Equivalent groups equating

This example illustrates the equating of equivalent groups with the BILOG-MG program. Two parallel test forms of 20 multiple-choice items were administered to two equivalent samples of 200 examinees drawn from the same population. There are no common items between the forms. Because the samples were drawn from the same population, GROUP commands are not required. The FORM1 command lists the order of the items in Form 1 and the FORM2 command lists the order of the items in Form 2. These commands follow directly after the TEST command as indicated by the NFORM=2 keyword on the INPUT command. As only one test is used, the vector of items per subtest given by the NITEMS keyword on the LENGTH command contains only one entry.

The SAVE option on the GLOBAL command is used in combination with the SAVE command to save item parameter estimates and scores to the external files **exampl04.par** and **exampl04.sco** respectively.

In this example, 40 unique item responses are given in the data file. The first few lines of the data file are shown below. The first record shown after the answer keys for the two forms, which should always appear first and in the same format as the data, contains responses to items 1 through 20 in the second line associated with this examinee. In the case of the data shown for another examinee who responded to the second form, responses in the same positions in the data file correspond to items 21 through 40. Keep in mind that the number of items read by the format statement is the total number of items in the form, when NFORM=1 and the total number of items in the longest form when NFORM>1.

```
1     11111111111111111111
2     11111111111111111111
1 001 11111111122212122111
1 002 11222212221222222112
1 003 12121221222222221222
1 004 11212212222222212222
…
2 198 11112211111222212211
2 199 21122222222222222122
2 200 11111111111111221111
```

The FLOAT option is used on the CALIB command to request the estimation of the means of the prior distributions of item parameters along with the parameters. This option should not be used when the data set is small and items few. Means of the item parameters may drift indefinitely during estimation cycles under these conditions. In the CALIB command, the FIXED option is also required to keep the prior distributions of ability fixed during the EM cycles of this example. In multiple-group analysis, the default is “not fixed”.

ML estimates of ability are rescaled to a mean of 250 and standard deviation of 50 in Phase 3 (METHOD=1, RSCTYPE=3, LOCATION=250, SCALE=50). By setting INFO to 1 on the SCORE command, the printing of test information curves to the phase 3 output file is requested. To request the calculation of expected information for the population, the POP option may be added to this command. In the case of multiple subtests, the further addition

of the YCOMMON option will request the expression of test information curves for the subtests in comparable units.

```
EXAMPL04.BLM - EQUIVALENT GROUPS EQUATING
               SIMULATED RESPONSES TO TWO 20-ITEM PARALLEL TEST FORMS
>GLOBAL  DFNAME='EXAMPL04.DAT', NPARM=2, SAVE;
>SAVE    SCORE='EXAMPL04.SCO', PARM='EXAMPL04.PAR';
>LENGTH  NITEMS=40;
>INPUT   NTOT=40, NFORM=2, KFNAME='EXAMPL04.DAT', NALT=5, ,NIDCHAR=5;
>ITEMS   INUM=(1(1)40), INAME=(T01(1)T40);
>TEST    TNAME=SIM;
>FORM1   LENGTH=20, INUM=(1(1)20);
>FORM2   LENGTH=20, INUM=(21(1)40);
(5A1,T1,I1,T7,20A1)
>CALIB   FIXED, FLOAT, NQPT=31, TPRIOR, PLOT=.05;
>SCORE   METHOD=1, RSCTYPE=3, LOCATION=250, SCALE=50, NOPRINT, INFO=1;
```

Phase 1 output

Because all examples are drawn from the same population, all responses are combined in the results. Since there are no common items between forms, the number tried for each item is 200. If there had been common items, their number tried would be 400. Results for the first 5 items are shown below.

```
       400 OBSERVATIONS READ FROM FILE:   EXAMPL04.DAT
       400 OBSERVATIONS WRITTEN TO FILE:  MF.DAT

 ITEM STATISTICS FOR SUBTEST SIM
                                                        ITEM*TEST CORRELATION
ITEM   NAME     #TRIED    #RIGHT   PCT    LOGIT/1.7  PEARSON  BISERIAL
-------------------------------------------------------------------------
    1   T01        200.0     165.0    82.5    -0.91     0.446    0.658
    2   T02        200.0     171.0    85.5    -1.04     0.416    0.642
    3   T03        200.0     150.0    75.0    -0.65     0.524    0.715
    4   T04        200.0     138.0    69.0    -0.47     0.448    0.588
    5   T05        200.0     149.0    74.5    -0.63     0.391    0.531
```

Phase 2 output

Item parameter estimation assumes a common latent distribution for the random equivalent groups administered the respective test forms. Empirical prior distributions are assumed for the slope and threshold parameters. The means of these priors are estimated concurrently with the item parameters.

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

 MAXIMUM NUMBER OF EM CYCLES:                 20
 MAXIMUM NUMBER OF NEWTON CYCLES:              2
 CONVERGENCE CRITERION:                      0.0100
 ACCELERATION CONSTANT:                      1.0000
 LATENT DISTRIBUTION:                      NORMAL PRIOR FOR EACH GROUP
 PLOT EMPIRICAL VS. FITTED ICC's:          YES, FOR ITEMS WITH FIT
                                           PROBABILITY
                                           LESS THAN 0.05000
 DATA HANDLING:                            DATA ON SCRATCH FILE
 CONSTRAINT DISTRIBUTION ON SLOPES:        YES
 CONSTRAINT DISTRIBUTION ON THRESHOLDS:    YES
 SOURCE OF ITEM CONSTRAINT DISTRIBUTION
 MEANS AND STANDARD DEVIATIONS:            PROGRAM DEFAULTS
                                           ITEM CONSTRAINTS IF PRESENT
                                           WILL BE UPDATED EACH CYCLE
```

Final iterations of the solutions and some of the results are as follows. Indeterminacy of the origin and unit of the ability scale is resolved in Phase 2 by setting the mean and standard deviation of the latent distribution to zero and one, respectively.

```
 -2 LOG LIKELIHOOD =       8297.415

 UPDATED PRIOR ON LOG SLOPES; MEAN & SD =       -0.23882     0.50000
 UPDATED PRIOR ON THRESHOLDS; MEAN & SD =       -0.01801     2.00000

 CYCLE      5;   LARGEST CHANGE=   0.00752



[NEWTON CYCLES]

 UPDATED PRIOR ON LOG SLOPES; MEAN & SD =       -0.23457     0.50000
 UPDATED PRIOR ON THRESHOLDS; MEAN & SD =       -0.01751     2.00000

 -2 LOG LIKELIHOOD:         8297.4560

 CYCLE     6;   LARGEST CHANGE=   0.00489
```

After assigning cases to the intervals (shown below) on the basis of the EAP estimates of their scale scores, the program computes the expected number of correct responses in the interval by multiplying these counts by the response model probability at the indicated θ. The χ^2 is computed in the usual way from the differences between the observed and expected counts.

The counts are displayed so that the user can judge whether there are enough cases in each group to justify computing a χ^2 statistic. If not, the user should reset the number of intervals.

```
 INTERVAL COUNTS FOR COMPUTATION OF ITEM CHI-SQUARES
--------------------------------------------------------------------------
         15.    30.    36.    52.    70.    69.    48.    36.    44.
--------------------------------------------------------------------------

 INTERVAL AVERAGE THETAS
--------------------------------------------------------------------------
      -2.000 -1.520 -1.076 -0.648 -0.191  0.235  0.620  1.100  1.724
--------------------------------------------------------------------------

 SUBTEST SIM     ;  ITEM PARAMETERS AFTER CYCLE   6
ITEM      INTERCEPT    SLOPE    THRESHOLD   LOADING   ASYMPTOTE    CHISQ     DF
             S.E.       S.E.       S.E.       S.E.       S.E.      (PROB)
-------------------------------------------------------------------------------
T01      |   1.339  |   1.000  |  -1.338  |   0.707  |   0.000  |     2.3   5.0
         |   0.192* |   0.206* |   0.194* |   0.146* |   0.000* | (0.8044)
T02      |   1.488  |   0.961  |  -1.549  |   0.693  |   0.000  |     4.4   6.0
         |   0.211* |   0.199* |   0.218* |   0.144* |   0.000* | (0.6179)
(Similar output omitted)       |          |          |          |
T39      |   0.508  |   0.911  |  -0.557  |   0.673  |   0.000  |     1.8   6.0
         |   0.119* |   0.172* |   0.126* |   0.127* |   0.000* | (0.9334)
T40      |   0.525  |   0.675  |  -0.777  |   0.559  |   0.000  |     5.4   7.0
         |   0.107* |   0.130* |   0.175* |   0.108* |   0.000* | (0.6055)
-------------------------------------------------------------------------------
                                                          * STANDARD ERROR

 LARGEST CHANGE =    0.004890                          176.3 243.0
                                                       (0.9996)
```

```
PARAMETER       MEAN   STN DEV
 ----------------------------------
 SLOPE           0.809    0.153
 LOG(SLOPE)     -0.230    0.189
 THRESHOLD      -0.019    0.975
```

Phase 3 output

For purposes of reporting test scores, the ability scale is set so that the mean score distribution in the sample of examinees is 250 and the standard deviation is 50. The item parameters are rescaled accordingly.

```
>SCORE METHOD = 1, RSCTYPE = 3,LOCATION = 250, SCALE = 50,NOPRINT, INFO = 1;

PARAMETERS FOR SCORING, RESCALING, AND TEST AND ITEM INFORMATION
METHOD OF SCORING SUBJECTS:                MAXIMUM LIKELIHOOD
SCORES WRITTEN TO FILE                      EXAMPL04.SCO
TYPE OF RESCALING:                         IN THE SAMPLE DISTRIBUTION
REFERENCE GROUP FOR RESCALING:             GROUP:  0
```

Before rescaling, the sample mean score is essentially the same as that in the Phase 2 latent distribution. The standard deviation is larger, however, because the score distribution includes measurement error variance.

Summary statistics for each group include the following.

- The correlation matrix of the test scores (when there is more than one test).
- The mean, standard deviation and variance of the θ score estimates:
- Maximum Likelihood (ML) estimate
- Bayes Model (Maximum A Posteriori, MAP) estimate
- Bayes (Expected, EAP) estimate

The summary of the error variation depends on the type of estimate:

- Maximum Likelihood – Harmonic Root-Mean-Square standard errors: The error variance for each case is the reciprocal of the Fisher information at the likelihood maximum for the case. The standard error is the reciprocal square root of the average of these variances.
- MAP – Root-Mean-Square posterior standard deviation: The error variance for each case is the posterior information at the maximum of the posterior probability density of θ, given the response pattern of the case. The standard error is the square root of the average of these variances.
- EAP – Root-Mean-Square posterior standard deviation: The error variance for each case is the variance of the posterior distribution of theta, given the response pattern of the case. The standard error is the square root of the average of these variances.

The empirical reliability of the test is the θ score variance divided by the sum of that variance and the error variance.

Note:

The expected value of the sum of the θ score variance and the error variance is the variance of the latent distribution of the group. The sum of the corresponding sample variances should tend to that value as the sample size increases.

```
SUMMARY STATISTICS FOR SCORE ESTIMATES
======================================


CORRELATIONS AMONG TEST SCOREs

                SIM
SIM            1.0000


MEANS, STANDARD DEVIATIONS, AND VARIANCES OF SCORE ESTIMATES

TEST:               SIM
MEAN:             0.0057
S.D.:             1.1426
VARIANCE:         1.3054

HARMONIC ROOT-MEAN-SQUARE STANDARD ERRORS OF THE ML ESTIMATES

TEST:               SIM
RMS:              0.4203
VARIANCE:         0.1767

EMPIRICAL RELIABILITY:     0.8647

RESCALING WITH RESPECT TO SAMPLE DISTRIBUTION
-------------------------------------------------

         RESCALING  CONSTANTS
TEST       SCALE    LOCATION
SIM         43.762   249.749
```

The scaled scores are saved on an external file and their printing is suppressed in all but the first two cases.

```
GROUP   SUBJECT IDENTIFICATION
WEIGHT   TEST      TRIED  RIGHT  PERCENT     ABILITY      S.E.
-----------------------------------------------------------------
  1      1                              |                       |
  1.00   SIM          20     14    70.00 |  282.5091   17.5097 |
  1      1                              |                       |
  1.00   SIM          20      6    30.00 |  217.0505   16.8979 |
-----------------------------------------------------------------
```

The magnitudes of the rescaled item parameters reflect the new origin and unit of the scale. The thresholds center around 250 and the slopes are smaller by a factor of about 50. The slopes are printed here to only three decimal places but appear accurately in the saved items parameter file. If saved parameters are used to score other examinees, the results will be determined in the present sample.

```
TEST    SIM     ;  RESCALED ITEM PARAMETERS
ITEM        INTERCEPT   SLOPE    THRESHOLD   LOADING    ASYMPTOTE
               S.E.      S.E.       S.E.       S.E.        S.E.
-------------------------------------------------------------------
T01       |  -4.371  |   0.023  | 191.189  |   0.707  |   0.000
          |   1.190* |   0.005* |   8.501* |   0.146* |   0.000*
          |          |          |          |          |
T02       |  -3.994  |   0.022  | 181.956  |   0.693  |   0.000
          |   1.157* |   0.005* |   9.357* |   0.144* |   0.000*
          |          |          |          |          |
(Similar output omitted)
          |          |          |          |          |
          |          |          |          |          |
T40       |  -3.327  |   0.015  | 215.726  |   0.559  |   0.000
          |   0.748* |   0.003* |   7.651* |   0.108* |   0.000*
-------------------------------------------------------------------
```

```
PARAMETER       MEAN  STN DEV
-----------------------------
SLOPE          0.018    0.003
LOG(SLOPE)    -4.009    0.189
THRESHOLD    248.921   42.657

MEAN & SD OF SCORE ESTIMATES AFTER RESCALING:   250.000    50.000
```

Results of the information analysis are depicted in the following line printer plot. Points indicated by + and * represent the information and measurement error functions, respectively. This plot applies to all 40 items and not to the separate test forms. Because the item thresholds are normally distributed with mean standard similar to that of the score distribution, the precision of the item set is greatest toward the middle of the scale.

TEST INFORMATION CURVE FOR TEST: SIM FORM: 1