Item calibration and examinee Bayes scoring with the rating-scale graded model

This example illustrates calibration and scoring of a test or scale containing 20 multiple category items. The simulated data represent responses of 1000 examinees drawn randomly from a population with a mean trait score of 0.0 and standard deviation of 1.0.

Data are read from the file **exampl01.dat** in the **examples** folder using the DFNAME keyword on the FILES command. The first few lines of the data file are shown below. The generating trait value of each examinee is the second column of information in the data file. The case ID, given at the beginning of each line, is 4 characters long and is indicated as such using the NIDCHAR keyword on the INPUT command. It is also reflected in the format statement as 4A1.

```
0001   .44739 42444232223343433332
0002  -.93465 12221121122324121432
0003  -.56465 32212212213342314121
0004  -.58622 13222111113224221111
0005  -.35223 21211122313132312131
```

All 20 items are used in a single test (NTEST=1 on INPUT command, with LENGTH=20). All 20 items have common categories and are assigned to the same BLOCK (NBLOCK=1 on TEST; NITEMS=20 on BLOCK).

All items have four categories (NCAT=4 on BLOCK command) and varying difficulties and discriminating powers. The graded model is assumed (GRADED on CALIB command); and a logistic response model (LOGISTIC on CALIB command) is requested. The choice between a logistic or normal response function metric is effective only if the graded response model is used. The response function of the graded model can be either the normal ogive or its logistic approximation. Graded is the default. If logistic is selected, the item parameters can be in the natural or the logistic metric. Natural is the default. For the normal metric, set SCALE equal to 1.7. Neither LOGISTIC nor SCALE is needed when PARTIAL is selected. Because the generalized model allows for varying item discriminating powers, both a slope and threshold are estimated for each item. The CADJUST keyword on the BLOCK command is used to set the mean of the category parameters to 0 as simultaneous estimation of slope parameters and all category parameters is not obtainable.

The ITEMFIT keyword is used to set the number of frequency score groups for the computation of item fit statistics to 10. Note that there is no default value for the ITEMFIT keyword.

The CYCLES keyword specifies 25 EM iterations, with maximum 2 inner EM iterations for the item and category parameter estimation. Five Newton-Gauss iterations are requested (NEWTON=5 on CALIB). A convergence criterion of 0.005 is specified by using the CRIT keyword on CALIB.

30 quadrature points are to be used in the EM and Newton estimation instead of the default of 10 for cases where LENGTH less or equal to 50 in the INPUT command. The calibration procedure depends on the evaluation of integrals using Gauss-Hermite quadrature. In general, the accuracy of numerical integration increases with the number of quadrature points used.

The score estimation method is specified (EAP option on SCORE command). Scale scores for each subtest are estimated by the Bayes (EAP) method, and their posterior standard deviations serve as standard errors.

The scores, which are rescaled to zero mean and unit standard deviation in the sample (SMEAN and SSD on SCORE), are saved in the file **exampl01.sco** using the SCORE keyword on the SAVE command. The PFQ keyword is specified. This keyword is usually used to make ML scores more computable but would also improve the EAP estimates somewhat. In addition, the estimated item parameters are saved in the file **exampl01.par** (PARM keyword on the SAVE command).

The command file is shown below, with comments omitted.

```
EXAMPL01.PSL: ARTIFICIAL EXAMPLE:  MONTE CARLO DATA
GRADED RATING SCALE MODEL, NORMAL RESPONSE FUNCTION:  EAP SCALE SCORES
>FILES   DFNAME='EXAMPL01.DAT',SAVE;
>SAVE    PARM='EXAMPL01.PAR',SCORE='EXAMPL01.SCO';
>INPUT   NIDCHAR=4,NTOTAL=20,NTEST=1,LENGTH=(20),NFMT=1;
(4A1,10X,20A1)
>TEST1   TNAME=SCALE1,ITEM=(1(1)20),NBLOCK=1;
>BLOCK1 BNAME=SBLOCK1,NITEMS=20,NCAT=4, CADJUST=0.0;
>CAL     GRADED,NQPTS=30,CYCLE=(25,2,2,2,2),
         NEWTON=5,CRIT=0.005,ITEMFIT=10;
>SCORE   EAP,NQPTS=30,SMEAN=0.0,SSD=1.0,NAME=EAP,PFQ=5;
```

Phase 0 output

At the beginning of the output for Phase 0, the command file is echoed. Information on the number of tests, items, and type of model to be fitted as interpreted by PARSCALE is also given.

```
SINGLE MAIN TEST IS USED.
 NUMBER OF ITEMS:    20

 FORMAT OF DATA INPUT IS
 (4A1,10X,20A1)

 >TEST1   TNAME=SCALE1,ITEM=(1(1)20),NBLOCK=1;
```

```
BLOCK CARD:   1
>BLOCK1 BNAME=SBLOCK1,NITEMS=20,NCAT=4,CADJ=0.0;
>CAL    GRADED,LOGISTIC,SCALE=1.7,NQPTS=30,CYCLE=(25,2,2,2,2),
        NEWTON=5,CRIT=0.005,ITEMFIT=10;

MODEL SPECIFICATIONS
====================

LOGISTIC - GRADED ITEM RESPONSE MODEL IS SPECIFIED.
           SCALE CONSTANT  1.70 FOR SLOPE PARAMETERS.
```

This section of the output file contains information on the settings to be used during the item parameter estimation in Phase 2.

```
CALIBRATION PARAMETERS
======================

MAXIMUM NUMBER OF EM CYCLES:                  25
MAXIMUM INNER EM CYCLES:                       2
MAXIMUM CATEGORY ESTIMATION CYCLES:            2
MAXIMUM ITEM PARAMETER ESTIMATION CYCLES:      2
MAXIMUM NUMBER OF NEWTON CYCLES:               2
CONVERGENCE CRITERION FOR EM CYCLES:         0.0050
CONVERGENCE CRITERION FOR SLOPE:             0.0050
CONVERGENCE CRITERION FOR THRESHOLD:         0.0050
CONVERGENCE CRITERION FOR CATEGORY:          0.0050
CONVERGENCE CRITERION FOR GEUSSING:          0.0050
ORDER OF INNER EM CYCLES:                   CATEGORY - ITEM PARAMETERS
ESTIMATION ACCELERATOR:                     NO (DEFAULT)
RIDGE METHOD:                               NO (DEFAULT)
```

No prior distribution was requested in the CALIB command, and consequently the default prior, a normal distribution on equally spaced points, will be used (DIST=2 on CALIB). The number of quadrature points to be used during item parameter estimation was set to 30 (NQPT on CALIB). The program-generated quadrature points and weights are printed to the Phase 0 output file, as shown below.

```
THE FIXED PRIOR DISTRIBUTION FOR LATENT TRAITS
                                          MEAN     :  0.0000
                                          S.D.     :  1.0000

QUADRATURE POINTS AND PRIOR WEIGHTS (PROGRAM-GENERATED NORMAL APPROXIMATION):

               1           2           3           4           5
POINT    -0.4000E+01 -0.3724E+01 -0.3448E+01 -0.3172E+01 -0.2897E+01
WEIGHT    0.3692E-04  0.1071E-03  0.2881E-03  0.7181E-03  0.1659E-02

               6           7           8           9          10
POINT    -0.2621E+01 -0.2345E+01 -0.2069E+01 -0.1793E+01 -0.1517E+01
WEIGHT    0.3550E-02  0.7042E-02  0.1294E-01  0.2205E-01  0.3481E-01

              11          12          13          14          15
POINT    -0.1241E+01 -0.9655E+00 -0.6897E+00 -0.4138E+00 -0.1379E+00
WEIGHT    0.5093E-01  0.6905E-01  0.8676E-01  0.1010E+00  0.1090E+00
```

```
              16          17          18          19          20
POINT     0.1379E+00  0.4138E+00  0.6897E+00  0.9655E+00  0.1241E+01
WEIGHT    0.1090E+00  0.1010E+00  0.8676E-01  0.6905E-01  0.5093E-01

              21          22          23          24          25
POINT     0.1517E+01  0.1793E+01  0.2069E+01  0.2345E+01  0.2621E+01
WEIGHT    0.3481E-01  0.2205E-01  0.1294E-01  0.7042E-02  0.3550E-02



              26          27          28          29          30
POINT     0.2897E+01  0.3172E+01  0.3448E+01  0.3724E+01  0.4000E+01
WEIGHT    0.1659E-02  0.7181E-03  0.2881E-03  0.1071E-03  0.3692E-04

TOTAL WEIGHT: 1.00000
MEAN        : 0.00000
S.D.        : 0.99970
```

The control settings to be used during calibration are followed by settings to be used during the scoring phase (Phase 3). The EAP method of scoring is requested (EAP option) and, as in the calibration phase, 30 quadrature points were requested. Since no prior distribution was requested using the DIST keyword, by default a normal distribution on equally spaced points will be used (DIST=2 on SCORE). Note that the DIST keyword applies only when EAP scoring has been selected.

```
>SCORE  EAP,NQPTS=30,SMEAN=0.0,SSD=1.0,NAME=EAP,PFQ=5;

PARAMETERS FOR SCORING AND TEST AND ITEM INFORMATION
====================================================

METHOD OF SCORING SUBJECTS:                EXPECTATION A POSTERIORI
                                           (EAP; BAYES ESTIMATES)

TYPE OF PRIOR:                             NORMAL APPROXIMATION

NUMBER OF QUADRATURE POINTS                 30
SCORES WRITTEN TO FILE                     EXAMPL01.SCO

QUADRATURE POINTS AND PRIOR WEIGHTS (PROGRAM-GENERATED NORMAL APPROXIMATION):

               1           2           3           4           5
POINT    -0.4000E+01 -0.3724E+01 -0.3448E+01 -0.3172E+01 -0.2897E+01
WEIGHT    0.3692E-04  0.1071E-03  0.2881E-03  0.7181E-03  0.1659E-02

               6           7           8           9          10
POINT    -0.2621E+01 -0.2345E+01 -0.2069E+01 -0.1793E+01 -0.1517E+01
WEIGHT    0.3550E-02  0.7042E-02  0.1294E-01  0.2205E-01  0.3481E-01

              11          12          13          14          15
POINT    -0.1241E+01 -0.9655E+00 -0.6897E+00 -0.4138E+00 -0.1379E+00
WEIGHT    0.5093E-01  0.6905E-01  0.8676E-01  0.1010E+00  0.1090E+00

              16          17          18          19          20
POINT     0.1379E+00  0.4138E+00  0.6897E+00  0.9655E+00  0.1241E+01
WEIGHT    0.1090E+00  0.1010E+00  0.8676E-01  0.6905E-01  0.5093E-01
```

```
              21          22          23          24          25
POINT     0.1517E+01  0.1793E+01  0.2069E+01  0.2345E+01  0.2621E+01
WEIGHT    0.3481E-01  0.2205E-01  0.1294E-01  0.7042E-02  0.3550E-02

              26          27          28          29          30
POINT     0.2897E+01  0.3172E+01  0.3448E+01  0.3724E+01  0.4000E+01
WEIGHT    0.1659E-02  0.7181E-03  0.2881E-03  0.1071E-03  0.3692E-04

TOTAL WEIGHT: 1.00000
MEAN        : 0.00000
S.D.        : 0.99970
```

The values assigned to the rescaling constants SMEAN and SSD in the SCORE command are shown:

```
SET NUMBER      :    1
SCORE NAME      : EAP
NUMBER OF ITEMS :   20
RESCALE CONSTANT: MEAN =       0.00   S.D. =       1.00

ITEMS           :   1    2    3    4    5    6    7    8    9   10
                   11   12   13   14   15   16   17   18   19   20

                  0001 0002 0003 0004 0005 0006 0007 0008 0009 0010
                  0011 0012 0013 0014 0015 0016 0017 0018 0019 0020
```

Input and output files as requested with the DFNAME keyword on the FILES command and the PARM and SCORE keywords on the SAVE command are listed:

```
FILE ASSIGNMENTS AND DISPOSITIONS
=================================

[INPUT FILES]

SUBJECT DATA INPUT FILE                    EXAMPL01.DAT
                                           SINGLE-SUBJECT DATA
                                           NO CASE WEIGHTS


[OUTPUT FILES]

ITEM PARAMETERS FILE                       EXAMPL01.PAR
SUBJECT SCALE-SCORE FILE                    EXAMPL01.SCO

[SCRATCH FILES]

PARSCALE SYSTEM BINARY DATA FILE           Exampl01.MFL
TEMPORARY FILE                             Exampl01.T99
TEMPORARY FILE                             Exampl01.T98
TEMPORARY FILE                             Exampl01.T97
TEMPORARY FILE                             Exampl01.T96
```

To allow the user to verify that data have been read in correctly from the raw data file, the first two records from the data file are echoed in the output. The INPUT RESPONSES fields give the original responses while the RECODED RESPONSES reflect any recoding of the responses.

Recoding of responses is controlled by the ORIGINAL and MODIFIED keywords on the BLOCK command.

```
 INPUT AND RECODED RESPONSE OF FIRST AND SECOND OBSERVATIONS

 OBSERVATION #      1
 GROUP:  1
 ID: 0001
INPUT RESPONSES:  4  2  4  4  4  2  3  2  2  2  3  3  4  3  4  3  3  3  3  2
RECODED RESPONSES:4  2  4  4  4  2  3  2  2  2  3  3  4  3  4  3  3  3  3  2

 OBSERVATION #      2
 GROUP:  1
 ID: 0002
INPUT RESPONSES:  1  2  2  2  1  1  2  1  1  2  2  3  2  4  1  2  1  4  3  2
RECODED RESPONSES:1  2  2  2  1  1  2  1  1  2  2  3  2  4  1  2  1  4  3  2
```

Finally, the number of observations to be used in the analysis is recorded. By default, all observations will be used. The number of observations to be used can be manipulated using the SAMPLE or TAKE keywords on the INPUT command.

```
      1000 OBSERVATIONS READ FROM FILE:   EXAMPL01.DAT
      1000 OBSERVATIONS WRITTEN TO FILE:  Exampl01.MFL
```

Phase 1 output

The title given in the TITLE command and name assigned to the test in the TEST command in the command file are echoed in the output file.

```
EXAMPLE 1: ARTIFICIAL EXAMPLE:  MONTE CARLO DATA
           GRADED MODEL, NORMAL METRIC:  EAP SCALE SCORES

 MAINTEST: SCALE1
```

The master file created during Phase 0 is used as input. Note that the master file **exampl01.mfl** may be saved using the MASTER keyword on the SAVE command for use as input in a subsequent analysis (MFNAME keyword on the FILES command). The keywords TAKE and SAMPLE on the INPUT command control the number of records read from the raw data file. As the default value of SAMPLE is 100%, neither keyword was used and all data were used by default.

```
 1000 OBS.(WEIGHTS:  1000.000) WERE READ FROM Exampl01.MFL
```

Summary item statistics for the 20 items are given next. Since no not-represented (NFNAME on FILES) or omit key (OFNAME on FILES) was used, no frequencies or percentages are reported under the "NOT PRESENT" or "OMIT" headings. Under the "CATEGORIES" heading, frequencies and percentages of responses for each of the 4 categories are given item-by-item. Cumulative frequencies and percentages for the categories over all items are given at the end of the table.

Note that, if empty categories are encountered, the user has to recode the corresponding items on which this occurs before proceeding with the analysis.

```
BLOCK NO.:   1     NAME: SBLOCK1
 ------------------------------------------------------------------
 ITEM    | TOTAL    NOT     OMIT |          CATEGORIES
         |        PRESENT        |
         |                       |   1       2       3       4
 ------------------------------------------------------------------
 0001    |                       |
   FREQ.|   1000       0       0|    194     303     313     190
   PERC.|            0.0     0.0|   19.4    30.3    31.3    19.0
         |                       |


0002    |                       |
   FREQ.|   1000       0       0|    204     284     310     202
   PERC.|            0.0     0.0|   20.4    28.4    31.0    20.2
       …
 0020    |                       |
   FREQ.|   1000       0       0|    305     211     212     272
   PERC.|            0.0     0.0|   30.5    21.1    21.2    27.2
         |                       |
 ------------------------------------------------------------------
 CUMMUL.|                       |
   FREQ.|                       |   4844    5186    5204    4766
   PERC.|                       |   24.2    25.9    26.0    23.8
 ------------------------------------------------------------------
```

Item means, initial slope estimates, and Pearson and polyserial item-test correlations are shown in the next table.

Pearson

The sample product-moment correlation of the test score,

$$t_i = \sum_{j=1}^{J} s_{ij},$$

and *m*-category polytomous item score, $s_{ij} = 1, 2, ..., m$, is the point polyserial correlation $r_{PP,j}$, where

$$r_{PP,j} = \frac{\sum_{i=1}^{n} t_i s_{ij} - n\bar{t}\bar{s_j}}{(\sum_{i=1}^{n} t_i^2 - n\bar{t^2})(\sum_{i=1}^{n} s_{ij}^2 - n\bar{s_j^2})}$$

where *n* is the sample size, $\bar{t}$ is the mean test score and $\bar{s_j}$, the mean item score. In this example *n* = 1000. For item 1,

$$\sum s_{i1} = (1\times194) + (2\times303) + (3\times313) + (4\times190),$$

so that

$$\overline{s}_1 = \frac{\sum s_{i1}}{n} = \frac{2502}{1000} = 2.502.$$

Also

$$\sum s_{i1}^2 = (1^2 \times 194) + (2^2 \times 303) + (3^2 \times 313) + (4^2 \times 190) = 7263$$

so that

$$S.D.(item\ 1) = \sqrt{\frac{7263 - (1000 \times 2.502^2)}{1000}}$$

$$1.0015$$

Polyserial correlation

The polyserial correlation r_P can be expressed in terms of the point polyserial correlation as

$$r_{P,j} = \frac{r_{PP,j}\sigma_j}{\sum_{k=1}^{m-1} h(z_{jk})}$$

where

- z_{jk} is the scoring corresponding to the cumulative proportion, p_{jk} of the *k*-th response category to item *j* (for item 1, for example, the cumulative proportions are 0.194, 0.497, and 0.81 for categories 1,2, and 3), σ_j is the standard deviation of item scores for item *j* (1.0015 for item 1), and $r_{PP,j}$ is the point-polyserial correlation.
- $h(z_{jk})$ is the ordinate of the normal distribution at the point z_{jk}; that is

$$h(z_{jk}) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} z_{jk}^2).$$

Initial slopes and location

The polyserial correlation estimates the item factor loading, α_j, say. If the arbitrary scale of the item latent variable, y_j, is chosen so that the variance y_j equals 1, then

$$y_j = \alpha_j(\theta - b_{jk}) + \varepsilon_j,$$

where θ is the factor score with mean 0 and variance 1, and the error, ε_j, has mean 0 and variance $1-r_{p,j}^2$.

For purposes of MML parameter estimation in IRT, it is convenient to rescale the item latent variable so that the error variance equals 1. The factor loading then becomes the item slope,

$$a_j = r_{p,j} / \sqrt{1-r_{p,j}^2}.$$

This provisional estimate of the slope is then used as the starting value in the iterative EM solutions of the marginal maximum likelihood equations for estimating the parameters of the polytomous item response models. The initial locations shown in the last column of the table are the averages of the category thresholds for each item.

Initial item-category threshold parameters

Item-category threshold parameters can be calculated once the polyserial coefficients have been obtained. The expression for the threshold parameter in terms of the cumulative category proportions and the biserial correlation coefficient (Lord & Novick, 1968) is

$$\hat{b}_{jk} = \frac{\hat{z}_{jk}}{r_{B,j}}$$

with $r_{B,j}$ the biserial correlation for item j and $\hat{z}_{jk}$ the z score that cuts off $\hat{p}_{jk}$ proportion of the cases to item j in a unit-normal distribution; that is

$$p_k = \frac{\sum_{j=1}^{n} n_{jk}}{\sum_{j=1}^{n}\sum_{k=1}^{m} n_{jk}}.$$

where n_{jk} is the frequency of the categorical response for item j and category k. These provisional thresholds of the categories serve as starting values in MML estimation of the corresponding item parameters. For the rating-scale model, whether or not all items have the same thresholds, the category proportions are computed from frequencies accumulated over all items; *i.e.*,

$$p_k = \frac{\sum_{j=1}^{n} n_{jk}}{\sum_{j=1}^{n}\sum_{k=1}^{m} n_{jk}}.$$

In Muraki's (1990) formulation of the rating-scale model, the category threshold parameter, c_k, is expressed as a deviation from the item threshold parameter, b_j; that is

$$y_j = \alpha(\theta - b_j + c_k) + \varepsilon_j$$

under the constraint that $\sum_{j=k-1}^{m-1} c_j = 0.$

In the context of the rating-scale model, b_j is referred to as a "location" parameter. The INITIAL LOCATION column provides the values of the average of the category thresholds for each item.

BLOCK ITEM	RESPONSE MEAN / S.D.*	TOTAL SCORE MEAN / S.D.*	PEARSON & POLYSERIAL CORRELATION	INITIAL SLOPE	INITIAL LOCATION
SBLOCK1					
1 0001	2.499	49.892	0.778	1.488	-0.017
	1.009*	14.754*	0.830		
2 0002	2.510	49.892	0.797	1.628	-0.036
	1.030*	14.754*	0.852		
3 0003	2.481	49.892	0.785	1.545	0.013
	1.031*	14.754*	0.839		
4 0004	2.515	49.892	0.805	1.695	-0.053
	1.037*	14.754*	0.861		
5 0005	2.511	49.892	0.811	1.739	-0.038
	1.032*	14.754*	0.867		
6 0006	2.137	49.892	0.728	1.293	0.837
	1.037*	14.754*	0.791		
7 0007	2.118	49.892	0.735	1.336	0.855
	1.033*	14.754*	0.801		
8 0008	2.144	49.892	0.754	1.426	0.758
	1.029*	14.754*	0.819		
9 0009	2.136	49.892	0.736	1.329	0.830
	1.029*	14.754*	0.799		
10 0010	2.128	49.892	0.730	1.293	0.882
	1.002*	14.754*	0.791		
11 0011	2.870	49.892	0.645	0.985	-1.168
	1.041*	14.754*	0.702		
12 0012	2.874	49.892	0.655	1.029	-1.094
	1.071*	14.754*	0.717		
13 0013	2.874	49.892	0.690	1.144	-1.017
	1.053*	14.754*	0.753		
14 0014	2.831	49.892	0.673	1.072	-0.953
	1.057*	14.754*	0.731		
15 0015	2.847	49.892	0.679	1.114	-0.938
	1.094*	14.754*	0.744		
16 0016	2.492	49.892	0.590	0.839	0.010

```
          |     1.161*     14.754*  |     0.643   |
   17 0017 |     2.541      49.892   |     0.548   |     0.738      -0.173
          |     1.125*     14.754*  |     0.594   |
   18 0018 |     2.463      49.892   |     0.589   |     0.834       0.102
          |     1.152*     14.754*  |     0.641   |
   19 0019 |     2.470      49.892   |     0.573   |     0.798       0.085
          |     1.160*     14.754*  |     0.624   |
   20 0020 |     2.451      49.892   |     0.583   |     0.830       0.048
          |     1.184*     14.754*  |     0.639   |
 ------------------------------------------------------------------------
  CATEGORY |            |    MEAN    |     S.D.    | PARAMETER
     1     |            |    36.116  |    10.656   |     0.927
     2     |            |    46.091  |    11.156   |     0.002
     3     |            |    54.107  |    11.165   |    -0.930
     4     |            |    63.427  |    10.739   |     0.000
 ------------------------------------------------------------------------
```

At the end of this table, descriptive statistics for the raw total scores of examinees who responded in each of the 4 categories are given. The highest average total score of 63.427 was for respondents who responded in the 4th category.

Phase 2 output

An MML approach is used for estimation, and either a normal or empirical latent distribution with mean 0 and standard deviation 1 is assumed. The type of distribution used is controlled by the DIST keyword on the CALIB command. By default, a normal distribution with equally spaced points is used and, for analyses where the LENGTH keyword on the INPUT command is set to a value less than or equal to 50, 10 quadrature points will be used.

Because of the potentially wide spacing of category boundary parameters on the latent dimension, it is advisable to use a greater number of quadrature points than in BILOG-MG. In this example, the number of quadrature points was set to 30 (NQPT on the CALIB command).

The EM algorithm is used in the solution of the maximum likelihood equations for parameters, starting from the initial values described in the Phase 1 output. At each iteration, the -2 ln L is given, along with information on the parameter for which the largest change between cycles was observed. The number of EM cycles is controlled by the CYCLE keyword on the CALIB command, and the convergence criterion may be set using the CRIT keyword on the same command. By default, 10 EM cycles would be performed when LENGTH≤50 on the INPUT command. In this example, 25 EM cycles with a maximum of 2 inner EM iterations for the item and category parameter estimation were specified. The default convergence criterion is 0.001. For this example, it was set to 0.005.

```
[E-M CYCLES]   GRADED RESPONSE MODEL

CATEGORY AND ITEM PARAMETERS AFTER CYCLE    0

   LARGEST CHANGE=  0.000

   -2 LOG LIKELIHOOD =      46371.421

 CATEGORY AND ITEM PARAMETERS AFTER CYCLE    1
```

```
  LARGEST CHANGE=  0.636 ( -1.168-> -0.532) at Location of Item: 11 0011

  -2 LOG LIKELIHOOD =      44229.018

 CATEGORY AND ITEM PARAMETERS AFTER CYCLE    2

  LARGEST CHANGE=  0.033 (  0.989->  1.022) at Slope    of Item: 13 0013

  -2 LOG LIKELIHOOD =      44224.943
```

The EM algorithm converged after 3 cycles were completed. After reaching either the maximum number of EM cycles or convergence, the program will perform the Newton-Gauss (Fisher scoring) cycles requested through the NEWTON keyword on the CALIB command. In this example, NEWTON was set to 5. The information matrix for all item parameters is approximated during each Newton step and then used at convergence to provide large-sample standard errors of estimation for the item parameter estimates.

```
[NEWTON CYCLES]   GRADED RESPONSE MODEL

 CATEGORY AND ITEM PARAMETERS AFTER CYCLE    0

  LARGEST CHANGE=  0.000
  -2 LOG LIKELIHOOD =      44224.833

 CATEGORY AND ITEM PARAMETERS AFTER CYCLE    1

  LARGEST CHANGE=  0.004 ( -0.536-> -0.533) at Location of Item: 11 0011
```

The Newton cycles converged after 2 iterations. As all items were assigned to the same BLOCK, only one table is printed to the output file.

At the top of the table, the estimated category parameters are given. For each *m* category item, there are *m-1* category threshold parameters with

$$b_{j1} \le b_{j2} \ldots \le b_{jm-1}.$$

For a polytomous item response model, the discriminating power of a specific categorical response depends on the width of the adjacent category thresholds as well as a slope parameter. Because of this property, the simultaneous estimation of the slope parameter and all m_j category parameters is not obtainable. If the model includes the slope parameter for each item *j* as in this example, the location of the category parameters must be fixed. The CADJUST keyword on the BLOCK command was set to 0, and thus the mean of the category parameters is 0.

For each item, the slope and location parameters, along with corresponding standard errors, are given. All guessing parameters are zero for this model.

```
ITEM BLOCK   1  SBLOCK1

   CATEGORY PARAMETER  :     1.024     0.005    -1.030
   S.E.                :     0.011     0.009     0.011
```

ITEM	BLOCK	SLOPE	S.E.	LOCATION	S.E.	GUESSING	S.E.
0001	1	1.486	0.063	0.006	0.042	0.000	0.000
0002	1	1.526	0.067	-0.012	0.040	0.000	0.000
0003	1	1.472	0.065	0.022	0.041	0.000	0.000

[Similar output omitted]

0019	1	0.699	0.030	0.048	0.060	0.000	0.000
0020	1	0.665	0.029	0.085	0.062	0.000	0.000

The average parameter estimates over all 20 items are given next. If the items are regarded as random samples from a real or hypothetical universe, these quantities estimate the means and standard deviations of the parameters. They could serve as item parameter priors in future item calibrations in this universe.

SUMMARY STATISTICS OF PARAMETER ESTIMATES

PARAMETER	MEAN	STN DEV	N
SLOPE	1.111	0.317	20
LOG(SLOPE)	0.065	0.296	20
THRESHOLD	0.003	0.370	20
GUESSING	0.000	0.000	0

The estimated latent distribution is given next. This distribution is the sum of the posterior distributions of θ for all respondents in the sample. It is represented here as point masses, scaled to sum to 1.0, at 30 equally spaced points on the θ dimension. If the population distribution is normal and the test is sufficiently informative over the range of θ, the posterior distributions for all respondents will approach normality and the latent distribution will approach normality.

	1	2	3	4	5
POINT	-0.4000E+01	-0.3724E+01	-0.3448E+01	-0.3172E+01	-0.2897E+01
WEIGHT	0.6912E-04	0.1967E-03	0.5110E-03	0.1201E-03	0.2420E-02

	6	7	8	9	10
POINT	-0.2621E+01	-0.2345E+01	-0.2069E+01	-0.1793E+01	-0.1517E+01
WEIGHT	0.4662E-02	0.7645E-02	0.1189E-01	0.2005E-01	0.3585E-01

	11	12	13	14	15
POINT	-0.1241E+01	-0.9655E+00	-0.6897E+00	-0.4138E+00	-0.1379E+00
WEIGHT	0.5568E-01	0.7094E-01	0.8078E-01	0.9708E+00	0.1104E+00

	16	17	18	19	20
POINT	0.1379E+00	0.4138E+00	0.6897E+00	0.9655E+00	0.1241E+01
WEIGHT	0.1086E+00	0.9806E+00	0.8301E-01	0.6999E-01	0.5416E-01

	21	22	23	24	25
POINT	0.1517E+01	0.1793E+01	0.2069E+01	0.2345E+01	0.2621E+01
WEIGHT	0.3797E-01	0.2403E-01	0.1328E-01	0.6619E-02	0.2962E-02

```
              26          27          28          29          30
POINT     0.2897E+01  0.3172E+01  0.3448E+01  0.3724E+01  0.4000E+01
WEIGHT    0.1197E-03  0.4451E-03  0.1547E-04  0.5062E-04  0.1563E-05

TOTAL WEIGHT: 1.00000
MEAN        : 0.00000
S.D.        : 0.99970
```

The goodness-of-fit of the polytomous item response model can be tested item by item. Summation of the item fit can also be used for the goodness-of-fit for the test as a whole. The fit statistics are useful in evaluating the fit of models to the same response data when models are nested in their parameters.

Respondents are assigned to H intervals on the θ-continuum. The number of intervals is set using the ITEMFIT keyword on the CALIB command. The expected a posteriori (EAP) score of each respondent is used for assigning respondents to the H intervals. The observed frequency r_{hjk} of the k-th category response to item j in interval h, and N_{hj}, the number of respondents assigned to item j in the h-th interval, are computed. The estimated θ s are rescaled so that the variance of the sample distribution equals that of the latent distribution on which the MML estimation of the parameters is based.

Thus an H by m_j contingency table is obtained for each item j. In order to avoid expected values less than 5, neighboring intervals and/or categories may be merged. For each interval, the interval mean, θ_h, and the value of the fitted response function $P_{jk}(\theta_h)$, are computed.

Finally, a likelihood ratio χ^2-statistic for each item is computed by

$$G_j^2 = 2\sum_{h=1}^{H_j}\sum_{k=1}^{m_j} r_{hjk} \ln \frac{r_{hjk}}{N_{hj}P_{jk}(\theta_h)},$$

where H_j is the number of intervals left after neighboring intervals are merged. The degrees of freedom is $\sum_{j=1}^{H_j}(m_j^* - 1)$ where m_j^* is the number of categories left after merging.

The likelihood ratio χ^2-statistic for the test as a whole is simply the summation of the separate χ^2-statistics. The number of degrees of freedom is also the summation of the degrees of freedom for each item.

```
                ITEM FIT STATISTICS

-----------------------------------------------
|  BLOCK   | ITEM | CHI-SQUARE |  D.F. | PROB. |
-----------------------------------------------
| SBLOCK1  | 0001 |   25.00714 |   20. | 0.201 |
|          | 0002 |   23.18082 |   20. | 0.280 |
|          | 0003 |   25.66873 |   20. | 0.177 |
|          | 0004 |   31.56813 |   19. | 0.035 |
```

```
|          | 0005 |   19.88483 |   19. | 0.339 |
|          | 0006 |   13.51922 |   22. | 0.918 |
   …
|          | 0019 |   12.51549 |   25. | 0.982 |
|          | 0020 |   25.25502 |   25. | 0.448 |
-----------------------------------------------
|  TOTAL   |      |  492.43930 |  442. | 0.049 |
-----------------------------------------------
```

The null hypothesis tested here is that there are no significant differences between the expected and observed frequencies. A significant χ^2-statistic indicates that item parameters differ across the raw score groups and that the assumed model is not appropriate for the data. In this case, no item showed poor fit to the assumed model.

Phase 3 output

The first information given in the output from the scoring phase is on the scoring function used for scaling. The default function is STANDARD, and thus the standard scoring function (1.0, 2.0) will be used even though a different scoring function may be used for calibration. The scoring function may also be set to CALIBRATION (SCORING keyword on the SCORE command) to use the calibration scoring function specified on the BLOCK command instead. Note that the scoring function only applies to the partial credit model.

```
SCORING FUNCTION FOR SCALING

BLOCK:   1  SBLOCK1

         1     1.000
         2     2.000
         3     3.000
         4     4.000
```

Bayes estimates are computed for each examinee with respect to his or her group latent distribution (controlled by the EAP option on the SCORE command used here). A discrete distribution on a finite number of points (see below) is used as prior. The user may select the number of points and the type of prior using the NQPT and DIST keywords on the SCORE command.

```
[EAP SUBJECT ESTIMATION]

QUADRATURE POINTS AND PRIOR WEIGHTS:

               1           2           3           4           5
POINT    -0.4000E+01 -0.3724E+01 -0.3448E+01 -0.3172E+01 -0.2897E+01
WEIGHT    0.3692E-04  0.1071E-03  0.2881E-03  0.7181E-03  0.1659E-02

               6           7           8           9          10
POINT    -0.2621E+01 -0.2345E+01 -0.2069E+01 -0.1793E+01 -0.1517E+01
WEIGHT    0.3550E-02  0.7042E-02  0.1294E-01  0.2205E-01  0.3481E-01

              11          12          13          14          15
POINT    -0.1241E+01 -0.9655E+00 -0.6897E+00 -0.4138E+00 -0.1379E+00
WEIGHT    0.5093E-01  0.6905E-01  0.8676E-01  0.1010E+00  0.1090E+00
```

```
              16          17          18          19          20
POINT     0.1379E+00  0.4138E+00  0.6897E+00  0.9655E+00  0.1241E+01
WEIGHT    0.1090E+00  0.1010E+00  0.8676E-01  0.6905E-01  0.5093E-01

              21          22          23          24          25
POINT     0.1517E+01  0.1793E+01  0.2069E+01  0.2345E+01  0.2621E+01
WEIGHT    0.3481E-01  0.2205E-01  0.1294E-01  0.7042E-02  0.3550E-02

              26          27          28          29          30
POINT     0.2897E+01  0.3172E+01  0.3448E+01  0.3724E+01  0.4000E+01
WEIGHT    0.1659E-02  0.7181E-03  0.2881E-03  0.1071E-03  0.3692E-04



MEANS AND STANDARD DEVIATIONS OF ABILITY DISTRIBUTIONS

 SCORE        MEAN    STANDARD     TOTAL
 NAME                DEVIATION FREQUENCIES
 ---------------------------------------------
 EAP           0.000     0.985      1000.00
 ---------------------------------------------
```

In this example, the keywords SMEAN and SSD were set to 0 and 1 respectively on the SCORE command. As a result, the following output reflects the rescaling constants (0.000 and 1.015) used in this particular case.

```
RESCALING DONE WITH RESPECT TO USER SUPPLIED LINEAR TRANSFORMATION

 SCORE       LOCATION  SCALING      TOTAL
 NAME        CONSTANT  CONSTANT  FREQUENCIES
 ---------------------------------------------
 EAP           0.000     1.015      1000.00
 ---------------------------------------------
```

Scores are saved to an external file (keyword SCORE on SAVE command), but the first three scores are printed to the output file for purposes of checking. When EAP is used for scoring, the S.E. column represents the posterior standard deviation.

```
SUBJECT IDENTIFICATION                 WEIGHT/FREQUENCY
 SCORE NAME    GROUP    WEIGHT MEAN CATEGORY ATTEMPTS    ABILITY       S.E.
 ------------------------------------------------------------------------------
      .447          |        1  GROUP 01      1.00
   1  EAP         1  |     1.00       3.00     1.00        0.6435      0.2193
------------------------------------------------------------------------------
     -.934          |        2  GROUP 01      1.00
   1  EAP         1  |     1.00       1.95     1.00       -0.7442      0.2164
 ------------------------------------------------------------------------------
     -.564          |        3  GROUP 01      1.00
   1  EAP         1  |     1.00       2.10     1.00       -0.4392      0.2115
------------------------------------------------------------------------------

 MEANS AND STANDARD DEVIATIONS OF ABILITY DISTRIBUTIONS

 SCORE        MEAN    STANDARD     TOTAL
 NAME                DEVIATION FREQUENCIES
 ---------------------------------------------
```

```
EAP            0.000     1.000      1000.00
-------------------------------------------
```

When EAP is selected, an estimate of the population distribution of ability in the form of a discrete distribution of a finite number of points is obtained by accumulating the posterior densities over the subjects at each quadrature point. These sums are then normalized to obtain the estimated probabilities at the points. Improved estimates of the latent distribution may be obtained after one more iteration of the solution.

The program also computes the mean and standard deviation for the estimated latent distribution. Sheppard's correction for coarse grouping is used in the calculation of the standard deviation. The EAP estimate is the mean of the posterior distribution while the standard error is the standard deviation of the posterior distribution. Posterior weights are only given when EAP is used. Note that it is based on all cases, and not just on those cases used in calibration.

```
QUADRATURE POINTS AND POSTERIOR WEIGHTS: SCORE SET #    1
              1           2           3           4           5
 POINT    -0.4000E+01 -0.3724E+01 -0.3448E+01 -0.3172E+01 -0.2897E+01
 WEIGHT    0.6822E-04  0.1942E-03  0.5048E-03  0.1187E-03  0.2494E-02

              6           7           8           9          10
 POINT    -0.2621E+01 -0.2345E+01 -0.2069E+01 -0.1793E+01 -0.1517E+01
 WEIGHT    0.46622-02  0.7591E-02  0.1180E-01  0.1987E-01  0.3555E-01

             11          12          13          14          15
 POINT    -0.1241E+01 -0.9655E+00 -0.6897E+00 -0.4138E+00 -0.1379E+00
 WEIGHT    0.5541E-01  0.7082E-01  0.8069E-01  0.9694E+00  0.1105E+00

             16          17          18          19          20
 POINT     0.1379E+00  0.4138E+00  0.6897E+00  0.9655E+00  0.1241E+01
 WEIGHT    0.1088E+00  0.9832E+00  0.8323E-01  0.7015E-01  0.5431E-01

             21          22          23          24          25
 POINT     0.1517E+01  0.1793E+01  0.2069E+01  0.2345E+01  0.2621E+01
 WEIGHT    0.3809E-01  0.2411E-01  0.1333E-01  0.6645E-02  0.2974E-02

             26          27          28          29          30
 POINT     0.2897E+01  0.3172E+01  0.3448E+01  0.3724E+01  0.4000E+01
 WEIGHT    0.1202E-03  0.4470E-03  0.1554E-04  0.5083E-04  0.1569E-05

 TOTAL WEIGHT: 1.00000
 MEAN        : 0.00012
 S.D.        : 1.01246
```

The mean and standard deviation of the latent posterior distribution calculated from posterior weights at quadrature points are also given. In these calculations, the formulas for the variance of grouped data are used, with quadrature points as class marks and posterior weights as class frequencies.