Example: Application of the NR Model to a Science Test, MMLE, MULTILOG

The data for our example come from the Third International Mathematics and Science Study (TIMSS; Gonzalez et al., 1998) database. Using the responses from the 1995 Canadian examinees, four physical science items were selected. Three of the items are in a multiple-choice format (four options), whereas the fourth item is open-ended and scored using a four-point rubric. The three multiple-choice items' options constitute nominal response categories, and we initially treat the fourth item as such. The responses to each item are labeled 1 through 4. It should be noted that the toolbox of model–data fit methods summarized in Chapter 6 is still relevant and would be used in practice. We assume a unidimensional latent space and conditional independence.

Rather than analyzing individual case data in this example, we use pattern data. With four items, each with four possible "options," there are $4^4 = 256$ possible patterns (i.e., number of patterns = m^L). Of these 256 possible patterns we observed 233. The observed patterns indicate that for each item there is a response to each option. We use MULTILOG for calibrating these data. As mentioned in Chapter 8, when pattern data are calibrated MULTILOG automatically produces the EAP $\hat{\theta}$.

The command file for the NR model calibration of the physical science test is shown in Table 1. Most of this file is described in Chapters 7 and 8. On the `PROBLEM` line we use `PATTERNS` to indicate the calibration of pattern data and `NPATTERNS` to specify the number of patterns. On the `TEST` line we specify that each of the four items consists of four response categories (i.e., `NC=(4(0)4)`) as well as which option has the highest frequency for each item (i.e., `HIGH=(2,3,4,3)`). For example, the `HIGH` subcommand indicates that item 1's highest frequency is category 2; that for item 2, category 3 has the highest frequency; and so on. (These

de Ayala

highest frequency categories specify the item's baseline response category.) In our example the HIGH category is also the correct response. We set the number of EM cycles to 500 and the number of M-STEPS to 25 by using subcommands NC=500 and IT=25, respectively, on the EST line. Similar to what has been done in previous examples, the unique response pattern and its frequency are used as the person identification field. Therefore, the FORTRAN format's first field (9A1) refers to the person identification field, the actual item responses occupy the first four columns and are read by 4A1, and the field F5.0 is used to read the pattern frequencies.

Table 1. Command file for MULTILOG NR model calibration example.[a]

```
MULTILOG for Windows 7.00.2327.2
NR CALIBRATION, 4 PHYSICAL SCIENCE ITEMS
>PROBLEM RANDOM,
          PATTERNS,                    ⇐ Specification that pattern data is being used
          DATA = 'C:SCIENCE.PAT',
          NITEMS = 4,
          NGROUPS = 1,
          NPATTERNS = 256,             ⇐ Number of possible patterns[b]
          NCHARS = 9;
>TEST ALL,
      NOMINAL,                         ⇐ Specification of NR model
      NC = (4(0)4),
      HIGH = (2,3,4,3);                ⇐ Ordinal position of highest frequency category
>EST NC=500 IT=25;                     ⇐ Changing the number of EM and M-Step iterations
>END ;
4                                      ⇐ Number of response categories
1234                                   ⇐ The response code line
1111                                   ⇐ The 1s are coded to be 1s for all four items
2222                                   ⇐ The 2s are coded to be 2s for all four items
3333                                   ⇐ The 3s are coded to be 3s for all four items
4444                                   ⇐ The 4s are coded to be 3s for all four items
(9A1,T1,4A1,F5.0)
```

[a]The text following the '⇐' is provided to help the reader understand the corresponding input.
[b]The nonobserved patterns are added to the data file with corresponding frequencies of zero.

Table 2 contains the abridged output. With four 4-option items there are L*2(m – 1) = 4*2(3) = 24 parameters being estimated and we see that this is the value on the NUMBER OF FREE PARAMETERS line. The identification field shows that the first observation is the pattern 1111 and that this pattern occurred once in the data. Convergence is achieved in 29 cycles.

de Ayala

Table 2. Abridged output from NR model calibration example.

```
    :
<echo of command file>
  NUMBER OF FREE PARAMETERS IS:   24
    :
 FIRST OBSERVATION AS READ-

 ID    1111    1
 ITEMS 1111
 NORML      0.000
 WT/CR       1.00

 FINISHED CYCLE 29
 MAXIMUM INTERCYCLE PARAMETER CHANGE=   0.00082 P(  21)
    :
 ITEM SUMMARY
    :
 ITEM   1:       4 NOMINAL CATEGORIES,  2 HIGH
  CATEGORY(K): 1       2       3       4
    A(K)       0.18    0.61   -0.37   -0.42
    C(K)       0.18    0.55   -0.56   -0.16

           CONTRAST-COEFFICIENTS (STANDARD ERRORS)
  FOR:              A                       C
  CONTRAST P(#)   COEFF.[ DEV.]   P(#)   COEFF.[ DEV.]
      1       1    0.44 (0.09)      4    0.37 (0.08)
      2       2   -0.55 (0.12)      5   -0.74 (0.11)
      3       3   -0.60 (0.11)      6   -0.34 (0.10)

 @THETA:      INFORMATION:   (Theta values increase in steps of 0.2)
 -3.0 - -1.6  0.077  0.086  0.096  0.107  0.118  0.129  0.140  0.151
 -1.4 -  0.0  0.161  0.170  0.178  0.183  0.187  0.188  0.187  0.184
  0.2 -  1.6  0.179  0.172  0.163  0.153  0.143  0.132  0.121  0.110
  1.8 -  3.0  0.100  0.090  0.081  0.073  0.065  0.058  0.052

  OBSERVED AND EXPECTED COUNTS/PROPORTIONS IN
  CATEGORY(K):  1       2       3       4
  OBS. FREQ.    456     723     245     375
  OBS. PROP.  0.2535 0.4019 0.1362 0.2084
  EXP. PROP.  0.2535 0.4020 0.1362 0.2084
    :
TOTAL TEST INFORMATION
@THETA:      INFORMATION:
 -3.0 - -1.6  1.380  1.424  1.471  1.521  1.572  1.624  1.674  1.721
 -1.4 -  0.0  1.763  1.797  1.822  1.837  1.841  1.834  1.817  1.792
  0.2 -  1.6  1.759  1.722  1.681  1.638  1.595  1.551  1.509  1.468
  1.8 -  3.0  1.429  1.392  1.358  1.325  1.295  1.268  1.242
    :
MARGINAL RELIABILITY:    0.4142
    :
 OBSERVED(EXPECTED)    STD.  :     EAP (S.D.)   :  PATTERN
                       RES.  :                  :

      1.0(     0.7)    0.40  :   -0.97 ( 0.75) :  1111
      2.0(     0.8)    1.33  :   -0.45 ( 0.75) :  1112
      0.0(     1.0)   -1.01  :   -0.61 ( 0.75) :  1113
      2.0(     0.5)    2.00  :   -0.18 ( 0.75) :  1114
    :
 NEGATIVE TWICE THE LOGLIKELIHOOD=       288.4
 (CHI-SQUARE FOR SEVERAL TIMES MORE EXAMINEES THAN CELLS)
```

In the `ITEM SUMMARY` section we find our item parameter estimates. For item 1 there are

de Ayala

four possible options (i.e., `4 NOMINAL CATEGORIES`) and the response of 2 has the highest frequency (i.e., `2 HIGH`); this option is also the correct option. The constrained category discrimination parameter estimates are read from the `A(K)` line. That is, $\hat{\alpha}_{11} = 0.18$, $\hat{\alpha}_{12} = 0.61$, $\hat{\alpha}_{13} = -0.37$, and $\hat{\alpha}_{14} = -0.42$ or $\underline{\hat{\boldsymbol{\alpha}}} = (0.18, 0.61, -0.37, -0.42)$. The constrained intercept estimates are listed on the `C(K)` line with $\hat{\gamma}_{11} = 0.18$, $\hat{\gamma}_{12} = 0.55$, $\hat{\gamma}_{13} = -0.56$, and $\hat{\gamma}_{14} = -0.16$ or $\underline{\hat{\gamma}}_{-} = (0.18, 0.55, -0.56, -0.16)$. The unconstrained discrimination and intercept parameters are found in the section labeled `CONTRAST-COEFFICIENTS (STANDARD ERRORS)`. The columns labeled `A` and `C` contain the $m_j - 1$ discrimination and intercept parameter estimates, respectively. For example, for item 1 option 1 we have $\hat{\alpha}_{11}^{u} = 0.44$ and $\hat{\gamma}_{11}^{u} = 0.37$.

Following the unconstrained parameter estimate section are the item's information values. These would be interpreted as done in Chapter 7. For this item as well as for the other items, we can see that the expected and observed proportions show good agreement for all categories (i.e., the difference between `OBS. FREQ.` and `EXP. PROP.` is about 1/10,000th).

In some situations one or more of an item's response categories may not be attractive and may never be chosen. These are sometimes referred to as *null* categories. In these cases one does not have data to estimate the category's parameters. In short, the item is functioning with fewer categories than are specified for the calibration. This situation would reveal itself by the null category's corresponding observed proportion and frequency being zero, as well as by the absence of an ORF for the null category in the item's ORF plot. If a null category occurs, then one should ignore the null category's parameter estimates and recalibrate the item set specifying the appropriate number of observed categories for each item. For instance, assume that item 4's fourth option is not selected by any individuals. Therefore, this item is functioning as a three-

de Ayala

category item. To specify that item 4 has only three observed categories, we change the `NC` specification on the `TEST` line to indicate three categories for item 4 (i.e., `NC = (4,4,4,3)`); in this case the corresponding `HIGH` category identification does not need to be changed (i.e., `HIGH = (2,3,4,3)`).

The parameter estimates for all the items are presented in Table 3. As would be expected, the sum of the constrained $\hat{\alpha}_{jk}$ s for an item is 0.0 and this is also the case for an item's $\hat{\gamma}_{jk}$ s. Looking at item 1, one sees that category 2 does the best of the item's response categories in discriminating among the individuals ($\hat{\alpha}_{12}$ = 0.61) and it is also the most attractive ($\hat{\gamma}_{12}$ = 0.55). Because the γ s are associated with category frequencies, the category with the largest frequency has the largest positive γ and the category with smallest frequency is associated with the most negative γ. (If a category has its α equal to zero, then its corresponding γ is poorly defined and has a tendency to drift. In this situation the calibration may not converge, but according to Thissen there is little loss of fit by allowing the calibration to stop because of reaching the maximum number of iterations.) The –2ln*L* for this calibration is 288.4 (BIC = 468.2797); we use the -2ln*L* value below.

The corresponding ORFs for item 1 (Figure 1, left panel) shows that category 2 is most useful for persons located above approximately –0.69 and that category 4 is attractive for persons located below approximately –0.69. That is, the intersection between categories 2 and 4 is

$$\delta_{2,4} = \frac{0.55 - (-0.16)}{-0.42 - 0.61} = -0.69$$

Table 3. Constrained Item Parameter Estimates for Science Items - NR Model.

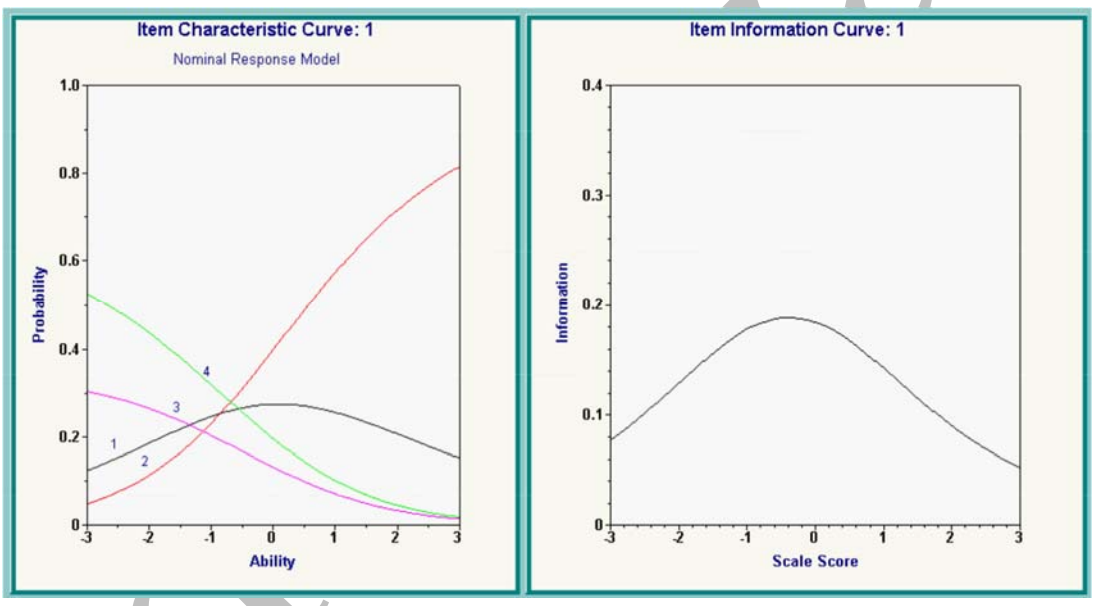
	Discrimination				Intercept			
Item	$\hat{\alpha}_{j1}$	$\hat{\alpha}_{j2}$	$\hat{\alpha}_{j3}$	$\hat{\alpha}_{j4}$	$\hat{\gamma}_{j1}$	$\hat{\gamma}_{j2}$	$\hat{\gamma}_{j3}$	$\hat{\gamma}_{j4}$

de Ayala

1	0.18	0.61	-0.37	-0.42	0.18	0.55	-0.56	-0.16
2	-0.07	-0.29	0.54	-0.18	-1.19	-0.03	0.76	0.46
3	-0.45	-0.66	0.29	0.82	-0.76	-0.47	0.53	0.70
4	-0.73	-0.18	-0.10	0.66	-0.58	0.25	0.33	0.00

MULTILOG's graph uses a red line to identify the `HIGH` category (i.e., category 2). According to the ORF pattern this item functions primarily as a two-category item with categories 2 and 4 being the primary categories. The item's information function (Figure 1, right panel) shows that this item is most useful for estimating persons located around –0.5.

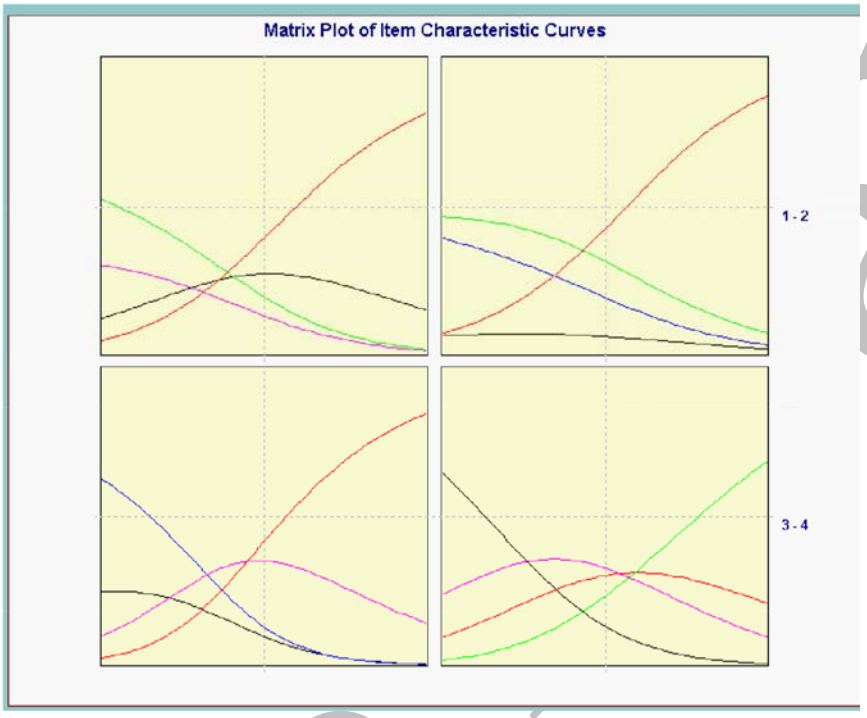
Figure 1. ORFs and item information function for item 1.



The matrix plot feature simultaneously presents the ORFs for all the items and is shown in Figure 2. This figure shows that item 2 (top right) is behaving as a binary item. The relatively "flat" ORF reflects a category (specifically, category 1) that is not very discriminating (this is also reflected in its $\hat{\alpha}_{21}$ = –0.07) and is not, comparatively speaking, attracting very many respondents; the latter interpretation comes from the magnitude of its $\hat{\gamma}_{21}$ = –1.19. Item 3 is

functioning as a three-, not four-, category item. As mentioned in previous chapters, a fit analysis should involve a comparison of empirical and predicted ORFs. The program MODFIT could be used for this purpose with the NR model.

Figure 2. ORFs for all four items, NR model calibration.



Example: Mixed Model Calibration of the Science Test—NR and PC Models, MMLE, MULTILOG

The calibration examples in the previous chapters have applied a single model to an instrument because the instrument used a single-item response format. However, some instruments may contain various response formats. In these cases, and provided that the IRT assumptions are met for the entire instrument, one may apply multiple models to the data and still perform a single calibration of the instrument. In this example we apply two models, the NR

and PC models, to the science test. As previously mentioned, the PC model is subsumed by the NR model. Therefore, this mixed model calibration involves a nested model and we are able to compare the fit of this mixed model calibration with that of the NR model calibration.[1]

In the previous example we mentioned that item 4 is an open-ended question. This question is concerned with whether the combined weight of two objects would change if one object is placed inside the other as opposed to remaining separate. The rubric identified the correct answer and various incorrect answers. One may consider some of these incorrect answers to reflect a better understanding of the physical sciences than do the other incorrect answers. For example, answering that the combined weight is zero when one of the objects is placed inside the other reflects a greater lack of understanding than some nonzero value. Therefore, responses to this item are scored by giving no credit, varying degrees of partial credit, and full credit. In this example we treat this item as a four-category partial credit item. Specifically, a value of 4 is assigned to a correct response, a value of 3 for a partially correct response, a value of 2 for a response that does not indicate an understanding as good as a value of 3, and a 1 for an incorrect response. We could have coded these as 3, 2, 1, and 0, respectively. However, we would then have had to recode these values to be 1 through 4 to perform a PC model calibration because a code of 0 is used internally by MULTILOG. To summarize our data, items 1 through 3 are considered to be nominally scored and item 4 is scored in a graded fashion. To accommodate these different scoring paradigms we apply two models to the four-item instrument. The NR model is used for items 1 through 3 and the PC model is applied to item 4. The command file for this calibration is shown in Table 4.

Table 4. Command file for the NR and PC mixed model calibration.[a]

```
MULTILOG for Windows 7.00.2327.2
MIX NR & PC MODELS; ITEM 4 - PC MODEL
```

de Ayala

```
>PROBLEM RANDOM,
         PATTERNS,
         DATA = 'C:SCIENCE.PAT',
         NITEMS = 4,
         NGROUPS = 1,
         NPATTERNS = 256,
         NCHARS = 9;
>TEST ALL,
      NOMINAL,
      NC = (4(0)4),                          ⇐ That is, NC = (4, 4, 4, 4)
      HIGH = (2,3,4,4);
>TMATRIX ITEMS=4,AK,POLYNOMIAL;              ⇐ This line and the next 2 impose the constraints
>FIX     ITEMS=4,AK=(2,3),VALUE=0.0;           on the NR model to obtain the PC model
>TMATRIX ITEMS=4,CK,TRIANGLE;
>EST NC=500 IT=25;
>END;
   :
```

[a]The text following the '⇐' is provided to help the reader understand the corresponding input.

Because the PC model is nested within the NR model we use a single `TEST` line to identify the model, but then invoke the PC model constraints on the appropriate item (i.e., item 4). These constraints (i.e., the `TMATRIX` and `FIX` lines) are similar to those that we used in Chapter 7 (Table 7.1), although the `ITEMS` subcommand is used in lieu of `ALL` to specify that the constraints should be applied only to item 4. (If we had more than one item to be calibrated using the PC model and wanted them to all have a common $\hat{\alpha}$, then it would be necessary to also include the command line `>EQUAL ITEMS=`<item list>, `AK=1;` as well as to specify the items with the `ITEMS` subcommand.) The corresponding output is found in Table 5.

MULTILOG took 36 cycles to achieve convergence. For the NR model we have 2*(4 – 1) = 6 parameters per item, and with three items this yields 18 parameters. With the PC model there are 4 parameters, one α and 3 transition location parameters. Therefore, the total number of parameters estimated by both models is 18 + 4 = 22. This corresponds to the `NUMBER OF FREE PARAMETERS` listed in the output. The estimates for items 1 through 3 are not dramatically different from those observed when all the items were treated nominally. Using the PC model, item 4 has an estimated item discrimination of 0.33 with transition location estimates of $\hat{\delta}_{41} = -0.68$, $\hat{\delta}_{42} = -0.10$, and $\hat{\delta}_{43} = 0.28$.

de Ayala

We can conceive of the NR–PC mixed model calibration as nested within the NR model calibration from the previous example. Calibrating all four items using only the NR model produced a –2ln*L* of 288.4 with 24 free parameters (or 256 – 24 – 1 = 231 degrees of freedom) and a BIC of 468.2797. With the mixed NR–PC model, calibration –2ln*L* increased to 320.5 with 22 free parameters or 233 degrees of freedom; BIC = 485.3897. The difference between these two -2ln*L*s is distributed as an χ^2 with two degrees of freedom (i.e., 233 – 231 = 24 – 22 = 2); we assume that the Full (nesting) model holds for the data. Because the critical χ^2 with two degrees of freedom (α = 0.05) is 5.99 and the difference between the –2ln*L*s is 32.1, we observe a significant increase in misfit by using the PC model for the last item. Furthermore, according to the BIC values the NR model is favored over the NR–PC mixed model calibration. Therefore, from a statistical perspective the NR model calibration is preferred to the NR–PC mixed model calibration. However, in some applications there may be pragmatic reasons for preferring the NR–PC mixed model calibration. For example, conceptually it may be more appealing to the various constituencies to treat item 4 in a graded fashion, particularly in light of the fact that the NR model does not fit the data in an absolute sense.

Figure 3 contains the ORFs for all the items using this mixed model approach; Figure 2 contains the corresponding ORFs based on the NR calibration. As would be expected, the ORFs for items 1 through 3 are very similar to those using the NR model. Item 4's ORFs show a slight spreading out, such that some of the response categories (e.g., category 3) have a wider θ range over which they are the most probable response category than when the item is treated nominally.

Table 5. Abridged output for the NR and PC mixed model calibration example.

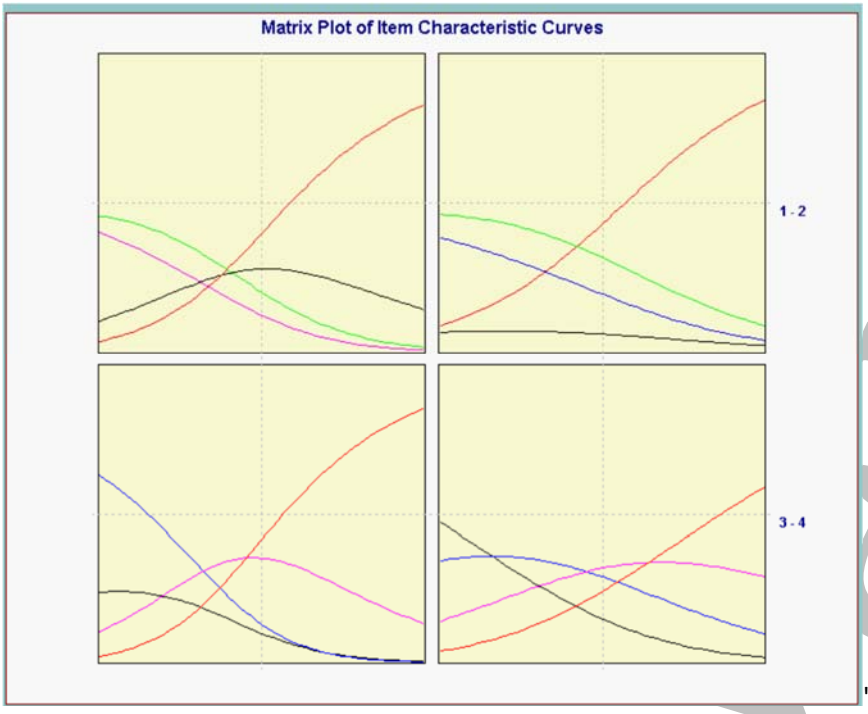
:

de Ayala

```
 NUMBER OF FREE PARAMETERS IS:   22
   :
FINISHED CYCLE 36
   :
ITEM SUMMARY
   :
ITEM   1:       4 NOMINAL CATEGORIES,  2 HIGH
 CATEGORY(K): 1       2       3       4
   A(K)       0.21    0.68   -0.51   -0.38
   C(K)       0.20    0.55   -0.62   -0.13

          CONTRAST-COEFFICIENTS (STANDARD ERRORS)
 FOR:              A                       C
 CONTRAST P(#)  COEFF.[ DEV.]  P(#)  COEFF.[ DEV.]
     1       1   0.47 (0.09)      4   0.36 (0.07)
     2       2  -0.72 (0.12)      5  -0.82 (0.12)
     3       3  -0.60 (0.11)      6  -0.33 (0.10)

@THETA:      INFORMATION:   (Theta values increase in steps of 0.2)
-3.0 - -1.6  0.083  0.094  0.106  0.119  0.133  0.148  0.163  0.177
-1.4 -  0.0  0.191  0.203  0.214  0.221  0.226  0.228  0.226  0.221
 0.2 -  1.6  0.213  0.203  0.190  0.177  0.162  0.148  0.134  0.120
 1.8 -  3.0  0.108  0.096  0.085  0.075  0.067  0.059  0.052

 OBSERVED AND EXPECTED COUNTS/PROPORTIONS IN
 CATEGORY(K):  1      2      3      4
 OBS. FREQ.    456    723    245    375
 OBS. PROP.  0.2535 0.4019 0.1362 0.2084
 EXP. PROP.  0.2535 0.4020 0.1362 0.2084
   :
ITEM   4:       4 NOMINAL CATEGORIES,  4 HIGH
 CATEGORY(K): 1       2       3       4
   A(K)      -0.50   -0.17    0.17    0.50
   C(K)       0.00    0.68    0.78    0.50

          CONTRAST-COEFFICIENTS (STANDARD ERRORS)
 FOR:              A                       C
 CONTRAST P(#)  COEFF.[POLY.]  P(#)  COEFF.[ TRI.]
     1      19   0.33 (0.04)     20  -0.68 (0.11)
     2      32   0.00 (0.00)     21  -0.10 (0.09)
     3      33   0.00 (0.00)     22   0.28 (0.09)

@THETA:      INFORMATION:   (Theta values increase in steps of 0.2)
-3.0 - -1.6  0.078  0.082  0.086  0.089  0.093  0.097  0.100  0.103
-1.4 -  0.0  0.105  0.108  0.109  0.111  0.111  0.112  0.111  0.110
 0.2 -  1.6  0.109  0.107  0.105  0.102  0.099  0.096  0.092  0.088
 1.8 -  3.0  0.084  0.080  0.076  0.072  0.068  0.065  0.061

 OBSERVED AND EXPECTED COUNTS/PROPORTIONS IN
 CATEGORY(K):  1      2      3      4
 OBS. FREQ.    288    507    551    453
 OBS. PROP.  0.1601 0.2818 0.3063 0.2518
 EXP. PROP.  0.1601 0.2818 0.3063 0.2518
   :
MARGINAL RELIABILITY:    0.3945
   :
 OBSERVED(EXPECTED)    STD.  :     EAP (S.D.)  :  PATTERN
                       RES.  :                 :
      1.0(     0.6)    0.51  :   -0.83 ( 0.77) :  1111
      2.0(     0.9)    1.10  :   -0.64 ( 0.76) :  1112
      0.0(     0.9)   -0.93  :   -0.44 ( 0.76) :  1113
      2.0(     0.6)    1.85  :   -0.25 ( 0.77) :  1114
   :
 NEGATIVE TWICE THE LOGLIKELIHOOD=       320.5
 (CHI-SQUARE FOR SEVERAL TIMES MORE EXAMINEES THAN CELLS)
```

de Ayala

Figure 3. ORFs for all four items from the NR and PC mixed model calibration.



Example: NR and PC Mixed Model Calibration of the Science Test, Collapsed Options, MMLE, MULTILOG

Inspection of the science test items' ORFs (e.g., Figure 2) indicate that item 2 contains a response option that is not very informative. Specifically, option 1 shows little discrimination capacity, and this option's frequency is substantially smaller than those of the item's other response options. As such, it is not very attractive to the examinees. Therefore, unless there are substantive reasons for maintaining four options for this item, one might consider collapsing option 1 with another option. In this example, we collapse item 2's option 1 with its option 2. As a result, the test consists of items 1 and 3, each with four nominal response options, item 2 with three nominal response options, and item 4 with four graded options. As in the previous example, the NR model is used for items 1 through 3, whereas the PC model is applied to item 4. Although we can edit the data file and recode item 2's options (i.e., option 2 becomes 1, option 3

de Ayala

becomes 2, and option 4 becomes 3), it is less error-prone to have MULTILOG do the recoding. The recoding within MULTILOG is performed similarly to what is done in the PC model calibration example (see MULTILOG_PCMcalibrationEx.pdf). That is, in the command file section that specifies which codes to use for which items, one substitutes the new codes for the old response codes. Recall that this section's format is that the columns represent the items and the rows represent each identified response code. Here is the basic structure of the command file's response code section:

```
                                                                    :
Number of response categories ⇒                                     4
Response code line (i.e., the codes in the data) ⇒                  1234
Codes to use for the identified response of "1" for each item ⇒     1111
Codes to use for the identified response of "2" for each item ⇒     2122
Codes to use for the identified response of "3" for each item ⇒     3233
Codes to use for the identified response of "4" for each item ⇒     4344
                                                                    :
                                                                    ⇑
                                                  Columns represent items 1 through 4
```

In the current context, the second column (representing item 2) reflects the recoding of the observed codes of 4, 3, and 2 to be 3, 2, and 1, respectively (i.e., item 2's original response codes become 1, 2, and 3). After this recording, item 2 consists of three response categories, and this is reflected in the NC as well as the `HIGH` specifications on the `TEST` command line. The command file for this analysis is shown in Table 6, with the corresponding abridged output in Table 7.

Convergence is achieved in 36 cycles. The number of parameters estimated is 20 (i.e., 2(3 – 1) = 4 for item 2, 4 for item 4 , plus 12 for items 1 and 3). The item parameter estimates for items 1, 3, and 4 are, within rounding, the same as those from the NR–PC mixed model calibration. The combining of item 2's options 1 and 2 results in the item's OBS. FREQ of 457. Because this number reflects the sum of the uncombined options 1 and 2's frequencies, we have

de Ayala

assurance that the options were correctly combined. The previous calibration showed that the standard errors for the item parameter estimates were in the range of 0.15 to 0.16. As can be seen, by combining these two options the standard errors have decreased to the range of 0.08 to 0.09. Therefore, by combining these low-frequency options we are able to obtain more accurate estimates. For completeness the ORFs are presented in Figure 4. As would be expected, except for item 2, these match those presented in Figure 3. Items 1 and 2 still show that they are behaving in a binary fashion (i.e., in general, examinees are either correctly answering the question or they are not). In the Appendix Example: Mixed Model Calibration of the Science Test—NR and 2PL Models, MMLE, MULTILOG we explore the effect of dichotomizing items 1 and 2.

Table 6. Command file for the NR and PC mixed model calibration; Collapsed Options.[a]

```
MULTILOG for Windows 7.00.2327.2
MIX NR & PC MODELS; ITEM 4 PC MODEL; ITEM 2-3 OPTIONS
>PROBLEM RANDOM,
         PATTERNS,
         DATA = 'C:SCIENCE2.PAT',
         NITEMS = 4,
         NGROUPS = 1,
         NPATTERNS = 256,
         NCHARS = 9;
>TEST ITEMS=(1,2,3,4),
      NOMINAL,
      NC = (4,3,4,4),
      HIGH = (2,2,4,4);
>TMATRIX ITEMS=4,AK,POLYNOMIAL;
>FIX      ITEMS=4,AK=(2,3),VALUE=0.0;
>TMATRIX ITEMS=4,CK,TRIANGLE;
>EST NC=500 IT=25;
>END ;
4                                   ⇐ Number of response categories
1234                                ⇐ The response code line
1111                                ⇐ Response code substitution for item 2: 1 stays as 1
2122                                ⇐ Response code substitution for item 2: 2 becomes a 1
3233                                ⇐ Response code substitution for item 2:  3 becomes a 2
4344                                ⇐ Response code substitution for item 2: 4 becomes a 3
(9A1,T1,4A1,F5.0)
```

[a]The text following the '⇐' is provided to help the reader understand the corresponding input.

Rather than using MULTILOG to recode item 2, if one edits the original response data to

de Ayala

recode item 2 to have the three response categories described above, then one obtains 192 patterns. The calibration of these data yields the same item parameter estimates as shown in Table 7. However, we now obtain a valid $-2\ln L$ of 219.6 with a BIC of 369.4997 and 20 estimated parameters. These values reflect a substantial improvement in fit over simply using only the NR model or the mixed NR–PC model's calibrations.

Table 7. Abridged output for the NR and PC mixed model calibration example; Collapsed Options.

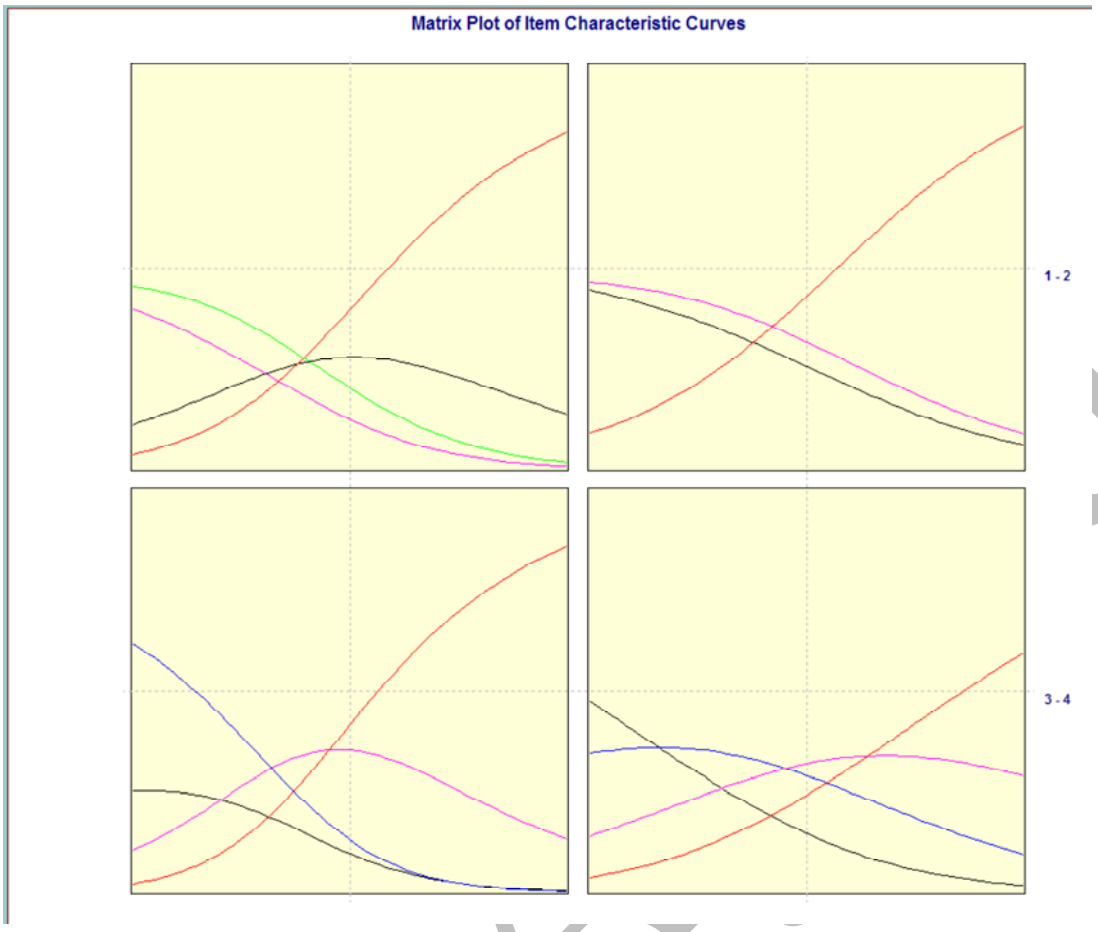
```
  :
 NUMBER OF FREE PARAMETERS IS:   20
  :
FIRST OBSERVATION AS READ-

ID    1111    4
ITEMS 1111
NORML      0.000
WT/CR       4.00

FINISHED CYCLE  36
  :
ITEM SUMMARY
  :
ITEM   1:       4 NOMINAL CATEGORIES,  2 HIGH
 CATEGORY(K): 1      2      3      4
   A(K)      0.20   0.68  -0.50  -0.38
   C(K)      0.20   0.55  -0.62  -0.13

          CONTRAST-COEFFICIENTS (STANDARD ERRORS)
 FOR:              A                     C
 CONTRAST P(#)  COEFF.[ DEV.]  P(#)  COEFF.[ DEV.]
     1       1   0.48 (0.09)      4   0.36 (0.09)
     2       2  -0.70 (0.12)      5  -0.81 (0.12)
     3       3  -0.58 (0.11)      6  -0.33 (0.10)

@THETA:       INFORMATION:   (Theta values increase in steps of 0.2)
-3.0 - -1.6  0.081  0.092  0.104  0.117  0.130  0.144  0.159  0.173
-1.4 -  0.0  0.187  0.199  0.210  0.218  0.223  0.225  0.224  0.220
 0.2 -  1.6  0.212  0.203  0.191  0.178  0.164  0.150  0.136  0.123
 1.8 -  3.0  0.110  0.098  0.087  0.077  0.068  0.061  0.054

 OBSERVED AND EXPECTED COUNTS/PROPORTIONS IN
 CATEGORY(K):  1      2      3      4
 OBS. FREQ.    456    723    245    375
 OBS. PROP.  0.2535 0.4019 0.1362 0.2084
 EXP. PROP.  0.2535 0.4020 0.1362 0.2084


ITEM   2:       3 NOMINAL CATEGORIES,  2 HIGH
 CATEGORY(K): 1      2      3
   A(K)     -0.25   0.45  -0.20
   C(K)     -0.25   0.28  -0.03

          CONTRAST-COEFFICIENTS (STANDARD ERRORS)
 FOR:              A                     C
 CONTRAST P(#)  COEFF.[ DEV.]  P(#)  COEFF.[ DEV.]
     1       7   0.71 (0.09)      9   0.52 (0.08)
     2       8   0.06 (0.09)     10   0.21 (0.09)
```

de Ayala

```
@THETA:      INFORMATION:   (Theta values increase in steps of 0.2)
-3.0 - -1.6  0.038  0.043  0.047  0.052  0.058  0.063  0.069  0.075
-1.4 -  0.0  0.081  0.087  0.092  0.098  0.102  0.106  0.110  0.112
 0.2 -  1.6  0.114  0.114  0.114  0.112  0.110  0.106  0.102  0.098
 1.8 -  3.0  0.092  0.087  0.081  0.075  0.069  0.063  0.058

 OBSERVED AND EXPECTED COUNTS/PROPORTIONS IN
 CATEGORY(K):  1      2      3
 OBS. FREQ.    457    785    557
 OBS. PROP.  0.2540 0.4364 0.3096
 EXP. PROP.  0.2540 0.4364 0.3096
   :
ITEM   4:       4 NOMINAL CATEGORIES,  4 HIGH
 CATEGORY(K): 1      2      3      4
   A(K)     -0.50  -0.17   0.17   0.50
   C(K)      0.00   0.68   0.78   0.50

          CONTRAST-COEFFICIENTS (STANDARD ERRORS)
 FOR:              A                    C
 CONTRAST P(#)  COEFF.[POLY.]  P(#)  COEFF.[ TRI.]
     1      17   0.33 (0.04)     18  -0.68 (0.11)
     2      29   0.00 (0.00)     19  -0.10 (0.09)
     3      30   0.00 (0.00)     20   0.28 (0.09)

@THETA:      INFORMATION:   (Theta values increase in steps of 0.2)
-3.0 - -1.6  0.078  0.082  0.086  0.090  0.094  0.097  0.101  0.104
-1.4 -  0.0  0.106  0.108  0.110  0.111  0.112  0.112  0.112  0.111
 0.2 -  1.6  0.110  0.108  0.106  0.103  0.100  0.096  0.093  0.089
 1.8 -  3.0  0.085  0.081  0.077  0.073  0.069  0.065  0.061

 OBSERVED AND EXPECTED COUNTS/PROPORTIONS IN
 CATEGORY(K):  1      2      3      4
 OBS. FREQ.    288    507    551    453
 OBS. PROP.  0.1601 0.2818 0.3063 0.2518
 EXP. PROP.  0.1601 0.2818 0.3063 0.2518
   :
@THETA:      INFORMATION:
-3.0 - -1.6  1.315  1.353  1.395  1.440  1.487  1.536  1.586  1.633
-1.4 -  0.0  1.677  1.714  1.743  1.763  1.772  1.770  1.758  1.737
 0.2 -  1.6  1.708  1.673  1.634  1.593  1.551  1.510  1.469  1.430
 1.8 -  3.0  1.393  1.359  1.327  1.297  1.270  1.244  1.222
   :
 MARGINAL RELIABILITY:    0.3938
   :
 OBSERVED(EXPECTED)    STD.  :     EAP (S.D.)   :  PATTERN
                       RES.  :                  :

      1.0(     3.0)    -1.14  :    -0.94 ( 0.77) :  1111
      2.0(     4.4)    -1.15  :    -0.75 ( 0.77) :  1112
      0.0(     3.9)    -1.99  :    -0.55 ( 0.77) :  1113
      2.0(     2.6)    -0.35  :    -0.35 ( 0.77) :  1114
   :
NEGATIVE TWICE THE LOGLIKELIHOOD=      -171.2
   :
```

Figure 4. ORFs for all four items from the NR and PC mixed model calibration; Item 2: Collapsed Options.

de Ayala


Matrix Plot of Item Characteristic Curves
1 - 2
3 - 4

NOTES

1. It is possible to use non-nested mixed models in a single calibration. For instance, we might use the GR model instead of the PC model for this example. However, because the example treats only one item in a graded fashion, there is no inherent benefit in using a model that allows items to vary in their discrimination over a model that constrains them to all be the same. We present the NR and GR mixed model calibration for the science test in the Appendix Example: Mixed Model Calibration of the Science Test—NR and GR Models, MMLE, MULTILOG.

de Ayala

Appendix

Example: Mixed Model Calibration of the Science Test—NR and 2PL Models, MMLE, MULTILOG

The analysis of the ORFs for the item set in Chapter 9 (Figure 2) indicates that the first two items were behaving in a binary fashion. Therefore, in an attempt to improve the fit to the data the first two items are scored correct (i.e., a coded response of 1) and incorrect (i.e., a coded response of 0); we assume that there are no other reasons to keep these items' responses as polytomous data. As a result, the first two items are modeled with the 2PL model and the last two are modeled with the NR model. Although theoretically the 2PL model is subsumed by the NR model, MULTILOG's implementation does not allow the NR model calibration of items with fewer than three options; MULTILOG estimates the 2PL model as a special case of the GR model. Therefore, it is necessary to use two TEST lines to identify the items associated with the 2PL and NR models (see Table 8 for the command file). The first TEST line specifies the use of the 2PL model with items 1 and 2, whereas the second TEST line indicates the use of the NR model with items 3 and 4. After the dichotomization of the responses for items 1 and 2 there are 64 possible patterns across the item set, all of which are observed. Table 9 contains the abridged output.

Table 8. Command file for the NR and 2PL mixed model calibration.[a]

```
MULTILOG for Windows 7.00.2327.2
2PL CALIBRATION OF ITEMS 1 & 2; NR FOR ITEMS 3 & 4
>PROBLEM RANDOM,
         PATTERNS,
         DATA = 'C:CompPatMix.PAT',
         NITEMS = 4,
         NGROUPS = 1,
         NPATTERNS = 64,
         NCHARS = 9;
>TEST ITEMS = (1, 2),                ⇐ Identifying items 1-2  for calibration with 2PL
      L2;
>TEST ITEMS = (3, 4),                ⇐ Identifying items 3-4  for calibration with NR
      NOMINAL,
      NC = (4, 4),
```

de Ayala

```
      HIGH = (4, 3);
>EST NC=500 IT=25;
>END ;
4
1234
1111
2222
0033
0044
(9A1,T1,4A1,F5.0)
```

[a]The text following the '⇐' is provided to help the reader understand the corresponding input.

Convergence is achieved in 26 cycles. Using the 2PL model, item 1 has an estimated item discrimination of 0.74 and a location estimate of 0.60, whereas for item 2, $\hat{\alpha}_2$ = 0.81 and $\hat{\delta}_2$ = 0.36. Because of a difference in the metrics of the NR model-only calibration and the way this mixed model calibration is implemented, we do not expect the item parameter estimates for items 3 and 4 to be the same as those resulting from NR model-only calibration. The –2ln*L* for this NR-2PL mixed model calibration is 51.8 with 16 free parameters (BIC = 171.7198). This –2ln*L* of 51.8 reflects a substantial improvement in fit over simply using the NR model for calibrating all four items. If there were no substantive reasons for treating the first two items as polytomous, then from simply a fit perspective items 1 and 2 should be treated as binary items. However, the cost of doing this is an instrument that provides less information below approximately 0.8 than when we calibrate it using solely the NR model (see Tables 2 and 4's `TOTAL TEST INFORMATION` sections). (The similarity in the estimates for items 3 and 4 from this calibration and those from the NR model-only calibration indicates that the metrics are in close alignment to one another, so the value 0.8 would not change by very much if the metrics are linked to one another.) Therefore, the accuracy of the location estimates for individuals less than 0.8 would be less with the mixed model approach than with the single NR model item parameter estimates. This begs the question, "Is it better to use more accurate person location estimates based on a

de Ayala

comparatively poorer-fitting model or to use less accurate $\hat{\theta}_S$ based on a better-fitting model?" If the purpose of the calibration is to estimate θ (in contrast to, for example, item pool construction), then the answer to this question would hinge, at least in part, on the validity evidence for the $\hat{\theta}_S$ based on the mixed model and on the NR model calibration results.

Table 9. Abridged output for the NR and 2PL mixed model calibration example.

```
     :
 NUMBER OF FREE PARAMETERS IS:   16
     :
FIRST OBSERVATION AS READ-

ID    2211    1
ITEMS 2211
NORML      0.000
WT/CR       1.00
     :
 FINISHED CYCLE  26
  MAXIMUM INTERCYCLE PARAMETER CHANGE=   0.00077 P(   7)

  ITEM SUMMARY
     :
ITEM   1:       2 GRADED CATEGORIES
       P(#) ESTIMATE (S.E.)
A         1    0.74  (0.08)
B( 1)     2    0.60  (0.10)

@THETA:      INFORMATION:   (Theta values increase in steps of 0.2)
-3.0 - -1.6  0.033  0.038  0.043  0.048  0.054  0.061  0.067  0.075
-1.4 -  0.0  0.082  0.090  0.097  0.105  0.112  0.119  0.124  0.129
 0.2 -  1.6  0.133  0.135  0.136  0.135  0.133  0.129  0.125  0.119
 1.8 -  3.0  0.112  0.105  0.098  0.090  0.082  0.075  0.068

 OBSERVED AND EXPECTED COUNTS/PROPORTIONS IN
 CATEGORY(K):  1      2
 OBS. FREQ.   1076    723
 OBS. PROP.  0.5981 0.4019
 EXP. PROP.  0.5981 0.4019

ITEM   2:       2 GRADED CATEGORIES
       P(#) ESTIMATE (S.E.)
A         3    0.81  (0.08)
B( 1)     4    0.36  (0.08)

@THETA:      INFORMATION:   (Theta values increase in steps of 0.2)
     :
 OBSERVED AND EXPECTED COUNTS/PROPORTIONS IN
 CATEGORY(K):  1      2
 OBS. FREQ.   1014    785
 OBS. PROP.  0.5636 0.4364
 EXP. PROP.  0.5636 0.4364

ITEM   3:       4 NOMINAL CATEGORIES,  4 HIGH
 CATEGORY(K): 1      2      3      4
   A(K)     -0.49  -0.54   0.29   0.74
   C(K)     -0.80  -0.41   0.51   0.70
     :
 OBSERVED AND EXPECTED COUNTS/PROPORTIONS IN
```

de Ayala

```
  CATEGORY(K):  1      2      3      4
  OBS. FREQ.    192    293    564    750
  OBS. PROP.  0.1067 0.1629 0.3135 0.4169
  EXP. PROP.  0.1067 0.1628 0.3135 0.4170

ITEM   4:       4 NOMINAL CATEGORIES,  3 HIGH
  CATEGORY(K): 1      2      3      4
    A(K)      -0.77   0.19  -0.11   0.69
    C(K)      -0.59   0.26   0.34  -0.01
     :
TOTAL TEST INFORMATION
@THETA:      INFORMATION:
 -3.0 - -1.6  1.328  1.368  1.409  1.453  1.497  1.541  1.584  1.625
 -1.4 -  0.0  1.661  1.692  1.717  1.735  1.745  1.747  1.742  1.730
  0.2 -  1.6  1.712  1.689  1.662  1.631  1.598  1.563  1.527  1.490
  1.8 -  3.0  1.454  1.418  1.383  1.350  1.318  1.288  1.260
     :
NEGATIVE TWICE THE LOGLIKELIHOOD=         51.8
     :
```

Example: Mixed Model Calibration of the Science Test—NR and GR Models, MMLE, MULTILOG

To demonstrate a mixed model calibration using non-nested models, we calibrate our science test (Chapter 9) using the GR and the NR models. The command file for performing this calibration is shown in Table 10. As can be seen, two TEST lines are used to identify which items are associated with which model. The first TEST line specifies the use of the NR model with items 1 through 3, whereas the second TEST line indicates that the GR model should be used with item 4. The basic structure of the command file parallels that seen with the NR model (MULTILOG_NRMcalibrationEx.pdf's Table 1) and for the GR model (MULTILOG_GRMcalibrationEx.pdf's Table 2). Performing this mixed model calibration produces a $-2\ln L$ of 322.2 with 22 free parameters (BIC = 487.0897); for the NR model there are 2*(4 – 1) = 6 parameters per item times 3 items, or 18 parameters, and for the GR model there are 4 parameters (i.e., one α and 3 category boundary locations), for a total number of estimated parameters of 18 + 4 = 22. Using the GR model for item 4 we have an estimated item discrimination of 0.60 with category boundary locations of $\hat{\delta}_{41}$ = –2.94, $\hat{\delta}_{42}$ = –0.42, and $\hat{\delta}_{43}$ =

de Ayala

1.95.

Table 10. Command file for the NR and GR mixed model calibration.[a]

```
MULTILOG for Windows 7.00.2327.2
MIX NR & GR MODELS; ITEM 4 - 3 CATEGORIES
>PROBLEM RANDOM,
         PATTERNS,
         DATA = 'C:SCIENCE.PAT',
         NITEMS = 4,
         NGROUPS = 1,
         NPATTERNS = 256,
         NCHARS = 9;
>TEST ITEMS=(1,2,3),                          ⇐ Identifying items 1-3  for calibration
      NOMINAL,                                     with NR model
      NC = (4,4,4),
      HIGH = (2,3,4);
>TEST ITEMS=4,                                ⇐ Identifying item 4  for calibration
      GR,                                          with GR model
      NC = 4;
>EST NC=500 IT=25;
>END ;
   :
```

[a]The text following the '⇐' is provided to help the reader understand the corresponding input.

de Ayala