Two-stage spelling test

This example is based on a study by Bock and Zimowski (1998). The full document is available on the Internet from the American Institutes for Research. As a small computing example, we simulated two-stage testing in data for the "One-Hundred Word Spelling Test" previously analyzed by Bock, Thissen, and Zimowski (1997).

On the basis of item parameters they reported, we selected 12 first-stage items and 12 items for each of three levels of the second-stage test.

Because of the limited number of items in the pool, we could not meet exactly the requirements of the prototype design, but the resulting test illustrates well enough the main features of the analysis. The item numbers in this and a later example correspond to the words presented in Bock, Thissen, and Zimowski's (1977) Table 1 in the NAEP report. All computations in the analysis were carried out with the BILOG-MG program of Zimowski, Muraki, Mislevy and Bock (1996).

For assigning the cases in the data to second-stage levels under conditions that would apply in an operational assessment, we re-estimated the parameters for the 12 first-stage items, computed Bayes estimates of proficiency scale scores, and rescaled the scores to mean 0 and standard deviation 1 in the sample. The command file **step0.blm**, shown below, contains the necessary commands.

```
STEP0.BLM - A SIMULATED TWO-STAGE SPELLING TEST - Prototype 1 computing
            example. Estimation of the 12 first-stage item parameters.
>COMMENTS
From: "Feasibility Studies of Two-Stage Testing in Large-Scale Educational
Assessment: Implications for NAEP" by R. Darrel Bock and Michele F. Zimowski,
May 1998, American Institutes for Research.

Based on the 100-word spelling test data. N = 1000
   (See Bock, Thissen and Zimowski, 1997).

According to page 35 of the NAEP study, we first establish group membership by recalibrating
the parameters for the 12 first-stage items and compute EAP estimates of the proficiency
scale scores, rescaled to mean 0 and standard deviation 1 in the sample of 1000. Next, we
assign group membership based on scores at or below -0.67 (group 1), at or above +0.67 (group
3), and the remaining scores (group 2).

The resulting score file was manipulated per these instructions(see result in the STEP0.EAP
file) and the assigned group membership added to the original data file as column 12 (before
empty). The resulting split is: group 1 236, group 2 531, group 3 233.

>GLOBAL   NPARM=2,  DFNAME='SPELL1.DAT', SAVE;
>SAVE     PARM='STEP0.PAR', SCORE='STEP0.SCO';
>LENGTH   NITEMS=12;
>INPUT    NTOTAL=100, NIDCH=11, TYPE=1, SAMPLE=1000, KFNAME='SPELL1.DAT';
>ITEMS    INUM=(1(1)100), INAME=(SPELL001(1)SPELL100);
>TEST     TNAME=SPELLING, INUM=(1,4,8,10,23,25,28,29,39,47,59,87);
```

```
(11A1,1X,25A1,1X,25A1/12X,25A1,1X,25A1)
>CALIB    NQPT=20, CRIT=0.001, CYCLES=100, NEWTON=2, NOFLOAT;
>SCORE    IDIST=3, METHOD=2, NOPRINT, INFO=1, POP;
```

Cases with scores at or below -0.67 were assigned to group 1. Those at or above +0.67 were assigned to group 3, and the remainder to group 2. Of the 1000 cases in the original study, 274, 451, and 275 were assigned to groups 1, 2, and 3, respectively. With these assignment codes inserted in the case records, the latent distributions were estimated using the command file for the first-stage analysis shown below (**step1.blm**).

```
STEP1.BLM - ANALYSIS 1: A SIMULATED TWO-STAGE SPELLING TEST
            Estimation of first-stage item parameters and latent distributions.
>GLOBAL  DFNAME='SPELL2.DAT', NPARM=2, SAVE;
>SAVE    SCORE='STEP1.SCO', PARM='STEP1.PAR';
>LENGTH  NITEMS=12;
>INPUT   NTOT=100, SAMPLE=1000, NGROUP=3, KFNAME='SPELL2.DAT', NIDCHAR=11,
         TYPE=1;
>ITEMS   INUMBERS=(1(1)100), INAMES=(SPELL001(1)SPELL100);
>TEST    TNAME=SPELLING, INUM=(1,4,8,10,23,25,28,29,39,47,59,87);
>GROUP1  GNAME=GROUP1, LENGTH=12, INUM=(1,4,8,10,23,25,28,29,39,47,59,87);
>GROUP2  GNAME=GROUP2, LENGTH=12, INUM=(1,4,8,10,23,25,28,29,39,47,59,87);
>GROUP3  GNAME=GROUP3, LENGTH=12, INUM=(1,4,8,10,23,25,28,29,39,47,59,87);
(11A1,I1,25A1,1X,25A1,/T13,25A1,1X,25A1)
>CALIB   FIX, NOFLOAT, NQPT=20, CYCLE=35, SPRIOR, NEWTON=2, CRIT=0.001, REF=0;
>SCORE   IDIST=3, METHOD=2, NOPRINT, INFO=1, POP;
```

For the second-stage analysis, we used the latent distributions estimated in the first-stage analysis as the prior distributions for maximum marginal likelihood analysis of the combined first- and second-stage data. The points and weights representing the distributions are shown in the corresponding BILOG-MG command file.

Inasmuch as there are no second-stage link items in this example, we use the first-stage items as an anchor test. The six easiest of these items provide the links between levels 1 and 2; the six most difficult provide the links between levels 2 and 3.

The syntax for this analysis is given in **step2.blm**, as shown below.

```
STEP2.BLM - ANALYSIS 2: A SIMULATED TWO-STAGE SPELLING TEST. Estimated link
            and second-stage item parameters, and latent distributions.
>COMMENTS
         The points and weights are the posterior estimates from STEP1.PH2.

>GLOBAL  DFNAME='SPELL2.DAT', NPARM=2, SAVE;
>SAVE    SCORE='STEP2.SCO', PARM='STEP2.PAR';
>LENGTH  NITEMS=48;
>INPUT   NTOT=100, SAMPLE=1000, NGROUP=3, KFNAME='SPELL2.DAT', NIDCHAR=11,
         TYPE=1;
>ITEMS   INUM=(1(1)100), INAME=(SPELL001(1)SPELL100);
>TEST    TNAME=SPELLING,
         INUM=( 1, 4, 5, 6, 7, 8, 9,10,12,14,15,17,20,23,24,25,
               26,27,28,29,33,34,35,38,39,46,47,48,49,50,53,54,
               59,60,64,68,69,72,73,77,78,84,85,86,87,90,95,97);
>GROUP1  GNAME=GROUP1, LENGTH=18,
         INUM=( 1, 4, 5,14,24,26,29,38,39,46,53,59,68,78,85,87,90,95);
>GROUP2  GNAME=GROUP2, LENGTH=24,
         INUM=( 1, 4, 8, 9,10,15,20,23,25,27,28,29,33,34,39,47,48,49,
               50,54,59,64,72,87);
>GROUP3  GNAME=GROUP3, LENGTH=18,
         INUM=( 6, 7, 8,10,12,17,23,25,28,35,47,60,69,73,77,84,86,97);
(11A1,I1,25A1,1X,25A1,/T13,25A1,1X,25A1)
>CALIB   IDIST=1, FIX, NOFLOAT, CYCLE=35, SPRIOR, NEWTON=2, CRIT=0.001,
         NQPT=20, REF=0, PLOT=1.0, ACC=0.0;
>QUAD1 POINT=(-0.4081E+01, -0.3652E+01, -0.3222E+01, -0.2792E+01,
              -0.2363E+01, -0.1933E+01, -0.1504E+01, -0.1074E+01,
              -0.6443E+00, -0.2147E+00,   0.2150E+00, 0.6446E+00,
               0.1074E+01,  0.1504E+01,  0.1933E+01, 0.2363E+01,
```

```
                0.2793E+01,  0.3222E+01,  0.3652E+01,  0.4082E+01),
      WEIGHT= (0.2345E-03, 0.1159E-02, 0.4738E-02, 0.1624E-01,
               0.4605E-01, 0.1077E+00, 0.2023E+00, 0.2785E+00,
               0.2311E+00, 0.9390E-01, 0.1678E-01, 0.1320E-02,
               0.4924E-04, 0.9717E-06, 0.8556E-12, 0.0000E+00,
               0.0000E+00, 0.0000E+00, 0.0000E+00, 0.0000E+00);
>QUAD2 POINT=(-0.4081E+01, -0.3652E+01, -0.3222E+01, -0.2792E+01,
              -0.2363E+01, -0.1933E+01, -0.1504E+01, -0.1074E+01,
              -0.6443E+00, -0.2147E+00, 0.2150E+00, 0.6446E+00,
               0.1074E+01, 0.1504E+01, 0.1933E+01, 0.2363E+01,
               0.2793E+01, 0.3222E+01, 0.3652E+01, 0.4082E+01),
      WEIGHT=(0.0000E+00, 0.0000E+00, 0.0000E+00, 0.3055E-05,
               0.7882E-04, 0.1170E-02, 0.1119E-01, 0.6218E-01,
               0.1820E+00, 0.2791E+00, 0.2502E+00, 0.1451E+00,
               0.5407E-01, 0.1271E-01, 0.1945E-02, 0.2046E-03,
               0.8579E-06, 0.0000E+00, 0.0000E+00, 0.0000E+00);
>QUAD3 POINT=(-0.4081E+01, -0.3652E+01, -0.3222E+01, -0.2792E+01,
              -0.2363E+01, -0.1933E+01, -0.1504E+01, -0.1074E+01,
              -0.6443E+00, -0.2147E+00, 0.2150E+00, 0.6446E+00,
               0.1074E+01, 0.1504E+01, 0.1933E+01, 0.2363E+01,
               0.2793E+01, 0.3222E+01, 0.3652E+01, 0.4082E+01),
      WEIGHT= (0.0000E+00, 0.0000E+00, 0.0000E+00, 0.0000E+00,
               0.0000E+00, 0.3914E-11, 0.1006E-05, 0.5966E-04,
               0.1650E-02, 0.1943E-01, 0.9720E-01, 0.2237E+00,
               0.2717E+00, 0.2051E+00, 0.1111E+00, 0.4735E-01,
               0.1652E-01, 0.4763E-02, 0.1128E-02, 0.2324E-03);
>SCORE   IDIST=3, METHOD=2, NOPRINT, INFO=1, POP;
```

Since the spelling data contain responses of all cases to all items, we can examine the comparative accuracy of the estimates based on the 24 items per case in the two-stage data with those based on 48 items per case in a conventional one-stage test. Syntax is as given in **step3.blm**, shown below.

```
STEP3 - ANALYSIS 2: A SIMULATED TWO-STAGE SPELLING TEST
        Estimation of 48 one-stage item parameters, and latent distributions.
>GLOBAL  DFNAME='SPELL2.DAT', NPARM=2, SAVE;
>SAVE    SCORE='STEP3.SCO', PARM='STEP3.PAR';
>LENGTH  NITEMS=48;
>INPUT   NTOTAL=100, SAMPLE=1000, KFNAME='SPELL2.DAT', NIDCHAR=11, TYPE=1;
>ITEMS   INUMBERS=(1(1)100), INAMES=(SPELL001(1)SPELL100);
>TEST    TNAME=SPELLING, INUM=(1,4,5,6,7,8,9,10,12,14,15,17,20,23,24,25,
         26,27,28,29,33,34,35,38,39,46,47,48,49,50,53,54,59,60,64,68,69,72,73,
         77,78,84,85,86,87,90,95,97);
(11A1,1X,25A1,1X,25A1,/T13,25A1,1X,25A1)
>CALIB   IDIST=0, FIX, NOFLOAT, CYCLE=35, SPRIOR, NEWTON=2, CRIT=0.001,
         NQPT=20, REF=0, PLOT=1.0, ACC=0.0;
>SCORE   IDIST=3, METHOD=2, NOPRINT, INFO=1, POP;
```

The latter estimates are also shown in Table 1. Despite the small number of items and relatively small sample size in this computing example, the agreement between the estimates is reasonably good for the majority of items. There are notable exceptions, however, among the second-stage items: of these, items 6, 7, 77, and 84 show discrepancies in both slope and threshold; all of these are from level 3 and have extremely high thresholds in the one-stage analysis, well beyond the +1.5 maximum we are assuming for second-stage items. Items 12 and 17 from level 3 are discrepant only in slope, as are items 26 and 38 from level 2, and items 50 and 64 from level 1.

In all cases the two-stage slope is larger than the one-stage slope. This effect is balanced however, by the tendency of the first-stage items, 1, 4, 8, 10, 23, 25, 28, 29, 39, 47, 59, and 87, to show smaller slopes in the two-stage analysis. As a result, the average slope in the two-stage results is only slightly larger than the one-stage average.

Table 1: Comparison of two-stage and one-stage item parameter estimates in the spelling data (shown for first 10 items)

Item	Two-stage		One-stage	
	Slope (S.E.)	**Threshold (S.E.)**	**Slope (S.E.)**	**Threshold (S.E.)**
SPELL001	0.74191 0.10040	-0.22896 0.07910	0.84646 0.08642	-0.32964 0.06612
SPELL004	0.64140 0.08831	-0.45195 0.09150	0.71193 0.07347	-0.54128 0.08305
SPELL005	0.68036 0.19351	-1.47582 0.16286	0.69276 0.07525	-1.40895 0.13561
SPELL006	0.87969 0.24184	1.51254 0.13287	0.29534 0.04648	2.15957 0.37184
SPELL007	0.78362 0.24146	2.59105 0.37885	0.32823 0.06116	3.76009 0.67776
SPELL008	0.51257 0.07726	0.52107 0.11154	0.54531 0.06226	0.59135 0.10754
SPELL009	0.98121 0.19997	-0.28826 0.08066	0.68981 0.06884	-0.25449 0.07895
SPELL010	0.94877 0.10159	0.45341 0.06703	0.91421 0.08021	0.50198 0.06909
SPELL012	0.87810 0.23453	1.41514 0.11948	0.78199 0.09203	1.41415 0.13032
SPELL014	1.00579 0.28436	-1.99060 0.20872	0.72159 0.10121	-1.94803 0.20793

The average thresholds also show only a small difference. In principle, the parameters of a two-parameter logistic response function can be calculated from probabilities at any two distinct, finite values on the measurement continuum. Similarly, those of the three-parameter model can be calculated from three such points. This suggests that in fallible data estimation must improve, even in the two-stage case, as sample size increases. Some preliminary simulations we have attempted suggest that with sample sizes in the order of 5 or 10 thousand, and better placing of the items, the discrepancies we see in the prototype 1 results largely disappear.

The latent distributions estimated with items from both stages are depicted in Figure 1. The distributions for the three assignment groups are shown normalized to unity. The estimated population distribution, which is the sum of the distributions for the individual groups weighted proportional to sample size, is constrained to mean 0 and standard deviation 1 during estimation of the component distribution. It is essentially normal and almost identical to the population distribution estimated in the one-stage analysis.

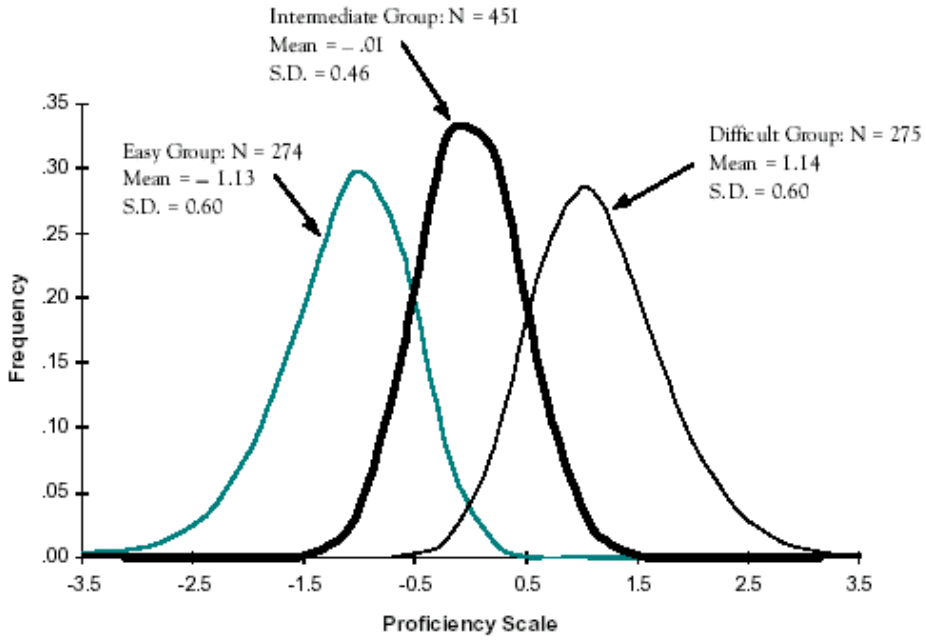

Figure 1. Prototype 1: estimated latent distributions from two-stage and one-stage spelling data

One may infer the measurement properties of the simulated two-stage spelling test from the information and efficiency calculations shown in Figure 2 and Figure 3, respectively. When interpreting information curves, the following rules of thumb are helpful. An information value of 5 corresponds to a measurement error variance of 1/5 = 0.2. In a population in which the score variance is set to unity, the reliability of a score with this error variance is 1.0 - 0.2 = 0.8. Similarly, the reliability corresponding to an information value of 10 is 0.9. In the context of low-stakes score reporting, we are aiming for reliabilities anywhere between these figures. As is apparent in Figure 2, this range of reliability is achieved in the two-stage results for spelling over much of the latent distribution.

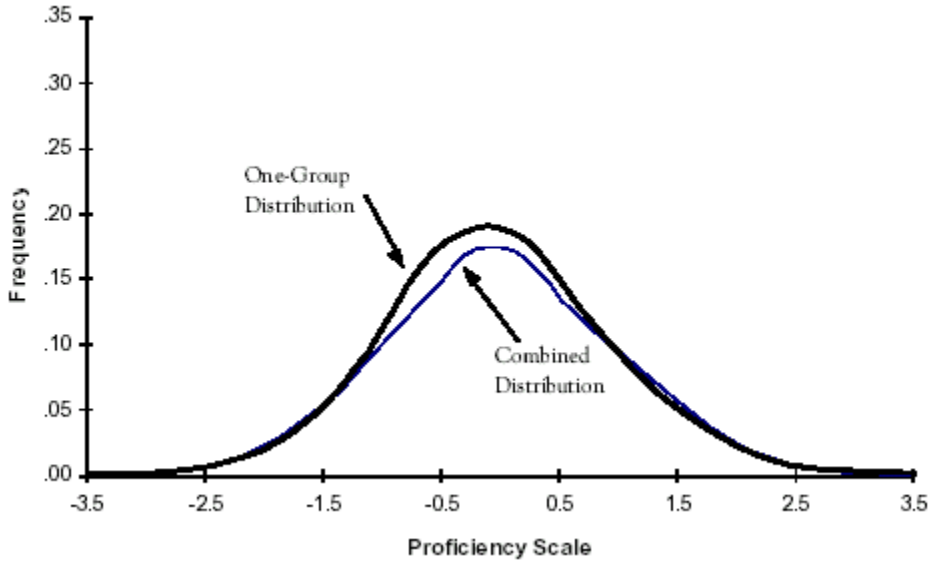

Figure 2. Prototype 1: two-stage spelling test

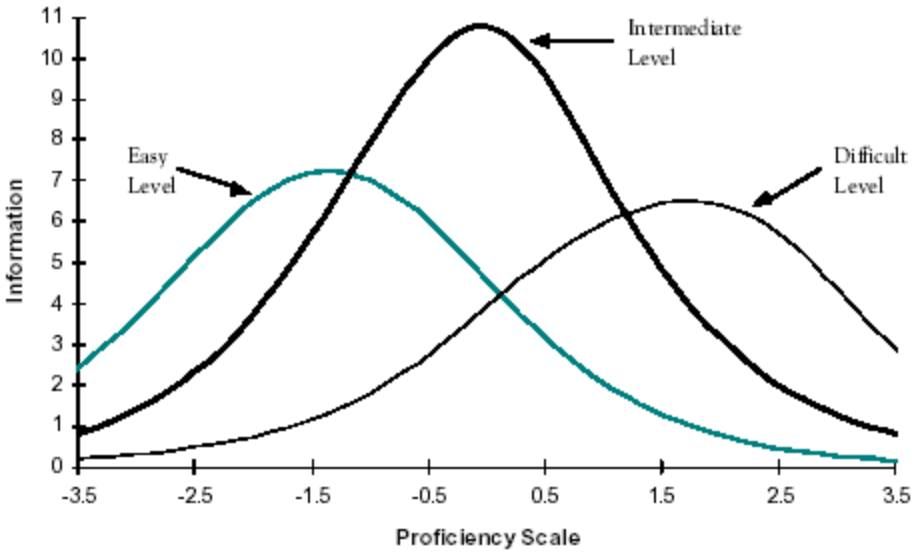

Figure 3. Prototype 1: efficiencies of the two-stage spelling tests

Finally, the efficiency curves in Figure 3 for the three levels show us the saving of test length and administration time, including both first- and second-stage testing, due specifically to the two-stage procedure in comparison with a one-stage test of the same length and item content.

In this case we hope to see efficiencies greater than 2.0, at least away from the population mean where conventional tests with peaked centers typically have reduced precision. The prototype 1 design and analysis meet this criterion.

To increase generalizability of group-level mean scores in assessment applications of the prototype 1 design, the second-stage tests will of course have to exist in multiple stratified randomly parallel forms. As with matrix sampling designs, these forms will be administered in random rotation to the examinees in each second-stage level. The sample data will then be suitable for equivalent-groups equating of the second-stage forms.