143
false positive error rate        = 0.11334
```

The format of the accuracy results is similar to that of the consistency results. The accuracy results contain a bivariate table of joint probabilities of observed and true classifications. Rows represent the classifications based on examinees' estimated true scores and true cut scores, whereas columns represent the classifications based on examinees' observed scores and observed cut scores. The statistics computed include the classification accuracy (i.e., sum of diagonal elements) and false negative and false positive error rates. Results are produced for the overall classification as well as all binary classifications.

The results for scale scores are produced next, with the exact same format as the raw score results as shown below.

```
********************** Results For Scale Scores **********************

               ************** Consistency ***************

Observed scale score cut scores
   120.00000   260.00000


Overall classification consistency

   0.02098   0.05238   0.00005
   0.05238   0.65610   0.06610
   0.00005   0.06610   0.08585
   -------   -------   -------
   0.07342   0.77458   0.15200

classification consistency (phi) = 0.76294
probability of misclassification = 0.23706
chance probability               = 0.62847
kappa                            = 0.36192


Less than or greater than cutoff  1

   0.02098   0.05243
   0.05243   0.87415
   -------   -------
   0.07342   0.92658

classification consistency (phi) = 0.89513
probability of misclassification = 0.10487
chance probability               = 0.86394
kappa                            = 0.22923


Less than or greater than cutoff  2

   0.78185   0.06615
   0.06615   0.08585
   -------   -------
   0.84800   0.15200

classification consistency (phi) = 0.86770
probability of misclassification = 0.13230
chance probability               = 0.74221
kappa                            = 0.48679


               **************** Accuracy ****************
```

```
Observed scale score cut scores
   120.00000   260.00000
True scale score cut scores
   120.00000   260.00000


row = true classification
col = obs classification

Overall classification accuracy

   0.01693   0.01617   0.00000   |   0.03310
   0.05649   0.75698   0.11334   |   0.92681
   0.00000   0.00143   0.03866   |   0.04010
   -------   -------   -------
   0.07342   0.77458   0.15200

classification accuracy (gamma)  = 0.81257
false negative error rate        = 0.05793
false positive error rate        = 0.12951


Less than or greater than cutoff  1

   0.01693   0.01617   |   0.03310
   0.05649   0.91041   |   0.96690
   -------   -------
   0.07342   0.92658

classification accuracy (gamma)  = 0.92734
false negative error rate        = 0.05649
false positive error rate        = 0.01617


Less than or greater than cutoff  2

   0.84657   0.11334   |   0.95990
   0.00143   0.03866   |   0.04010
   -------   -------
   0.84800   0.15200

classification accuracy (gamma)  = 0.88523
false negative error rate        = 0.00143
false positive error rate        = 0.11334
```

The program also computes results for conditional classification consistency given theta. The conditional consistency provides useful information about how consistency varies across levels of theta values (or expected raw and scale scores).

```
***** Conditional Consistency When All Cutoffs Applied Together ******

    Theta   Exp_Raw  Exp_Scale     Raw_p   Raw_1-p   Scale_p Scale_1-p
 -4.00000   1.16395  106.78506   0.85208   0.14792   0.56021   0.43979
 -3.11000   1.30599  108.24453   0.78979   0.21021   0.52801   0.47199
 -2.22000   1.64730  111.94029   0.65784   0.34216   0.50000   0.50000
 -1.33000   2.48289  122.09275   0.50334   0.49666   0.59298   0.40702
 -0.44000   4.38571  150.96562   0.73552   0.26448   0.89898   0.10102
  0.44000   7.20514  204.17307   0.82992   0.17008   0.83591   0.16409
  1.33000   9.57230  250.65773   0.50429   0.49571   0.50429   0.49571
  2.22000  12.22861  289.94739   0.92363   0.07637   0.92363   0.07637
  3.11000  13.45456  299.33973   0.99867   0.00133   0.99867   0.00133
  4.00000  13.75910  299.94207   0.99998   0.00002   0.99998   0.00002
```

The conditional consistency results are based on the overall classification (i.e., all cutoff scores applied together). The first column shows the theta points read in from the theta input file. The second and third columns provide the expected raw and scale scores given theta. The fourth column labeled "Raw_p" gives the conditional probabilities of consistent classification for raw scores. The fifth column labeled "Raw_1-p" gives results for the conditional probabilities of misclassification for raw scores. The sixth and seventh columns are for the scale scores.

The results for conditional accuracy are produced next. As for the conditional consistency, this set of results is based on the overall classification (i.e., all cutoff scores applied together).

```
******* Conditional Accuracy When All Cutoffs Applied Together *******

     Theta    Raw_Acc     Raw_FN    Raw_FP Scale_Acc  Scale_FN  Scale_FP
  -4.00000   0.91957   0.00000   0.08043   0.67351   0.00000   0.32649
  -3.11000   0.88065   0.00000   0.11935   0.61834   0.00000   0.38166
  -2.22000   0.78093   0.00000   0.21907   0.50096   0.00000   0.49904
  -1.33000   0.54108   0.00000   0.45892   0.71564   0.28432   0.00004
  -0.44000   0.84370   0.15387   0.00243   0.94678   0.05079   0.00243
   0.44000   0.90656   0.00372   0.08972   0.90987   0.00041   0.08972
   1.33000   0.45366   0.00001   0.54633   0.45367   0.00000   0.54633
   2.22000   0.96023   0.03977   0.00000   0.96023   0.03977   0.00000
   3.11000   0.99933   0.00067   0.00000   0.99933   0.00067   0.00000
   4.00000   0.99999   0.00001   0.00000   0.99999   0.00001   0.00000
```

The first column shows the theta points read in from the theta input file. The second column provides the conditional accuracy index (gamma) given theta for raw scores. The third and fourth columns give results for the false negative and false positive error rates given theta for raw scores. The fifth through seventh columns are results for scale scores.

The fitted frequency distribution is provided next. The fitted frequency distribution could be compared with the actual observed score distribution (not computed by IRT-CLASS) to evaluate goodness of fit of a particular model.

```
************************ Fitted Frequencies ************************

          Raw       Scale Fitted_Freq
      0.00000   100.00000     0.01995
      1.00000   100.00000     0.05347
      2.00000   120.00000     0.07917
      3.00000   120.00000     0.09472
      4.00000   140.00000     0.10187
      5.00000   160.00000     0.10435
      6.00000   180.00000     0.10574
      7.00000   200.00000     0.10498
      8.00000   220.00000     0.09622
      9.00000   240.00000     0.08752
     10.00000   260.00000     0.07118
     11.00000   280.00000     0.03882
     12.00000   300.00000     0.01492
     13.00000   300.00000     0.01554
     14.00000   300.00000     0.01154
```

The last piece of information computed by the program is reliability and conditional standard errors of measurement (CSEMs) for raw and scale scores. The marginal results are provided first, and then the results for the CSEMs follow.

```
************************** Conditional SEMs **************************

Overall error variance for raw scores    =     2.84563
True score variance for raw scores       =     7.26337
Observed score variance for raw scores   =    10.10900
Reliability for raw scores               =     0.71851

Overall error variance for scale scores  =   915.99535
True score variance for scale scores     =  2333.01009
Observed score variance for scale scores =  3249.00543
Reliability for scale scores             =     0.71807

        Theta     Exp_Raw   Exp_Scale    Raw_CSEM  Scale_CSEM
     -4.00000     1.16395   106.78506     0.92901    10.04582
     -3.11000     1.30599   108.24453     1.01753    11.21284
     -2.22000     1.64730   111.94029     1.20674    14.09738
     -1.33000     2.48289   122.09275     1.53100    20.85554
     -0.44000     4.38571   150.96562     1.83183    32.12243
      0.44000     7.20514   204.17307     1.74204    34.61991
      1.33000     9.57230   250.65773     1.55192    29.41049
      2.22000    12.22861   289.94739     1.46551    17.54413
      3.11000    13.45456   299.33973     0.77158     4.27242
      4.00000    13.75910   299.94207     0.48910     1.15118
```

For each of the raw and scale scores, the marginal results include the overall error variance (i.e., integrated conditional error variances over the theta distribution), true score variance, observed score variance, and reliability. Note that the observed score variance is computed based on the model--it is not the sample variance. The expected raw scores, expected scale scores, raw-score CSEMs, and scale-score CSEMs are computed for each theta value.

Examples of Other IRT Models

Samejima's logistic graded response model is coded "SL." The parameters for each item are the scaling constant, *D*, the discrimination parameter, *a*, and difficulty parameters for the second through the last category. An example for a two-item test using this model is as follows:

```
Samejima's Logistic graded response model
1 SL 5
 0 1 2 3 4
 1.7  1.2  -.5  .6  1.1  1.3
2 SL 3
 0 1 2
 1.7  1.0  0.0  1.0
```

Note that the first item is a 5-category item, with the categories scored 0-4. The items use a scaling constant $D = 1.7$. The discrimination parameter for the first item is 1.2 and the four step parameters are -.5, .6, 1.1, and 1.3. The number of step parameters for the graded response model is one less than the number of categories. If the parameter estimates for this model are available in the form of the item location and category parameters, $b, d_2, ..., d_K$ (as done by PARSCALE), they can be transformed to the step parameters as follows: $b_2 = b - d_2, ..., b_K = b - d_K$.

Samejima's normal ogive graded response model is coded "SN." The parameters for each item are the discrimination parameter, *a*, and difficulty parameters for the second through the last category. An example for a two-item test using this model is as follows:

```
Samejima's Normal ogive graded response model
1 SN 5
 0 1 2 3 4
 1.2  -.5  .6  1.1  1.3
2 SN 3
 0 1 2
 1.0  0.0  1.0
```

Unlike the logistic model, no scaling constant is used with the normal ogive model.

Bock's nominal model is coded "BN." The parameters for each item are the scaling constant *D* (either 1 or 1.7), slope parameters for each category *a*, and the intercept parameter for each category *c*. If the slope parameters are increasing over categories, then Bock's nominal model is associated with a set of graded categories. An example for a two-item test using this model is as follows:

```
Bock's Nominal model
1 BN 4
 1 2 3 4
 1.0  1.7  3.4  5.1  6.8  0.0  2.55  -.85  -3.06
1 BN 2
 1 2
 1.0  1.7  3.4  1.1  -0.5
```

Note that the first item is a four category item and the second item is a dichotomous item.

Issues

Depending upon the item scoring, some raw-score points between the minimum and maximum raw scores may not be defined. However, those indefinable raw-score points must be present in the conversion file. They do not affect the final raw-score results because the probabilities associated with the undefinable raw-score points will be zero anyway. Since there are no scale scores corresponding to the undefinable raw-score points, pseudo scale-score values can be used in a way that it does not affect the scale-score results. For example, consider a two-item test when each item has two categories scored 1 and 3, which results in only three possible raw-score points, 2, 4, and 6. Suppose the scale-score cutoff is 2. Further, suppose the raw score of 2 converts to the scale score of 1, 4 converts to 2, and 6 converts to 3. In this case, the following conversion can be used. Note that the pseudo values are boldfaced and italicized.

```
Raw-to-scale score equivalents (with some pseudo values)
   2  1
   3  1.5
   4  2
   5  2.2
   6  3
```

Any integer-value scoring functions can be used with the items, including negative values. However, the minimum and maximum values in the conversion file must be consistent with item scoring.

Currently, the program employs the maximum numbers of items (500), cut scores (20), theta quadrature points (500), individual theta values (50,000), score categories per item

(120), and raw-score points (1000). The maximum numbers are sufficiently large for most practical purposes. However, if the user needs larger numbers than the program can provide, please contact the authors. All other problems should also be directed to the authors.

References

Bock, R. D. (1997). The nominal categories model. In W. J. van der Linden and R. K. Hambleton (Eds.) *Handbook of modern item response theory* (pp. 33-49). New York: Springer-Verlag.

Lee, W. (2008). *Classification consistency and accuracy for complex assessments using item response theory* (CASMA Research Report No. 27). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa. (Available from http://www.education.uiowa.edu/casma).

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden and R. K. Hambleton (Eds.) *Handbook of modern item response theory* (pp. 85-100). New York: Springer-Verlag.

Samejima, F. (1997). Graded response model. In W. J. van der Linden and R. K. Hambleton (Eds.) *Handbook of modern item response theory* (pp. 153-164). New York: Springer-Verlag.