Edited by Mathilda du Toit.

Cover by Clint Smith of Esse Group. Based on a design by Louis Sullivan for elevator grillwork in the Chicago Stock Exchange (1893).

1 2 3 4 5 6 7 8 9 0 08 07 06 05 04 03 02

ISBN: 0-89498-053-X

Please download the examples from our website and unzip them into a convenient folder location.

Table of Contents

1 BILOG-MG

1.1 Introduction

BILOG-MG is an extension of the BILOG program that is designed for the efficient analysis of binary items, including multiple-choice or short-answer items scored right, wrong, omitted, or not-presented. BILOG-MG is capable of large-scale production applications with unlimited numbers of items or respondents. It can perform item analysis and scoring of any number of subtests or subscales in a single program run. All the program output may be directed to text files for purposes of selecting items or preparing reports of test scores.

The BILOG-MG program implements an extension of Item Response Theory (IRT) to multiple groups of respondents. It has many applications in test development and maintenance. Applications of multiple-group item response theory in education assessment and other large-scale testing programs include:

- Nonequivalent groups equating for maintaining the comparability of scale scores as new forms of the test are developed.
- Vertical equating of test forms across school grades or age groups.
- Analysis of Differential Item Functioning (DIF) associated with demographic or other group differences.
- Detecting and correcting for item parameter trends over time (DRIFT).
- Calibrating and scoring tests in two-stage testing procedures designed to reduce total testing time.
- Estimating latent ability or proficiency distributions of students in schools, communities, or other aggregations.

In addition, the BILOG-MG program provides for "variant items" that are inserted in tests for purpose of estimating item statistics, but that are not included in the scores of the examinees. Additional information on these applications are provided in Section 1.2.

Syntax can be generated or adapted using menus and dialog boxes or, as before, with command files in text format. The interface has menu options in the order the user would most generally use: model specification is followed by data specification and technical specifications, etc. Each of the menu options provides access to a number of dialog boxes in which the user can make specifications. For an overview of the required and optional commands in BILOG-MG syntax, please see Section 3.2.1. For more information on which dialog box to use to specify a specific keyword or option, please see the location of keywords in the interface discussed in Section 2.13.

- **Filename length:** All filenames with path may now extend to 256 characters. The filename must be enclosed in single quotes. Note that each line of the command file has a maximum length of 80 characters. If the filename does not fit on one line of 80 characters, the remaining characters should be placed on the next line, starting at column 1.

- **Factor loadings:** The item dispersion (reciprocal of the item slope) previously listed among the parameter estimates has been replaced by the one-factor item factor loading given by the expression $Slope/\sqrt{1+Slope^2}$.
- **Average measurement error and empirical reliability for each subtest:** The mean-square error and root-mean-square error for the sample cases are listed for each test. In addition, the empirical reliability computed from the IRT scale score variance and the mean-square error is listed.

  Note that for EAP and MAP estimated ability the formula for this reliability differs from the formula for ML estimated ability (to account for the regression effect in EAP and MAP estimation). If there are multiple test forms, these test statistics are averages over the forms. If there are multiple groups, the statistics are listed for both the combined groups and the separate groups.

- **Reliabilities in connection with information plots:** The reliabilities given by the program in connection with the information plots of Phase 3 differ from empirical reliabilities in that they assume a normal distribution of ability in the population. They depend only on the parameters of the items and not on the estimated abilities in the sample. The program now computes and lists these theoretical reliabilities for both combined and separated test forms and sample groups. (For a discussion of empirical and theoretical reliability see Bock & Zimowski (1999).)
- **Information curves and reliabilities for putative test forms:** It may be useful in test development to preview the information and theoretical reliability of test forms that might be constructed from items drawn from a calibrated item bank. (For a discussion of this procedure, see Section 1.2.)
- **GLOBAL command—PRNAME keyword:** This keyword instructs the program to read the provisional values of parameters of selected items in the test forms from the specified file.
- **SAVE command—PDISTRIB keyword:** This keyword allows the user to save the points and weights of the posterior latent distribution at the end of Phase 2. These quantities can be included as prior values following the SCORE command for later EAP estimation of ability from previously estimated item parameters.
- **TEST command—FIX keyword:** This keyword allows the user to keep selected item parameters fixed at their starting values. Starting values may be entered on the SLOPE, THRESHLD, and GUESSING keywords on the same command or read from an existing item parameter file.
- **CALIB command—NOADJUST option:** BILOG-MG routinely rescales the origin and scale of the latent distribution, even in the one-group case. This option may be used to suppress this adjustment.
- **CALIB command—CHI keyword:** This keyword determines the number of items required and the number of intervals used for χ^2 computations.
- **CALIB command—FIXED option:** If this option is present, the prior distributions of ability in the population of respondents are kept fixed at the values specified in the IDIST keyword and/or the QUAD commands. It suppressed the updating of the means and standard deviations of the prior distribution at each EM cycle in the multiple-group case.

- **CALIB command—GROUP-PLOTS option:** By default, the program item plots show observed proportions of correct responses in the data combined for all groups. The GROUP-PLOTS option provides plots for each separate group, along with the combined plot.
- **CALIB command—RASCH option:** If this option is specified, the parameter estimates will be rescaled according to Rasch model conventions: that is, all the slopes will be rescaled so that their geometric mean equals 1.0, and the thresholds will be rescaled so that their arithmetic mean equals 0.0. If the 1-parameter model has been specified, all slope parameters will therefore equal 1.0.
- **PRIORS command—SMU and SSIGMA keywords:** Prior values for slope parameter means and sigma are now entered in arithmetic units rather than natural log units. The means for both forms are printed in the Phase 2 output, however. The default for SMU is 1.0 (log SMU = 0.0) and for SSIGMA the default is 1.64872127 (log SSIGMA = 0.5).
- **SCORE command—MOMENTS option:** Inserting the MOMENTS option in the SCORE command causes the program to compute and list the coefficients of skewness and kurtosis of the ability estimates and of the latent distribution.
- **SCORE command—DOMAIN keyword:** BILOG-MG now includes a procedure for converting the Phase 3 estimates of ability into domain scores if the user supplies a file containing the item parameters for a sample of previously calibrated items from the domain. Weights can be applied to the items to improve the representation of the domain specifications.
- **SCORE command—FILE keyword:** This keyword is used to supply the external file used to calculate the domain scores (see above).

1.2 Multiple-group analyses[1]

1.2.1 Background for multiple-group models

In the multiple-group case, it is assumed that the response function of any given item is the same for all groups of subjects. In the DIF and DRIFT applications, however, we allow the relative difficulties of the items to differ from one group to another or one occasion to another. In that case, the b_j parameters will differ between groups, and we will have to detect and estimate the differences. Even in the presence of DIF and DRIFT, however, it is assumed that the item discriminating powers are the same from one group to another. In the other applications, such as nonequivalent groups equating or two-stage testing, we assume that both the locations and the slope of items common to more than one group are equal. To satisfy this assumption, we would perform a preliminary DIF analysis and not use, in equating, items showing appreciable DIF.

The main difference between the single-group and multiple-group case is in the assumption about the latent distribution. In most equation situations, it is reasonable to assume that the re-

[1] This section was contributed by Michele Zimowski.

spondents in the sample groups are drawn from populations that are normal, but have different means and standard deviations (see Figure 1.1).

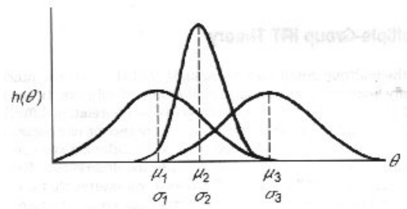


Figure 1.1: Normal latent densities in three populations

In that case, the item response data can be described completely by estimating the means and standard deviations of the groups along with the item parameters. One must, however, again contend with the arbitrary origin and unit of the latent continuum, and may resolve this indeterminacy either by setting the mean and standard deviation of one of the groups to any arbitrary values, or by setting the overall mean and variance of the combined distributions to arbitrary values. Both options are provided in BILOG-MG. The procedure for simultaneous estimation of item parameters and latent distributions in more than one group are described in Bock & Zimowski (1995) and in Mislevy (1987).

In two-stage testing applications, the situation is different. The groups correspond to examinees who have been selected on the basis of a first-stage test to receive second-stage test forms tailored to the provisional estimate of ability based on the first-stage test. Typically, the second-stage groups are determined by cutting points on the θ-scale of the pretest. Because the pretest score is a fallible criterion, the θ distributions of the second-stage groups may overlap to a considerable extent, but they cannot be expected to be normal even when the population from which the examinees originate is normal. More likely in these applications the latent distributions would appear as in Figure 1.2.

To accommodate such arbitrary shapes of distributions, one must make use of the empirical estimation procedure (see the section on estimation in the next chapter). As in the single-group case, these empirical distributions can be estimated along with the item parameters by marginal maximum likelihood. Again, the indeterminacy of location and scale must be resolved, either by setting the mean and standard deviation of one of the groups to convenient values, such as zero and one, or setting the overall mean and standard deviation of the combined distributions to similar values. In DIF analysis of ethnic effects, for example, the usual approach is to assign the mean and standard deviation of the reference group, which is usually the majority demographic group.

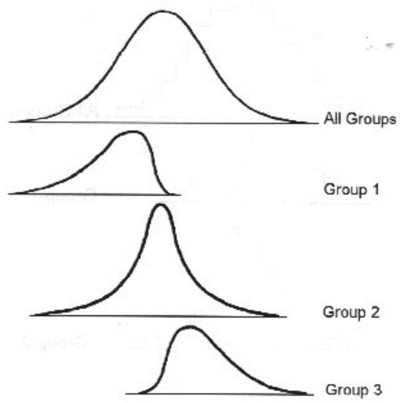


Figure 1.2: Two-stage testing: latent densities of three second-stage groups

In two-stage testing applications, where the groups represent an arbitrary partition of the original sample, assigning the overall mean and standard deviation is more reasonable. In vertical equating and DRIFT analysis, on the other hand, the groups correspond to distinct populations, so the best solution would be to choose a reference group, perhaps the youngest-age group or the first-year group, and assign the mean and standard deviation arbitrarily in that group. Comparing the estimated means and standard deviations of the remaining groups with the reference group would then show the trends in the mean and variability of test performance in successive age groups or year groups.

1.2.2 Equivalent groups equating

Equivalent groups equating refers to the equating of parallel test forms by assigning them randomly to examinees drawn from the same population. In educational applications, this type of assignment is easily accomplished by packaging the forms in rotation and distributing them across whatever seating arrangement exists in the classroom. Provided there are fewer forms than students per classroom, it is justifiable to assume that the abilities of the examinees who receive the various forms are similarly distributed in the population. This is the assumption on which the classical equi-percentile method of equating is based, and it applies also to IRT equating.

Indeed, the procedure is even simpler in IRT because the latent distribution of ability is invariant with respect to the distribution of item difficulties in the forms (this is not true of the number-right score of classical test theory: the test score distribution in the population of respondents is an artifact of the distribution of item difficulties (see Lord & Novick, 1968, pp. 387-392). The IRT scale scores computed from the various forms are therefore equated whenever their location and scale are set in the same way for all forms. There is no necessity for common items between forms, any more than there is for equi-percentile equating, but neither will they interfere with the equivalent groups equating if present.

The method of carrying out equivalent groups equating is somewhat different, however, according to whether common items between forms are or are not present. In both cases, the collection of forms may be treated as if it were one test with length equal to the number of distinct items over all forms. The data records are then subjected to a single-group IRT analysis and scoring.

When common items are *not* present, each form may also be analyzed as an independent test, with the mean and standard deviation of the scale scores of all forms set to the same values during the scoring phase.

Equivalent groups equating is especially well suited to matrix-sample educational assessment, where the multiple test forms are created by random assignment of items to forms within each of the content and process categories of the assessment design, and the forms are distributed in rotation in classrooms. Often as many as 30 forms are produced in this way in order to assure high levels of generalizability of the aggregate scores for schools or other large groups of students.

1.2.3 Nonequivalent groups equating

Nonequivalent groups equating is possible only by IRT procedures and has no counterpart in classical test theory. It makes stronger assumptions than equivalent groups equating, but it remains attractive because of the economy it brings to the updating of test forms in long-term testing programs. Either to satisfy item disclosure regulations or to protect the test from compromise, testing programs must regularly retire and replace some or all of the items with others from the same content and process domains. They then face the problem of equating the reporting scales of the new and old forms so that the scores remain comparable.

Although equivalent groups equating will accomplish this, it requires a separate study in which the new and old forms are administered randomly to examinees from the same population. A more economical approach is to provide for a subset of items that are common to the old and new forms, and to employ nonequivalent groups equating to place their scores on the same scale. These common or "link" items are chosen from the old form on the basis of item analysis results. Link items should have relatively high discriminating power, middle range difficulty, and should be free of any appreciable DIF effect. With suitable common items included, the old and new forms can be equated in data from the operational administration of the tests without an additional equating study. Only the BILOG-MG program can perform this type of equating.

Although the case records from the current administration of the new form and the earlier administration of the old form are subjected to a single IRT item analysis in nonequivalent equating, the test form is identified on each case record and separate latent distributions are estimated for examinees taking different forms. For typical applications of the procedure to unrestricted samples of examinees, the latent distributions may reasonably be considered normal. In that case, the estimation of the mean and standard deviation of each distribution jointly with the item parameters allows for the nonequivalence of the two equating groups. The common items provide the link between the two samples of data so that we may fix the arbitrary origin and unit of a single reporting scale. Simulation studies have shown that if the sample sizes for the two groups are large enough to ensure highly precise estimation of the item parameters, as few as four anchor items can accurately equate the reporting scales for the test forms (see Lord, 1980).

In the BILOG-MG procedure, this method of equating can be extended to nonequivalent groups equating of any number of such forms, provided there are common items linking the forms together in an unbroken chain. An example of a plan for common item linking of a series of test forms is shown in Figure 1.3.

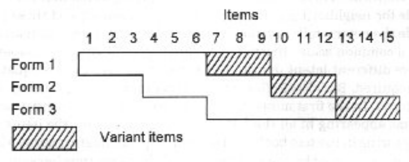


Figure 1.3: An item linking design for test forms updating

1.2.4 Variant items

If total disclosure of the item content of an educational test is required, a slightly different strategy is followed. Special items, called "variant" items, are included in each test form but not used in scoring the form in the current year. It is not necessary that all test booklets contain the same variant items; subsets of variant items may be assigned in a linked design to different test booklets in order to evaluate a large number of them without unduly increasing the length of a given test booklet. These variant items provide the common items that appear among the operational items of the new form, which itself includes other variant items in anticipation of equating to a later form. The item calibration of the old and new form then includes, in total, the response data in the case records for the operational items of the old form, for the linking variant items that appeared on the old form, and for all operational items from the new form. In this way, all of the items in the current test form can be released as soon as testing is complete.

1.2.5 Vertical equating

Vertical equating refers to the creation of a single reporting scale extending over a number of school grades or age groups. Because the general level of difficulty of finding items in tests intended for such groups must increase with the grade or age, the forms cannot be identical. There is little difficulty in finding items that are suitable for neighboring grades or age groups, however, and these provide the common items that can be used to link the forms together on a common scale. Inasmuch as these types of groups necessarily have different latent distributions, non-equivalent groups equating is required. BILOG-MG offers two methods for inputting the response records. In the first method, each case record spans the entire set of items appearing in all the forms, but the columns for the items not appearing in the test booklet of a given respondent are ignored when the data are read by the program. All of the items thus have unique locations in the INPUTrecords and are selected from each record according to the group code on the record. In the second method, the location of the items in the INPUTrecords is not unique. An item in one form may occupy the same column as a different item in another form. In this case, the items are selected from the record according to the form and the group codes on the record. These methods of inputting the response records apply in all applications of BILOG-MG.

1.2.6 Differential item functioning (DIF)

The purpose of differential item functioning analysis is to detect and estimate interactions between item difficulties and various subgroups within the population of respondents (see Thissen, Steinberg, & Wainer, 1993). It is most often applied to interactions with respect to demographic or ethnic groups and to gender, but any classification of the respondents could be investigated in a similar manner. Specifically, it is the interactions of the item location parameters, b_j, reflecting the item difficulties, that are in question. DIF includes only the relative differences in difficulties between the groups. Any reduction of the item percent corrects due to the average level of ability in the group, as indicated by the mean of the corresponding latent distribution, we attribute to the "adverse impact" of the test and do not regard it as DIF. Moreover, we assume that the differential item functioning does not extend to the item discriminating powers. The b_j parameters for the separate groups are estimated on the assumption that the slope parameters, a_j, are homogeneous across groups. (for an alternative form of DIF analysis that includes differential item discriminating power, see Bock, 1993).

DIF analysis is similar to nonequivalent groups equating in the sense that different latent distributions are assumed for the groups in question, but it differs because the same form of the test is administered in all of the groups. It also provides a large sample standard error estimate of the effect estimators. In addition, the program provides an overall marginal likelihood ratio test of the presence of differential item functioning in the data. To perform this test, first analyze the data in a single group as if they came from the same population and note the marginal maximum log likelihood of the item parameters in the final iteration (labeled –2 LOG LIKELIHOOD in the output). Then, analyze the data in separate groups using the DIF model and again note the final log likelihood. Under the null hypothesis of no DIF effects on item locations, the difference in these log likelihoods is distributed in large samples as χ^2 with $(n-1)(m-1)$ degrees of freedom, where n is the number of items and m is the number of groups. When this χ^2 is significant, there is evidence that differential item effects are present. Their interpretation usually becomes clear when the item content is examined in relation to the direction of the estimated contrasts in the b_j parameters, because these contrasts are interactions and must sum to zero (some are positive and others negative).

1.2.7 Item parameter drift (DRIFT)

As defined by Bock, Muraki & Pfiffenberger (1988), DRIFT is a form of DIF in which item difficulty interacts with the time of testing. It can be expected to occur in education tests when the same items appear in forms over a number of years and changes in the curriculum or instructional emphasis interact differentially with the item content (see Goldstein, 1983). Bock, Muraki & Pfiffenberger found numerous examples of DRIFT among the items of a form of the College Board's Advanced Placement Test in Physics that had been administered annually over a ten-year period (see Figure 1.4). DRIFT is similar to DIF in admitting only the item interaction: changes in the means of the latent distributions of successive cohorts are attributed to changes in the levels of proficiency of the corresponding population cohorts.

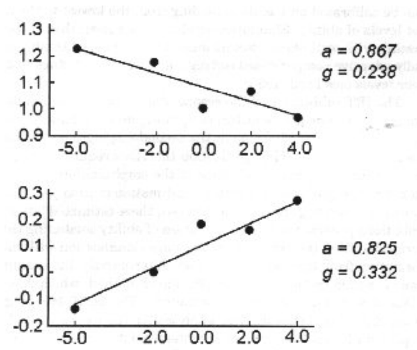


Figure 1.4: Drift of the location parameters of two items from a College Board Advanced Placement Examination in Physics

DRIFT differs from DIF in that the interaction of item location with time is assumed to be a continuous process that can be modeled by a straight or low degree polynomial regression line. Thus, in place of estimating contrasts between groups, we estimate the coefficients of the linear or polynomial function in time that describes the DRIFT in the b_j parameters. The significance of the trends can be judged from the size of the estimated regression coefficient relative to its large sample standard error estimate. The overall presence of DRIFT can be tested in a marginal likelihood ratio test similar to that for DIF.

As implemented in BILOG-MG, DRIFT analysis does not require all items to be included in each test form. The DRIFT regression functions are estimated for whatever time points are available for each item. In most DRIFT applications, it is satisfactory to assume that the latent distributions of the yearly cohorts are normal. The corresponding means and standard deviations estimated in the DRIFT analysis describe differences in the proficiencies of the cohorts.

1.2.8 Two-stage testing

Two-stage testing is a type of adaptive item presentation suitable for group administration. By tailoring the difficulties of the test forms to the abilities of selected groups of examinees, it permits a reduction in test length by a factor of a third or a half without loss of measurement precision. The procedure employs some preliminary estimate of the examinees' abilities, possibly from a short first-stage test or other evidence of achievement, to classify the examinees into three or four levels of ability. Second-stage test forms in which the item difficulties are optimally chosen are administered to each level. Forms at adjacent levels are linked by common items so that they can be calibrated on a scale extending from the lowest to the highest levels of ability. Simulation studies have shown that two-stage testing with well placed second-stage tests is nearly as efficient as fully adaptive computerized testing when the second-stage test has four levels (see Lord, 1980).

The IRT calibration of the second-stage forms is essentially the same as the nonequivalent forms equating described above, except that the latent distributions in the second-stage groups cannot be considered normal. This application therefore requires estimation of the location, spread, and shape of the empirical latent distribution for each group jointly with the estimation of item parameters. During the scoring phase of the analysis, these estimated latent distributions provide for Bayes estimation of ability combining the information from the examinee's first-stage classification with the information from the second-stage test. Alternatively, the examinees can be scored by the maximum likelihood method, which does not make use of the first-stage information. The BILOG-MG program is capable of performing these analyses for the test as a whole, or separately for each second-stage subtest and its corresponding first-stage test. For an example of an application of two-stage testing in mathematics assessment see Bock & Zimowski (1989).

1.2.9 Estimating latent distributions

An innovative application of the BILOG-MG program is the estimation, from matrix sampled assessment data, of the latent distributions for schools or other groups of students. Certain matrix sampling designs, such as those employed by the National Assessment of Educational Progress, include in each booklet a number of short scales, consisting of eight or nine items, in several subject-matter areas. These scales have too few items to permit reliable estimation of the proficiencies of individual examinees in each subject matter, but they do allow estimation of the latent distribution of each proficiency at the group-level if the number of respondents is sufficiently large. There is a tradeoff between the number of items for each scale in each test booklet and the number of respondents: the more items, the fewer respondents are needed for accurate estimation of the group latent distribution.

If each booklet contains perhaps 48 items, the latent distributions for six content areas could be estimated simultaneously. The results of the assessment could then be reported to the public in terms of the means and standard deviations of the achievement levels of the schools or groups. Alternatively, if achievement standards have been set in terms of IRT scale score levels, the percent of students attaining or exceeding each level can be computed from the latent distribution and reported. The latter form of reporting is often more easily understood than scale-dependent statistics such as the mean and standard deviation. Because the BILOG-MG program allows unlimited numbers of groups as well as unlimited numbers of items and respondents, it is well suited to the estimation of latent distributions for this form of reporting. The shape of the latent distributions may either be assumed normal or estimated empirically.

1.2.10 Technical details

The normal ogive model

A response to a binary test item j is indicated in these expressions by the item score,

$x_j = 1$ if the respondent answers correctly, or

$x_j = 0$ if the respondent answers incorrectly.

Let θ denote the ability of the person, and let the probability of a correct response to item j be represented by

$$P(x_j = 1 | \theta) = P_j(\theta);$$

and thus, the probability of an incorrect response is given by

$$P(x_j = 0 | \theta) = 1 - P_j(\theta).$$

In general, the response function also depends upon one or more parameters characteristic of the item, the values of which must be estimated.

The normal ogive model is defined as:

$$P_j(\theta) = \frac{1}{\sqrt{2\pi}} \int_{-(\theta - b_j)/\sigma_j}^{\infty} e^{-[1/2]t^2} dt,$$

where $\sigma_j = 1/a_j$ is called the item dispersion, a_j is the item discriminating power and b_j is an item location parameter. The normal ogive model is conventionally represented as $\Phi_j(\theta)$.

The logistic models for binary scored items

At present, the response models most widely used in applied work are the logistic models for binary scored items. The most important of these models are:

- The one-parameter (1PL, Rasch) model
- The two-parameter (2PL, Birnbaum) model
- The three-parameter (3PL, guessing) model

The one-parameter (1PL, Rasch) model

The one-parameter logistic model is defined as

$$P_{(1)j}(\theta) = \frac{1}{1 + \exp[-a(\theta - b_j)]}$$

where

$\exp(k) = e^k$	and
e = 2.718	is the base of the natural logarithm,
a	is a scale constant determining the units of θ, and
b_j	is a location parameter related to the difficulty of item j (also referred to as the

item “threshold”). Items with larger values of b_j are more difficult; those with smaller values are easier.

The two-parameter (2PL, Birnbaum) model

The two-parameter logistic model is defined as

$$P_{(2)j}(\theta) = \frac{1}{1 + \exp[-a_j(\theta - b_j)]}$$

where a_j is the item discriminating power, and b_j is an item location parameter as in the 1PL model.

The negative of the exponent in this model,

$$z_j = a_j(\theta - b_j),$$

is referred to as a logistic deviate, or logit. The logit can also be written as $z_j = a_j\theta + c_j$ where $c_j = -a_j b_j$. In this form, a_j is referred to as the item slope and c_j as the item intercept (see Figure 1.5).

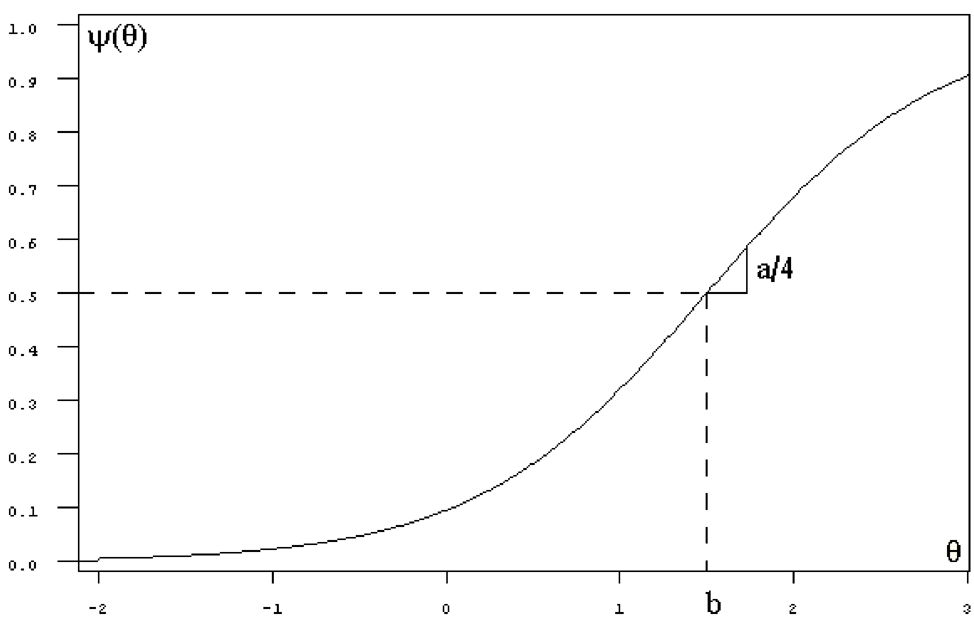


Figure 1.5: the two-parameter logistic model

The 2PL model is conventionally represented as

$$\Psi_j(\theta) = \frac{1}{1+e^{-z_j}}$$

If all a_j are equal, the model is reduced to a one-parameter logistic or Rasch model.

The three-parameter (3PL, guessing) model

In the case of multiple-choice items, an examinee who does not know the correct alternative may succeed in responding correctly by randomly guessing. If the examinee's ability is θ, the probability that the examinee will not know the answer, but guesses correctly (with probability g_j) is $g_j[1-\Psi_j(\theta)]$. The probability that the examinee will respond correctly either by knowledge or by random guessing is therefore

$$\begin{aligned} P_{3j}(\theta) &= g_j[1-\Psi_j(\theta)] + \Psi_j(\theta) \\ &= g_j + (1-g_j)\Psi_j(\theta), \end{aligned}$$

where g_j is the probability of a correct response to a multiple-choice item as a result of guessing. If the correct response alternative is randomly assigned, and all of the examinees guess blindly, the value of g_j is equal to 1/A, where A is the number of alternatives of the multiple-choice item.

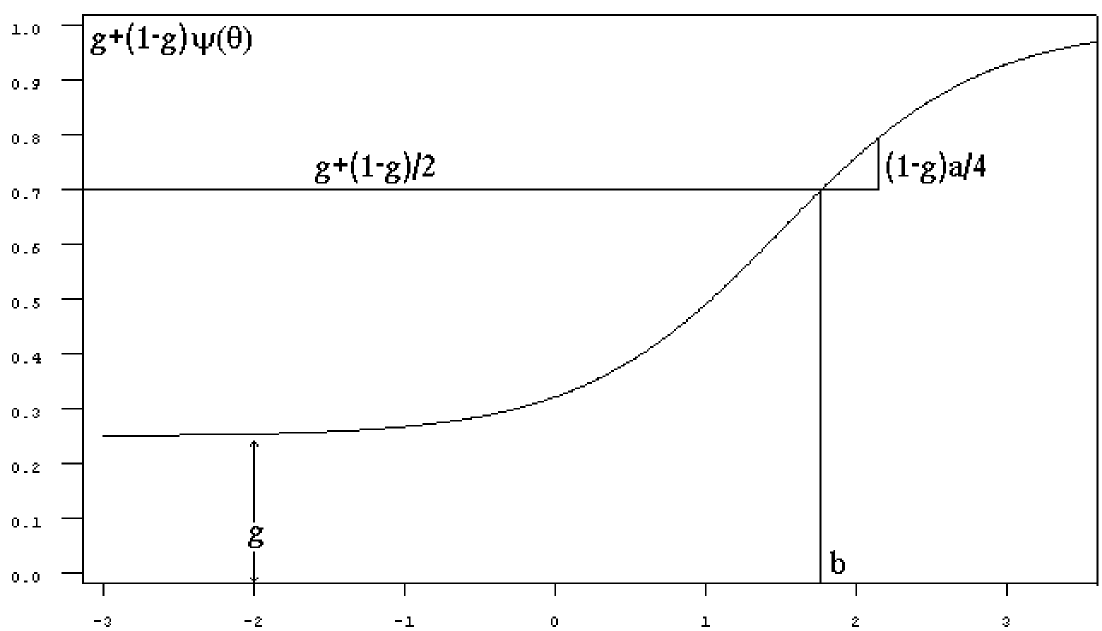


Figure 1.6: Three-parameter logistic model

If some of the examinees guess after eliminating one or more of the alternatives, g_j will be greater than 1/A by some amount that must be determined empirically along with the a_j and b_j or c_j parameters.

The parameter g_j corresponds to the lower asymptote of the item response function, $P_{3j}(\theta)$. This interpretation of g_j, as well as that of the other item parameters, is shown in Figure 1.6.

Relationship between normal ogive and logistic models

The logistic item response models are closely related to the normal ogive model. In order to bring the logistic models into close agreement with the normal ogive model, θ is multiplied by the factor $D = 1.7$. When $D = 1.7$ is used, the discrepancy between the normal response function and its logistic approximation is never greater than 0.01.

When the logit incorporates this factor, as in $z_j = a_j D(\theta - b_j)$, the models are said to be in the normal metric.

Classical item statistics

BILOG-MG computes and uses classical item statistics as starting values for the iterative estimation of the IRT parameters.

On the assumption that θ is distributed with zero mean and unit standard deviation in the population of respondents, the normal ogive item parameters are related to the classical item statistics as follows (see Lord & Novick, 1968, Sections 16.9 and 16.10).

Reliability index (item-trait correlation):

If one assumes a bivariate normal distribution of the population over the item and criterion variables, Richardson (1936) and Tucker (1946) have shown that

$$\rho_j = a_j / \sqrt{1 + a_j^2}, \quad 0 \le \rho_j \le 1$$

where ρ_j is the biserial correlation between ability and item *j*. In classical item analysis, ρ_j is estimated by the item-test correlation (the correlation between response to the item scored 1 or 0 and number-right score for the test).

We see from the equation above that an item with slope 1 (in the normal metric) has a reliability index equal to $1/\sqrt{2} = 0.707$. Items with slopes greater than 1 are more reliable (more discriminating measures of the trait represented by the test); those with slopes less than 1 but greater than zero are less reliable. Items with a negative slope are keyed in a direction opposite to that of the other items. The same relationships hold with good approximation for the logistic parameters expressed in the normal metric.

Item facility (p-value):

Tucker (1946) has expressed classical item difficulty P_j as a function of the item parameters a_j and b_j:

$$P_j = \Phi(-a_j b_j / \sqrt{1+a_j^2}),$$

that is, p_j is the value of the standard normal distribution function at the point

$$\frac{-a_j b_j}{\sqrt{1+a_j^2}} = -b_j \rho_j$$

i.e., the area to the left of the point under the normal curve.

From the equations above it follows that

$$\hat{a}_j = \frac{\hat{\rho}_j}{\sqrt{1-\hat{\rho}_j^2}}$$

and

$$\hat{b}_j = \frac{\hat{z}_j}{\hat{\rho}_j}$$

where $\hat{z}_j$ is calculated using the inverse normal distribution with $\hat{\rho}_j = P(z \le \hat{z}_j)$.

1.2.11 Statistical tests

Because BILOG-MG employs maximum likelihood estimation when fitting the IRT model, large-sample statistical tests of alternative models are available, provided one model is nested within each other. Two models are called "nested" when the larger model is formed from the smaller by the addition of terms and parameters. For example, the one-parameter logistic model is nested within the two-parameter model, which is in turn nested within the three-parameter model. Similarly, the single-group model is nested within the two-group model, and so on. The smaller of the nested models is referred to as the "null" model and the larger as the "alternative". The statistical test of the alternative model vs. the null model is equivalent to a test of the hypothesis that the additional parameters in the alternative are all zero and that no significant improvement in fit is obtained by including them.

At the end of the estimation cycles in the calibration phase, BILOG-MG prints the negative of the maximum marginal log likelihood. If the program is run, with the same data, once with the null model and once with the alternative model, the negative of the log likelihood of the former will always be larger than that of the latter. In large-samples, the positive difference of these log likelihoods is distributed as χ^2 on the null hypothesis. Its number of degrees of freedom is equal to the difference in the number of parameters in the null and alternative models. A model with more parameters should be adopted only when this test statistic is clearly significant. Otherwise, fitting of the additional parameters will needlessly reduce precision of estimation.

BILOG-MG also provides a large-sample test of the goodness-of-fit of individual test items in the analysis: this requires the test to have 20 or more items.

If the test is sufficiently long (more than 20 items), the respondents in a sample of size *N* can be assigned with good accuracy to intervals on the θ-continuum on the basis of their estimated value of θ (for this purpose, we use the EAP estimate with whatever prior is assumed for item calibration; see the section on test and item information to follow). Then the number of those in each interval who respond correctly to item *j* can be tallied from their item scores.

Finally, a likelihood ratio χ^2 test statistic can be used to compare the resulting frequencies of correct and incorrect responses in the intervals with those expected from the fitted model:

$$X_j^2 = 2\sum_{h=1}^{n_g}\left[r_{hj}\log_e \frac{r_{hj}}{N_h P_j(\bar{\theta}_h))} + (N_h - r_{hj})\log_e \frac{N_h - r_{hj}}{N_h(1-P_j(\bar{\theta}_h))}\right],$$

where n_g is the number of intervals, r_{hj} is the observed frequency of correct responses to item *j* in interval *h*, N_h is the number of respondents assigned to that interval, and $P_j(\bar{\theta}_h)$ is the value of the fitted response function for item *j* at $\bar{\theta}_h$, the average ability of respondents in interval *h*.

Because neither the MML or MAP methods of fitting the response functions actually minimize this χ^2, the residuals are not under linear constraints and there is no loss of degrees of freedom due to the fitting of the item parameters. The number of degrees of freedom is therefore equal to the number of intervals remaining after neighboring intervals are collapsed if necessary to avoid expected values less than 2.

1.3 Phases of the analysis: input, calibration and scoring

Phase 1: INPUT

The INPUT routine reads formatted data records. Data for each observation consist of subject identification, optional form number, optional group number, optional case weight, and item response data. Item responses of individual examinees comprise one character for each of *n* items. The answer key, not-presented, and omit codes are read in exactly the same format as the observations. For aggregate-level data, the "responses" consist of number of attempts and number correct for each item. If data are for the aggregate-level model, vectors of numbers of attempts and correct responses to the items are read in decimal format.

Omits and attempts

Omits may be scored "wrong", treated as fractionally correct, or omitted from calculations.

Items and forms

The INPUT routine accepts a list of numbers and corresponding names for all items to be read from the data records. The order in which the items appear in the data records is specified in a form key(s). If the data are collected with a multiple-form test, the program accepts a key for each form. Each respondent's data record is identified by its form number.

Multiple groups

When multiple-group IRT analysis is requested, the INPUT routine accepts a list of item numbers or names identifying the items administered to each group. Each respondent's data record is identified by its group number. The Phase 1 program computes classical item statistics separately for each group.

Subtests

The INPUT routine also accepts lists of item numbers or names, not necessarily mutually exclusive, describing *i* subtests. It scores each subtest and creates a file containing the item scores, item attempts, subtest scores, and other INPUT information for each respondent. Each subtest is calibrated separately in Phase 2. Each respondent is scored on all subtests in Phase 3.

Case weights

If there are case weights for respondents (because they were drawn in an allocation sample), the item responses and item attempts are multiplied by the weight. If the data consist of response patterns, the case weights are the frequencies of the patterns.

Samples

If there are a large number of respondents or aggregate-level records, the INPUT routine can be instructed to select a random sample of a specified size to be passed to CALIBRATE(Phase 2). The complete master file of cases will nevertheless be passed to Phase 3 for scoring.

Classical item statistics

While preparing the item-score file, the INPUT routine also accumulates, subtest by subtest, certain item and test statistics (accumulated from the sample file when the number of cases exceeds the user-specified sampling level). These statistics consist of

- item facilities (percent correct),
- item-subscore correlations, and
- the number of respondents attempting each item.

These quantities are listed and passed to the Phase 2 and Phase 3 routines to provide starting values for item parameter and respondent scale-score estimation.

Phase 2: CALIBRATE

The CALIBRATE routine fits a logistic item-response function to each item of each subscale. There are many options available to the user in this section of the program.

Item-response model

The response model may be the 1-, 2- or 3-parameter logistic response function. The scaling factor $D = 1.7$, employed to scale estimates in the normal metric, may be included or omitted at the user's option. Information that assists the user in model selection is provided in the marginal log likelihood and goodness of fit indices and statistics for individual items. The user may request plots of the observed and expected item-response curves.

Individual data or aggregate data

Item parameters may be estimated from either binary (right/wrong/omit) data or aggregate-level frequency data (number of correct responses, number of attempts) input from Phase 1. If aggregate-level data are used, it is assumed that each respondent in each group responds to only one item per subscale, as required in matrix-sampling applications (see Mislevy, 1983). The aggregate-level option can also be applied to individual data if weights are used and the binary responses take on fractional values. In this use of the aggregate-level option, each respondent responds to more than one item.

Marginal maximum likelihood (MML) estimation of item parameters

Estimation of item parameters by the method of marginal maximum likelihood is applicable to tests of three or more items. The solution assumes the respondents are drawn randomly from a population or populations of abilities, which may be assumed to have either a normal distribution, an arbitrary distribution specified by the user, or an arbitrary distribution to be estimated jointly with item parameters. The empirical distributions of ability are represented as discrete distributions on a finite number of points (histogram). In the case of multiple groups, the CALIBRATE routine also provides estimates of the means and standard deviations of the posterior distributions of ability.

The MML solution employs two methods of solving the marginal likelihood equations: the so-called EM method and Newton-Gauss (Fisher scoring) iterations. The default number of cycles for the EM algorithm is 10; the default for Newton steps is 2. Convergence in the EM steps is hastened by the accelerator described in Ramsay, 1975. Results of each cycle are displayed so that the extent of convergence can be judged. 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.

Item statistics supplied by CALIBRATE

Phase 2 provides the item parameters in the form of the lower asymptote and the item intercept (equal to minus the product of the slope and threshold) and so-called "slope" or "discrimination" parameter, and the item threshold (location) and loading (one-factor item factor loading = $Slope/\sqrt{1+Slope^2}$).

In the one-parameter solution, all slopes are equal. In both the one- and two-parameter solutions, all lower asymptotes are zero. In the three-parameter solution with a common lower asymptote, all lower asymptote parameters are equal. Otherwise, they are estimated separately for each item.

When an analysis of differential item functioning (DIF) is requested, the program provides estimates of the unadjusted and adjusted threshold parameters for each group along with their standard errors. Estimates of group differences in the adjusted threshold parameters are also provided. When an item parameter drift (DRIFT) analysis is selected, the program provides estimates of the coefficients of the linear or polynomial function.

In Phase 2, when there is a single group, the unit and origin of the scale on which the parameters are expressed are based on the assumption that the latent ability distribution has zero mean and unit variance. This is referred to as the "0, 1" metric. When there are multiple groups, the program provides the option of setting the mean and standard deviation of the combined estimated distributions of the group to zero and one.

The parameter estimates in Phase 3 can be rescaled according to scale conventions selected by the user. If the one-parameter model has been selected, the item slope estimates are uniformly 1.0. In other cases, the scores can be scaled to a specified mean and standard deviation in the sample. In both Phase 2 and Phase 3, the item parameter estimates can be saved before and after rescaling, respectively, in formatted external files.

Maximum marginal a posteriori estimation of item parameters

When some items are extremely easy or extremely difficult, there may be insufficient information in the sample to estimate their parameters accurately. This will be especially true if the number of respondents is only moderate (250 or fewer). As an alternative to deleting these items, prior distributions can be placed on the item parameters. The user may specify normal priors for item thresholds, log-normal priors for slopes, and beta priors for lower asymptotes. Each item may have a different specification for its prior.

Default specifications are for prior distributions on slopes under the two-parameter models, and on slopes and lower asymptotes under the three-parameter model. By specifying tight priors on selected item parameters, the user may hold these values essentially fixed while estimating other item parameters. This feature is useful in linking studies, where new test items are to be calibrated into an existing scale without changing parameter values for old items.

Item fit statistics

Approximate χ^2 indices of fit are computed for each item following the final estimation cycle. For the purpose of computing these χ^2, the scale score continuum is divided into a number of successive intervals convenient for displaying the response proportions (maximum of 20). Each respondent is assigned to the interval that includes the EAP estimate (based on the type of prior specified by the user) of his or her score. For the item in question, the expected response probabilities corresponding to the average EAP estimate of ability of cases that fall in the interval are used as the expected proportion for the interval.

A likelihood ratio χ^2 is then computed after combining extreme intervals so that the expected frequency exceeds five. Degrees of freedom are equal to the number of combined intervals. There is no reduction in degrees of freedom due to estimating the item parameters because the marginal maximum likelihood method does not place linear constraints on the residuals.

At the user's request, observed and expected item-response curves are plotted for each item.

Test of improved fit if the number of parameters is increased

When the expected frequencies of the individual response patterns are too small to justify the likelihood ratio test of goodness-of-fit, the change in likelihood ratio χ^2 between the 1- and 2-parameter models, or between the 2- and 3-parameter models, is a valid large-sample test of the hypothesis that the added parameters are null. The degrees of freedom of each of these change χ^2 are equal to the number of items.

Test of overall fit when the number of items is 10 or less

If the sample size is large and the number of items is small, the overall fit of the response functions of all items can be tested by comparing the observed frequencies of the patterns with the expected marginal frequencies computed from the fitted functions. The data must be in the form of response patterns and frequencies. The likelihood ratio χ^2 statistic for the test of fit is

$$G^2 = 2\sum_{i=1}^{2^n} r_i \log_e \frac{r_i}{N\bar{P}_i}$$

where 2^n is the number of possible patterns of the *n* binary item scores, r_i is the observed frequency of pattern *i*, *N* is the number of respondents, and P_i is the estimated marginal probability of pattern *i*.

The number of degrees of freedom is $2^n - kn - 1$, where k is the number of parameters in the response model.

This test should be used only when the number of respondents is large relative to the number of patterns. If a few patterns have zero observed frequency, ½ should be substituted as the frequency for those patterns and corresponding ½s subtracted from the frequency of the most frequent pattern (or 1 could be used for this purpose).

Phase 3: SCORE

The SCORE routine makes use of the master response file from Phase 1 and the item parameter estimate files from Phase 2 to compute estimated scale scores for respondents. The user may select one of the three methods described below for estimating scale scores.

In each of these methods the user has the option of bi-weight robustification to protect the estimates from spurious responses due to guessing or inattention. Because effects of guessing are suppressed by the robustification, the lower asymptote is not incorporated in the response model in Phase 3 when the bi-weight option is selected. Scores and standard errors for all subscales are calculated simultaneously for each respondent. Results may be printed and/or saved on an external file.

Maximum likelihood (ML)

ML estimates with or without robustification are computed by the Newton-Raphson method starting from a linear transformation of the logit of the percent-correct score for the subject. In those rare cases where the Newton iterations diverge, an interval-bisection method is substituted.

Estimates for respondents with all correct or all incorrect responses are attributed by the half-item rule. That is, respondents who score all incorrect are assigned one-half a correct response to the easiest item; respondents who score all correct are assigned one-half a correct response to the hardest item. The estimate is then computed from this modified response pattern.

Standard errors are computed as the square root of the negative reciprocal of the expected second derivative of the log likelihood at the estimate, *i.e.*, the square root of the reciprocal Fisher information.

Bayes or expected a posteriori (EAP)

EAP estimates with or without robustification are computed by quadrature using a discrete distribution on a finite number of points as the prior. The user may select the number of points and has the choice of a normal, locally uniform, or empirical prior. For the latter, the user may supply the values of the points and the corresponding empirical weights or may use the empirical weights generated in Phase 2.

The EAP estimate is the mean of the posterior distribution and the standard error is the standard deviation of the posterior distribution.

Bayes modal or maximum a posteriori (MAP)

MAP estimates with or without robustification are also computed by the Newton-Gauss method. This procedure always converges and gives estimates for all possible response patterns. A normal prior distribution with user-specified mean and variance is assumed [the default is N(0, 1)]. The estimate corresponds to the maximum of the posterior density function (mode); the standard error is the square root of the negative reciprocal of the curvature of the density function at the mode.

Estimated latent distribution

When EAP estimation is selected, the SCORE routine obtains an estimate of the population distribution of ability in the form of a discrete distribution on a finite number of points. This distribution 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. The program also computes the mean and standard deviation for the estimated latent distribution. Sheppard's correction for coarse grouping is used in calculating the standard deviation.

Rescaling

The ability estimates are calculated initially in the scale of the item parameter estimates from Phase 2. In addition, however, rescaled estimates may be obtained by one of the following options:

- the mean and standard deviation of the sample distribution of score estimates are set to arbitrary values specified by the user (default = 0, 1);
- a linear transformation of scale is provided by the user;
- if EAP estimation has been selected, the mean and standard deviation of the latent score distribution may be set to arbitrary values by the user (default = 0, 1). Any of these options may be applied to all subtests in the same computer run, or different rescaling parameters may be used for each subtest. Parameter estimates and standard errors for items from Phase 2 are rescaled for each subtest according to the selected option.

Marginal probabilities of response patterns

When EAP estimation is selected, the marginal probability of each response pattern in the sample is calculated and printed along with the corresponding number-right score and scale score.

Item and test information tables and curves

BILOG-MG provides at the user's request a number of indices and plots concerning item and test information:

- Plots of test information and standard error curves for each subtest.
- Tables of item information indices, including the point and value of maximum information.

Classical reliability

The classical definition of reliability is simply the ratio of the true score variance to the observed score variance, which is the sum of the true scores variance and the error variance. In an IRT context, the true scores are the unobservable theta values that are estimated with a specified standard error from item response patterns, as for example in Phase 3 of the BILOG-MG program.

Classical reliability is implemented in BILOG-MG in two different ways according to how the true score and error variances are estimated. To distinguish the two results, we refer to one as “theoretical” reliability and the other as “empirical” reliability. The result for theoretical reliability appears in connection with the test information plots in the Phase 3 output; the result for “empirical” reliability appears following the display of the means, standard deviations, and average standard error of the scores earlier in the Phase 3 output. The computation of these two quantities is carried out as follows.

Theoretical reliability

The theoretical reliability value applies to IRT scores estimated by the maximum likelihood method (METHOD=1 of the SCORE command). It is based only on the item parameters passed from Phase 2 and does not depend in any way on the ability scores computed in Phase 3. Instead, it assumes that the true ability scores are distributed normally with mean zero and variance one in the population of examinees. The test information function is integrated numerically with respect to this assumed distribution to obtain the average information expected when the test is administered in the population. The formulas for evaluating test information for any given value of ability, assuming a one, two, or three parameter logistic item response model, are as follows:

1PL:

$$S.E._{(1)}(\hat{\theta}) = \{1/D^2 a^2 \sum_{j=1}^{n} P_{(1)j}(\hat{\theta})[1 - P_{(1)j}(\hat{\theta})]\}^{\frac{1}{2}}$$

2PL

$$S.E._{(2)}(\hat{\theta}) = \{1/D^2 \sum_{j=1}^{n} a_j^2 P_{(2)j}(\hat{\theta})[1 - P_{(2)j}(\hat{\theta})]\}^{\frac{1}{2}}$$

3PL

$$S.E._{(3)}(\hat{\theta}) = \left\{ 1/D^2 \sum_{j=1}^{n} a_j^2 \frac{1-P_{(3)j}(\hat{\theta})}{P_{(3)j}(\hat{\theta})} \left(\frac{P_{(3)j}(\hat{\theta}) - g_j}{1-g_j} \right)^2 \right\}^{\frac{1}{2}}$$

Although the formulas are expressed in terms of standard errors, the information values can be obtained by taking the reciprocal of the squared standard error. Conversely, the reciprocal of the average information with respect to the ability distribution is the harmonic mean of the error variance. Since by assumption the variance of the true score (*i.e.*, ability) distribution is equal to one when expressed in the scale of the Phase 2 item parameter calibration, the theoretical reliability is one divided by the quantity one plus the error variance.

In the program the theoretical reliability is computed for each form of the test when there are multiple forms. Whether the analysis pertains to one group or multiple groups of examinees is not relevant; because the theoretical reliability is a function only of the item parameters, the presence of multiple groups has no effect on the results.

This version of BILOG-MG has provisions for computing information curves and reliability for any set of item parameters supplied in Phase 1 as starting values for item parameter estimation. If alternative forms are to be constructed from the items set, the user can insert forms commands following the score command to indicate the item composition of the forms. See the documentation of these score-forms commands for instructions on how to set up these calculations (REFERENCE, READF and NFORMS on the SCORE command) discussed in Section 3.2.14).

Empirical reliability

The formulas for estimating the error and true score variances for calculating empirical reliability differ depending on how the ability scores of the examinees in the sample (or in the samples in the case of a multiple-group analysis) are estimated:

- For maximum likelihood scores (method 1), the estimated error variance is the reciprocal of the mean of the best information evaluated at the ability estimates of all examinees in the sample or samples. The score variance is just the variance of the maximum likelihood scores in the sample or samples. The true score variance can therefore be estimated simply by subtracting the error variance from the score variance. The empirical reliability in each sample is then given by that value for the true score variance divided by the score variance.
- For Bayes EAP scores (METHOD=2 on the SCORE command), the estimate of the error variance is the mean of the variances of the posterior distribution's ability for all examinees in the sample or samples. Because the ability scores are regressed estimates in Bayes estimation, the true score variance is estimated directly by the variance of the means of the posterior distribution (*i.e.*, the EAP scores) in the sample or samples. An empirical reliability is therefore the true score variance divided by the sum of the true score variance and

the error variance. The formulas for computing, by numerical integration, the means and variances of the examinee posterior distributions of ability are as follows.

The Bayes estimate is the mean of the posterior distribution of θ, given the observed response pattern x_i (Bock & Mislevy, 1982). It can be approximated as accurately as required by the Gaussian quadrature,

$$\bar{\theta}_i \cong \frac{\sum_{k=1}^{q} X_k P(\mathrm{x}_i \mid X_k) A(X_k)}{\sum_{k=1}^{q} P(\mathrm{x}_i \mid X_k) A(X_k)}$$

This function of the response pattern x_i has also been called the *expected a posteriori* (EAP) estimator. A measure of its precision is the posterior standard deviation (PSD), approximated by

$$PSD(\bar{\theta}_i) \cong \frac{\sum_{k=1}^{q} (X_k - \bar{\theta}_i)^2 P(\mathbf{x}_i \mid X_k) A(X_k)}{\sum_{k=1}^{q} P(\mathbf{x}_i \mid X_k) A(X_k)}.$$

The weights, $A(X_k)$, in these formulas depend on the assumed distribution of θ. Theoretical weights, empirical weights $A^*(X_k)$, or subjective weights are possibilities.

The EAP estimator exists for any answer pattern and has a smaller average error in the population than any other estimator, including the ML estimator. It is in general biased toward the population mean, but the bias is small within $\pm 3\sigma$ of the mean when the PSD is small (*e.g.*, less than 0.2σ, see Bock & Mislevy, 1982).

Although the sample mean of the EAP estimates is an unbiased estimator of the mean of the latent population, the sample standard deviation is in general smaller than that of the latent population. In most applications, this effect is not apparent because the sample standard deviation is adjusted arbitrarily when the scale scores are standardized. Thus, the bias is not a serious problem if all the respondents are compared using alternative test forms that have much different PSDs. The same problem occurs, of course, when number-right scores from alternative forms with differing reliabilities are used to compare respondents. Users of tests should avoid making comparisons between respondents who have taken alternative forms that differ appreciably in their reliability or precision. A further implication is that, if EAP estimates are used in computerized adaptive testing, the trials should not terminate after a fixed number of items, but should continue until a prespecified PSD is reached.

For Bayes MAP scores, the estimated error variance is the mean of the reciprocal of the test information at the modes of the posterior distributions of all examinees in the sample or samples. Similarly the true score variance is estimated by the mean of the variances of the posterior distributions at the mode. As in the case of Bayes EAP scores, the empirical reliability for the MAP scores is equal to the true score variance divided by the sum of the true score variance and the error variance. The formulas for computing the posterior mode and test information at the mode are as follows.

Similar to the Bayes estimator but with a somewhat larger average error is the Bayes modal, or so-called *maximum a posteriori* (MAP) estimator. It is the value of θ that maximizes

$$P(\theta \mid x_i) = \sum_{j=1}^{n} \{x_{ij} \log_e P_i(\theta) + (1 - x_{ij}) \log_e [1 - P_i(\theta)]\} + \log_e g(\theta),$$

where $g(\theta)$ is the density function of a continuous population distribution of θ.

The stationary equation is

$$\sum_{j=1}^{n} \frac{x_{ij} - P_j(\theta)}{P_j(\theta)[1 - P_j(\theta)]} \cdot \frac{\partial P_j(\theta)}{\partial \theta} + \frac{\partial \log_e g(\theta)}{\partial \theta} = 0$$

Analogous to the maximum likelihood estimate, the MAP estimate is calculated by Fisher scoring, employing the *posterior information*,

$$J(\theta) = I(\theta) - \partial^2 \log_e g(\theta) / \partial \theta^2,$$

where the right-most term is the second derivative of the population log density of θ.

In the case of the 2PL model and a normal distribution of θ with variance σ^2, the posterior information is

$$J(\theta) = \sum_{j=1}^{n} a_j^2 P_j(\theta)[1 - P_j(\theta)] + \frac{1}{\sigma^2}.$$

The PSD of the MAP estimate, θ, is approximated by

$$PSD(\theta) = \sqrt{1 / J(\hat{\theta})}.$$

Like the EAP estimator, the MAP estimator exists for all response patterns but is generally biased toward the population mean.

Because empirical reliabilities are estimated from the results of test score estimation, they are reported separately for each group of examinees in a multiple-group analysis. Note, however, that the test forms are not distinguished in these computations. If there are multiple forms of the test, the empirical reliabilities are aggregations over the test forms.

Information curves and reliabilities for putative test forms

It may be useful in test development to preview the information and theoretical reliability of test forms that might be constructed from items drawn from a calibrated item bank. This can now be done using the FIX keyword on the TEST commands. Starting values for the item parameters are supplied to the program (see definition of the FIX keyword) or the parameters may be read from an IFNAME file. Then all of the items are designated as fixed using the FIX keyword. If the INFO keyword appears in the SCORE command, the required information and reliability analysis will be performed in Phase 3.

In order for this procedure to work, however, the program must have data to process in Phases 1 and 2 for at least a few cases. Some artificial response data can be used for this purpose. The only calculations that will be performed in Phase 2 are preparations for the information analysis in Phase 3. The number of EM cycles in the CALIB command can therefore be set to 1 and the number of Newton cycles to 0. The NOADJUST option must also be invoked.

Output files

Phase 1 results appear in the ***.ph1** file. They include test and item identification and classical item statistics.

Phase 2 results appear in the ***.ph2** file. They include assumed prior distributions, estimated item parameters, standard errors and goodness-of-fit statistics, DRIFT parameters, estimates of differential item functioning, posterior distributions for the groups, group means, and standard deviations, and estimates of their standard errors.

Phase 3 results appear in the ***.ph3** file. They include assumed prior distributions of the scale scores for MAP and EAP estimation, correlations among the subtest scores, rescaling constants, rescaled item parameters, scale scores for the subjects, test information plots, and parameters of the rescaled latent distribution.

1.4 Data preparation[2]

1.4.1 Data characteristics: What kind of data can I use?

The only type of data that BILOG-MG currently can handle is fixed format with one or more lines per record (case) and one-character response codes. Fixed format means that the variables occupy the same column positions throughout the data file. The only acceptable values in such a data file are the upper- and lowercase characters a through z, the digits 0 through 9, and any of the special characters like +-.*&. Tab characters (^t) and other control characters that are usually embedded in files from word processing (*e.g.*, **doc**), database (*e.g.*, **dbf**), spreadsheet (*e.g.*, **xls**), and statistical applications (*e.g.*, **sav**) are not acceptable and data files with such extraneous char-

[2] This section was contributed by Leo Stam.

acters will produce unexpected program behavior that may be difficult to trace. Section 1.4.5 illustrates the conversion of an Excel[3] file to a fixed format file.

In its simplest form the data file contains *individual response data*. Such a *flat* file usually has one line per record, starting with a subject ID (identification field) and followed by a number of one-character response codes for the items in the test. Spaces in between fields and/or items are permitted, as long as those blanks maintain the column positions of the item responses throughout the file.

Example:

```
John     abbac aaacc
Mary-Ann bcabb bbcaa
```

Mary-Ann selected response category a for items 3, 9, and 10, while John answered b, c, and c, respectively.

The item response codes may represent right/wrong answers, selected response categories, nominal category codes, ordinal variable values, ratings, etc. The maximum number of different codes per item is dependent on the program used for analysis. BILOG-MG analyzes binary (dichotomous) responses only. The data may be multiple-category (1,2,3,4 or a,b,c,d,e, etc.), but the program reduces that to right/wrong data with the correct response code key that the user provides.

Besides a subject ID with up to 30 characters and the single-character item response codes, other fields that may be present in the records are:

- A case weight or a frequency
- A subtest number
- A group identifier
- A form number
- A rater code

The specific requirements for these fields can be found in the *Command Reference* section for the different programs. The group identifier in BILOG-MG has to be a single digit (integer) , starting with 1.

Including the single-subject data described above, the program allow the following data types:

- Single-subject data with or without case weights
- Number tried/number right data with or without case weights
- Response patterns with frequencies

[3] Excel 2000 was used in the examples.

1.4.2 Format statement: How do I tell the program about my data?

This program is command-driven and runs in batch mode. That is to say that the user prepares a command file (either directly in an editor or through a dialog-box user interface, if present) and submits this command file to the program for execution (*Run*).

While it is true that command-driven programs were the standard before the "point-and-click" user interfaces ("GUI") entered the computing scene, maintaining this standard for the current programs was done deliberately. The dialog-box interfaces that have been added are merely a so-called front-end for the convenience of the user in building such a command file. Despite the progress that has been made with the graphical user interfaces, in our experience users who use a program routinely still prefer the ease of use of the command file. Moreover, such a file stores the particulars of an analysis in a very succinct way, such that making small changes to an analysis, retrieving an old analysis, or sharing the analysis with other users of the program (also: technical support) is a straightforward task. It is like giving somebody a MAP of how to get from A to B instead of having to describe the route with "take the first street to the right, then a left at the third traffic light", etc. Granted, learning and remembering the commands, keywords, and options used in a program requires a considerable effort (like learning how to read MAPs), while the click-and-point interface can lay claim to being intuitive for the user. The dialog-box user interface is especially helpful in that learning process or as a means to refresh the memory of the occasional user of the programs.

Besides the particular analysis specifications, the command file informs the program where the data file can be found and how it should be read. The location of the data file to be analyzed is simply a matter of specifying that location with a keyword.

For example:

```
>GLOBAL … DFNAME='F:\BILOGMG\EXAMPLES\EXAMPL06.RAW';
```

Note that the name of the data file must be enclosed in single quotes. The drive and directory path should be included if the data file is not in the same folder as the command file. It is also good practice to keep all the files, including the command file, for a particular analysis together in a separate folder. In that case, all that is needed is the filename.

Now that the program knows where to find the data, it needs to be told how to read those data. What part of a record has the subject ID, in which column is the response code for the first item to be read, where is the group code, if any, etc. To that end, the user includes a format statement in the command file.

Format statements are enclosed in parentheses. They are entered on a separate line in the command file and usually one line is all that is needed. However, if more lines are needed, the user can indicate that with a keyword (*e.g.*, NFMT=2 tells the program that the format statement occupies two lines).

The format statement for the simple example above is: (8A1,1X,5A1,1X,5A1).

Here is the file again, with a column counter added above for convenience:

```
12345678901234567890
John     abbac aaacc
Mary-Ann bcabb bbcaa
```

As can be seen, the total length of each record in the file is 20 columns. The first eight columns contain the ID field. This is specified in the format statement with “8A1.” That stands for “eight alphanumeric characters of length one.” The “A” is a format code and stands for *alphanumeric*. The 1 indicates the width and the 8 is a repeat count. Other possible format codes are “F” (for floating point, used to read real numbers) and “I” (for integer).

The next element in the format statement is an example of an operator, in this case “X”. The “X” is used to tell the program to skip one or more columns. The example specifies “1X” or skip one column. Next follows a block of five item responses to be read as “5A1”. Then, we instruct the program to skip another column and to read a second set of five alphanumeric characters: items 6 through 10. Thus, the complete format statement, (8A1,1X,5A1,1X,5A1), describes how to read each of the twenty columns in a record. Because the format statement describes one data record and that description is applied to the whole data file, all the records in the data file should look identical: the essence of a fixed format.

Instead of the “X” operator, the “T” (tab) operator can be used with the same result. The tab operator specifies the column position to tab to. Thus, the format statement (8A1,1X,5A1,1X,5A1) becomes (8A1,T10,5A1,T16,5A1) when using the tab operator. Tabbing backward is also possible. That is often used when the examinee records have the examinee ID at the end of each line, while the program wants it as the first thing being read. Here is our example in that format. The first line is a column counter added for your convenience. It is not part of the actual data file.

```
12345678901234567890
abbac aaacc John
bcabb bbcaa Mary-Ann
```

With the format statement (T13,8A1,T1,5A1,1X,5A1) we instruct the program to read the eight-character ID starting at column 13, then go back to column 1 and read two blocks of five items, skipping a blank column in the middle. This examples also illustrates that the “X” and “T” operators can be used within the same format statement. Obviously, the “T” operator can also be used to read the items in an order that is different from the order in the data file. For example, with (T13,8A1,T7,5A1,T1,5A1) we read the second block of 5 items before the first block of 5 items.

The final operator that the user should know about is the “/” (slash) operator. It instructs the program to go to the next line of the data file. Oftentimes, users have data where the record for each examinee spans more than one line in the data file. A simple example is as follows (again, with the column counter added for convenience).

```
1234567890123456
John     1 abbac
John     2 aaacc
Mary-Ann 1 bcabb
Mary-Ann 2 bbcaa
```

Here, each block of five items is given on a separate line. This could easily result from two different data files (each with an examinee ID and five items) that were concatenated into one file, then sorted on examinee ID. To keep the order of the item blocks the same for each examinee, a block number was added to the original data files.

The format statement (8A1,T12,5A1,/,11X,5A1) will read the examinee ID from the first line of the record (8A1), tab to column 12 and read the first five items (T12,5A1), then go to the next line of the record (/), skip the first 11 columns and read columns 12—16 as the responses to the second set of five items. Note that the examinee ID in the second line of each record is not needed.

A special use of the forward slash operator is to read every first, second, third, etc. record of a large data file. For example, (8A1,1X,20A1,/) reads every odd record of a data file, starting with the first one, while (/,8A1,1X,20A1) reads every even record of a data file, starting with the second one.

The examples that come with the program use a variety of format statements and it is a good idea to look for an example that resembles your data when in doubt about the right format statement. The chapters in this book that describe the examples also offer further details on the use of the format statement.

1.4.3 Telling right from wrong with the response key

When you are analyzing multiple-choice items that are either answered correctly or incorrectly, the program needs to know the item response code for each item that represents a correct answer. The user provides that information with a response key.

Users of BILOG-MG should specify in the command file where the response key can be found (unless the data are already coded as 1 for a right and 0 for a wrong answer). The response key is a record with the exact same format as the data records. It can be in its own file, or it can be part of the data file. The latter option makes it easier to check that the format is indeed identical.

```
key      acaab baaba
John     abbac aaacc
Mary-Ann bcabb bbcaa
```

The file has the response key as the first record. The word *key* is used in the ID field for convenience. It is not needed and will not be read by the program. BILOG-MG will apply the response key to the data records and it will convert John's responses to 1001001100 and Mary-Ann's responses to 0110110001.

1.4.4 What about missing data?

In educational assessment, the reason for an item response in a data file to be coded as missing is generally limited to two possibilities. The specific item was not presented to the examinee or the examinee did not respond to the specific item. The former occurs when examinees answer different forms (selection of items) of the same test and all the items of the test are included in the data file. The importance of the differentiation in missing codes lies in the fact that omitted items can be treated as a wrong response, a fractionally correct response, or the same as a not-presented item, *i.e.*, excluded from the calculations.

Using the simple example again, the data file with not-presented items could look like:

```
John     abbac xxxxx
Mary-Ann xxxxx bbcaa
```

John took a different form of the test than Mary-Ann. They both responded to the five items in their form and all ten items of the two forms are included in the data file. Although the example uses the same not-presented code for all items, note that with BILOG-MG the not-presented (or omitted item code) may vary among items. BILOG-MG accommodates both omitted and not-presented codes. The format of not-presented and omitted keys is as described in Section 1.4.3. Note that, if more than one key is used as part of the data file, the keys should follow the order as described in the *Command Reference* sections for the respective programs.

1.4.5 Data import: What if my data are different?

BILOG-MG expects plain text (ascii) data files with a fixed format. Because the program does not include an import facility to handle various file formats, the user with data in such a format faces the task of converting the dataset to the plain text, fixed format. Spreadsheet, database, and statistical applications generally offer the user some form of data export (or **Save As**) that includes the plain text format. In this section we will illustrate such a conversion with an Excel dataset as starting point. We selected Excel, because it has a format that other applications include in their export formats, and it is a widely used program. This way, users that are unclear about how to convert a specific data format to plain text format may convert to Excel, then follow one of the two methods described below.

The user is advised always to use copies of the original dataset. With Excel, for example, the **Save As** operation uses a format that can only save the active worksheet, so some of your work may get lost.

1.4.6 Using the print format

Using this simple format can only be done with files up to 240 columns (after conversion). In other words, if your Excel worksheet has more than 240 (minus maximum ID length minus possible form and/or group indicators) items, this method will not work.

In Excel, highlight all the columns with the item response codes and set the column width of the highlighted columns to 1. This assumes that your response codes are already one-character codes. If not, you should use the recode capabilities of Excel. For example, if a twelve-category item is coded as 1 through 12, recode it as 1,2,3,4,5,6,7,8,9,A,B,C or as A through L. The column with the ID field should be set to the maximum length of the values appearing in that column. Form or group indicators are best coded as numbers, starting with one.

Now, save the data file as a "***.prn**" file. Excel calls that a **Formatted text (Space-delimited)** file. If you want your filename to have the extension **dat** (instead of the automatic **prn** extension), use double quotation marks (") around the name of the file you want to save it to. Answer **Yes** to the question about loosing special formatting features.

The resulting file should look as shown below, where the first 8 columns are the ID field, followed by 17 item responses. Note that the leading blanks in the first ID field are automatically included because the column width in Excel was set to 8 and the ID itself has only 4 characters. The alignment of the item responses is preserved.

```
    John01010101010101010
Mary-Ann10101010101010101
....
```

1.4.7 Using the tab-delimited format

Another option in Excel is to **Save As** txt format, which produces a tab delimited file. This method has no limitations on the maximum record length. However, the program stumbles on tab characters (does not know how to handle that) and they have to be removed. You can do that in MS Word, for example, by reading in the file as a plain text file, then do a global replace of the "^t" with either a blank or nothing. Then save the file. This works well if your ID field has the same number of characters. Otherwise, you can move the column to the end of the worksheet before you do the **Save As** operation.

A second problem occurs when your worksheet has cells with no entries at all (missing response). When exporting (**Save As**) this as a tab-delimited text file, a global replacement of the tab character with a blank will throw off the column alignment. In that case, you should replace all instances of tab-tab with tab-space-tab.

1.4.8 Data export: What if my data needs editing?

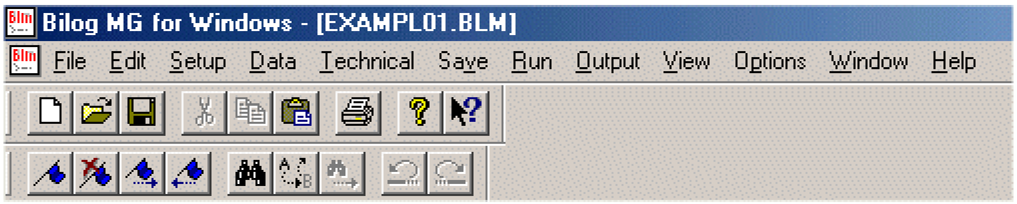
Going the other way, from a plain text, formatted data file to an Excel file has a number of usages. Foremost is data editing. The first attempt at analysis may reveal several difficulties in the data. Values that are out of range, negative item-test correlations, group codes that are coded with characters instead of numbers, etc. Importing the plain text data file into Excel or a similar application provides the user with powerful tools for data editing and data cleaning.

From within Excel, select **Get External Data** from the **Data** menu, then **Import Text File**. Select the data file to import. The **Text Import Wizard** opens with a preview of the data file. Select **Fixed width** as the type that best describes the data, then click the **Next** button. In the **Data**

Preview box, use the mouse to set break lines separating the data into columns. Once satisfied, click **Next**. The last step allows you to skip columns, if needed. Click **Finish**.

2 The BILOG-MG interface

When the BILOG-MG program is opened the first time, a blank window is displayed with only three active options: **File**, **View** and **Help**. By default, however, BILOG-MG will open with the last active file syntax displayed. In this case, or when a command file is opened, the main menu bar shown below is displayed.



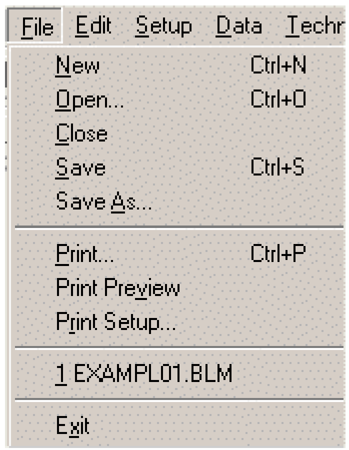
There are 12 menu titles available on the main menu bar. The main purpose of each is summarized in Table 2.1.

Table 2.1: Menu titles on the main menu bar

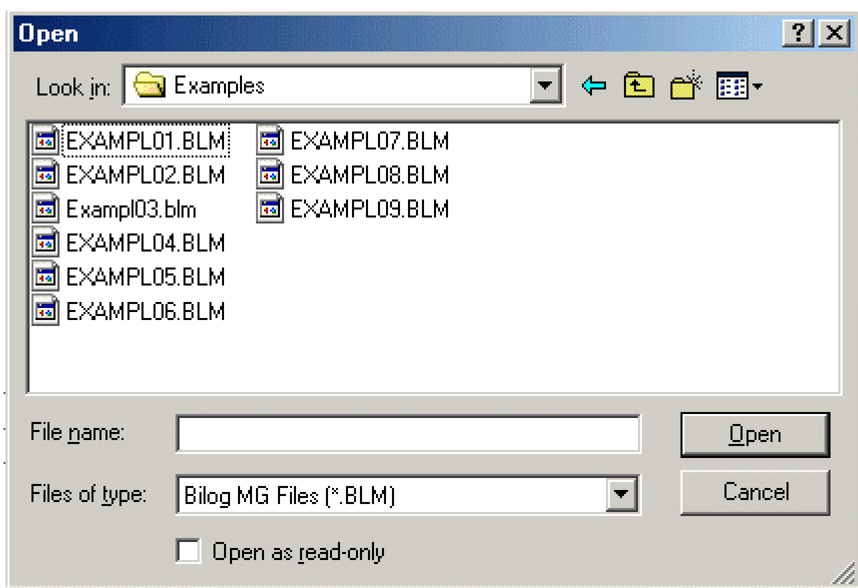
Menu title	Purpose
File	Creating or opening files, printing files and exiting the program
Edit	Standard Windows editing functions
Setup	Model specification
Data	Description of the data, with option to enter new data
Technical	Specifying starting values and priors for calibration and/or scoring
Save	Saving output to external files
Run	Generating syntax and running one or all phases of the program; accessing the graphics procedure.
Output	Viewing output files for the current analysis.
View	Show or hide the tool bar and status bar
Options	Changing program settings and user preferences
Window	Switching between open files
Help	Access to the online help, build number and contact information for SSI

2.1 File menu

The **File** menu provides the user with options to open an existing syntax or text file, to create a new file, to save or to print files.

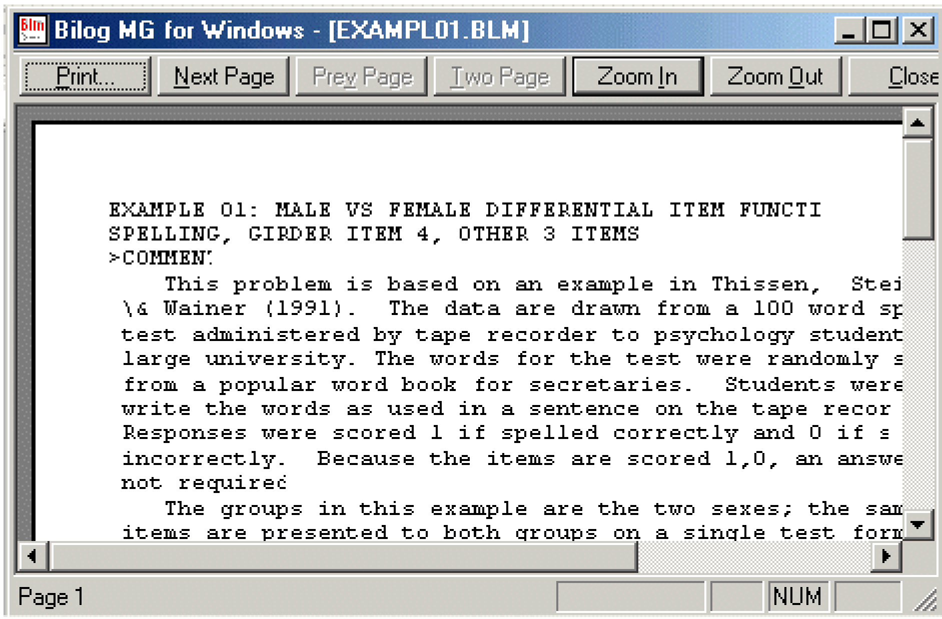


When the **New** or **Open** options are selected from the **File** menu, the user is prompted for the name of a command file. This can be either a new file, in which case a new name is entered in the **File name** field, or an existing file, in which case one can browse and select the previously created command file to be used as the basis for the current analysis.



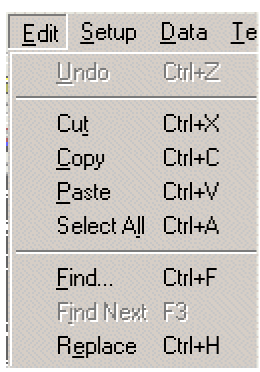
The **Close** option is used to close any file currently open in the main BILOG-MG window, while the **Save** option is used to save any changes made to the file since it was opened. With the **Save As** option a file may be saved under the same or a different name in a folder of the user's choosing.

The **Print** and **Print Setup** options represent the usual Windows printing options, while selection of the **Print Preview** option opens a new window, in which the current file is displayed in print preview mode. Options to move between pages and to zoom in and out are provided. The printing options are followed by the names of the last files opened, providing easy access to recently used files. The **Exit** option is used to exit the program and return to Windows.



2.2 Edit menu

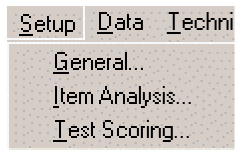
The **Edit** menu has the standard Windows options to select, cut, copy and paste contents of files. In addition, the user can search for text strings and/or replace them with new text using the **Find** and **Replace** options.



2.3 Setup menu

The **Setup** menu is used to provide general information to be used in the analysis. The three options on the **Setup** menu are:

- **General**: used for entering general information on the type of analysis required.
- **Item Analysis**: used to specify the allocation of items to forms, subtests, and/or groups and to control the item parameter estimation procedure.
- **Test Scoring**: used to request the scoring of individual examinees or of response patterns, item and test information and rescaling of scores.



The menu options are used to activate dialog boxes. The function of each dialog box is described below.

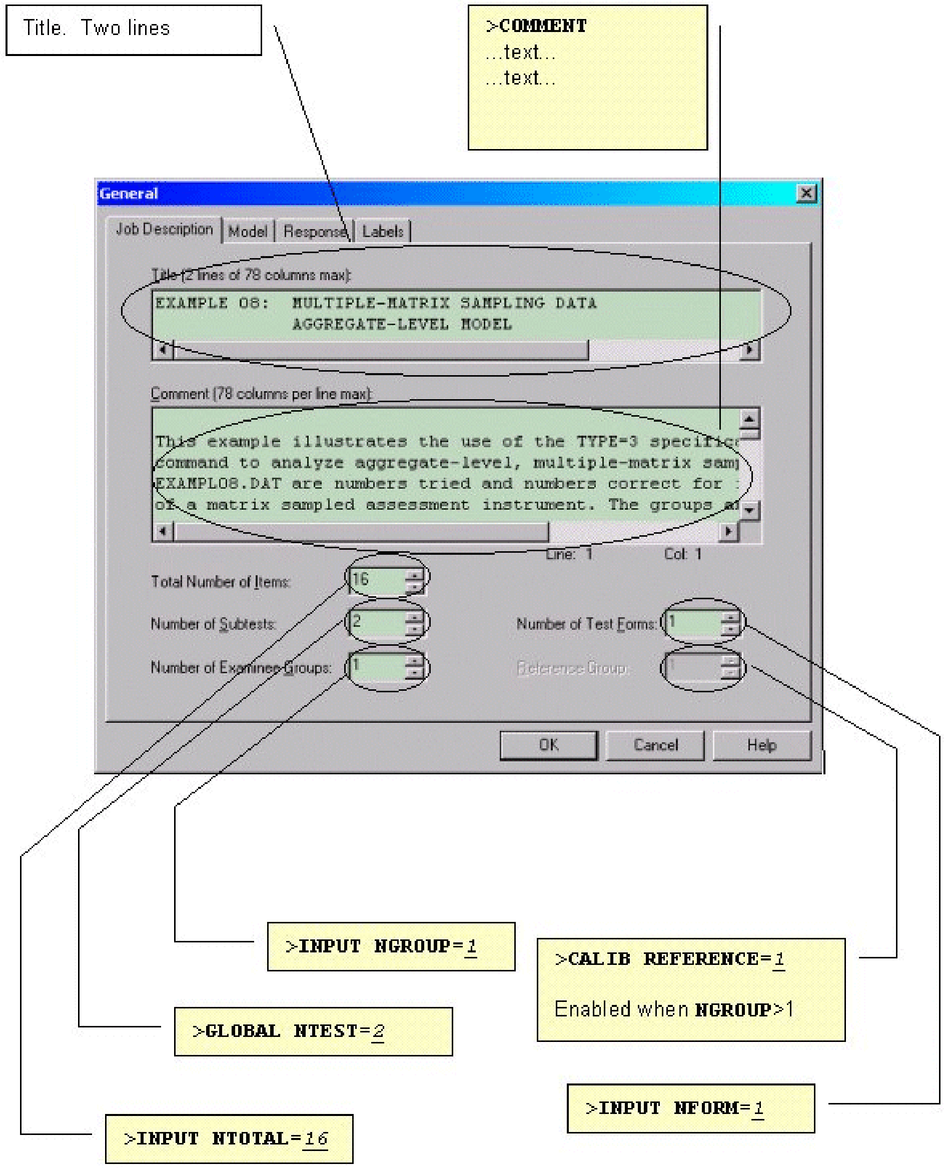
Setup menu: General dialog box

The **General** dialog box has four tabs on which the job description, model, type of response and test, group and item labels may be specified. The **Job Description** tab is shown below.

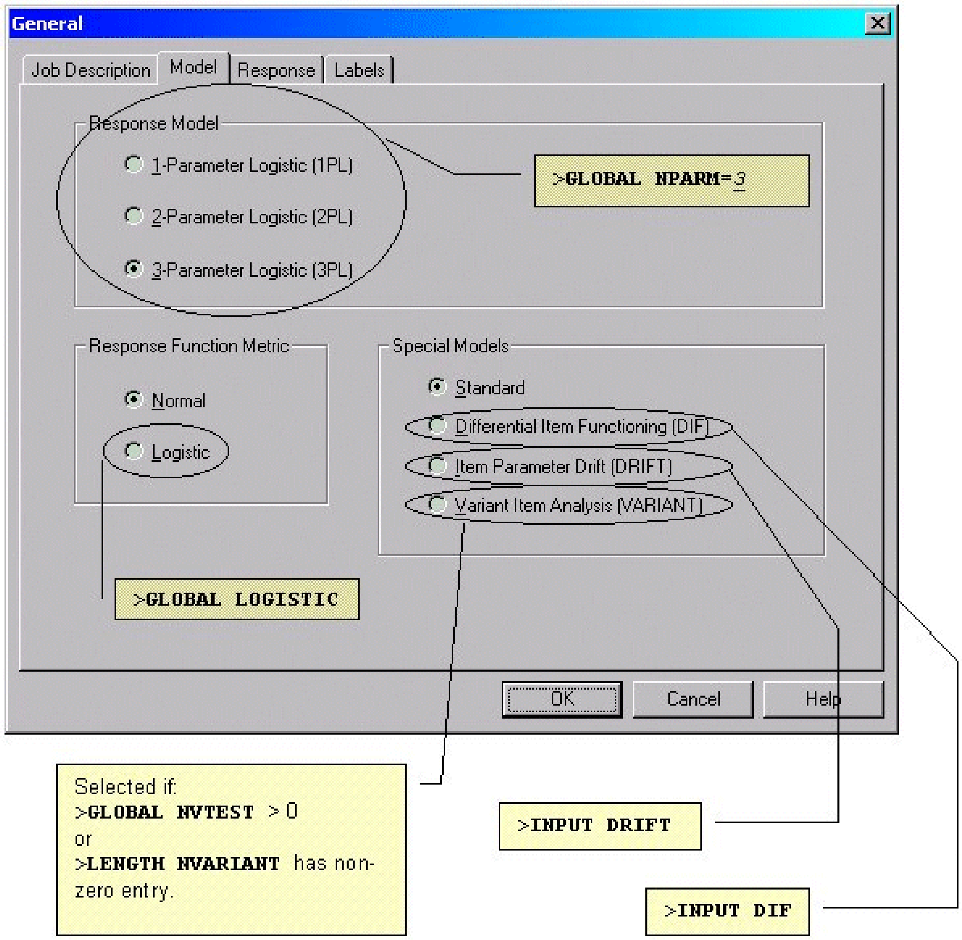
The top half of the **Job Description** tab on the **General** dialog box is used to provide a title and additional comments for the analysis. Below these fields, the number of items, subtests, groups and/or forms (if any), and the reference group in the case of a multiple-group analysis are entered. On the images shown here, links between the fields and the corresponding keywords are provided.

Related topics

- CALIB command: REFERENCE keyword (see Section 0)
- COMMENT command (see Section 0)
- GLOBAL command: NTEST keyword (see Section 3.2.5)
- INPUT command: NFORM, NGROUP, and NTOTAL keywords (see Section 3.2.7)
- TITLE command (see Section 3.2.16)



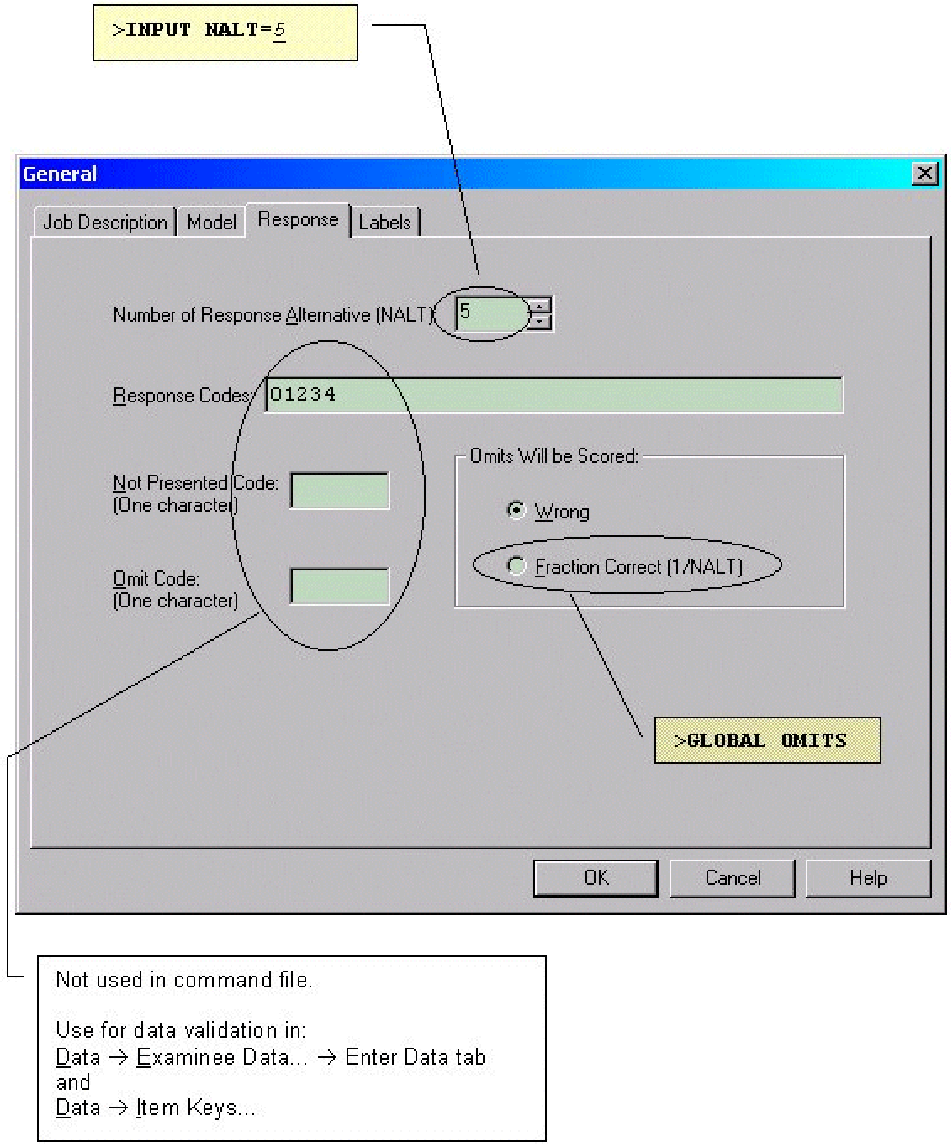
The second tab, **Model**, is used to select a 1-, 2-, or 3PL model and to specify the response function metric to be used. If variant items are to be included in the analysis, or a DIF or DRIFT multiple-group analysis is required, this is indicated in the **Special Models** group box.



Note that the selection of some models is dependent on the presence of other keywords in the syntax. For example, in order to request **Variant Item Analysis** the NVTEST keyword on the GLOBAL command should have a value larger than the default of 0, or the NVARIANT keyword on the LENGTH command should have a non-zero entry.

Related topics

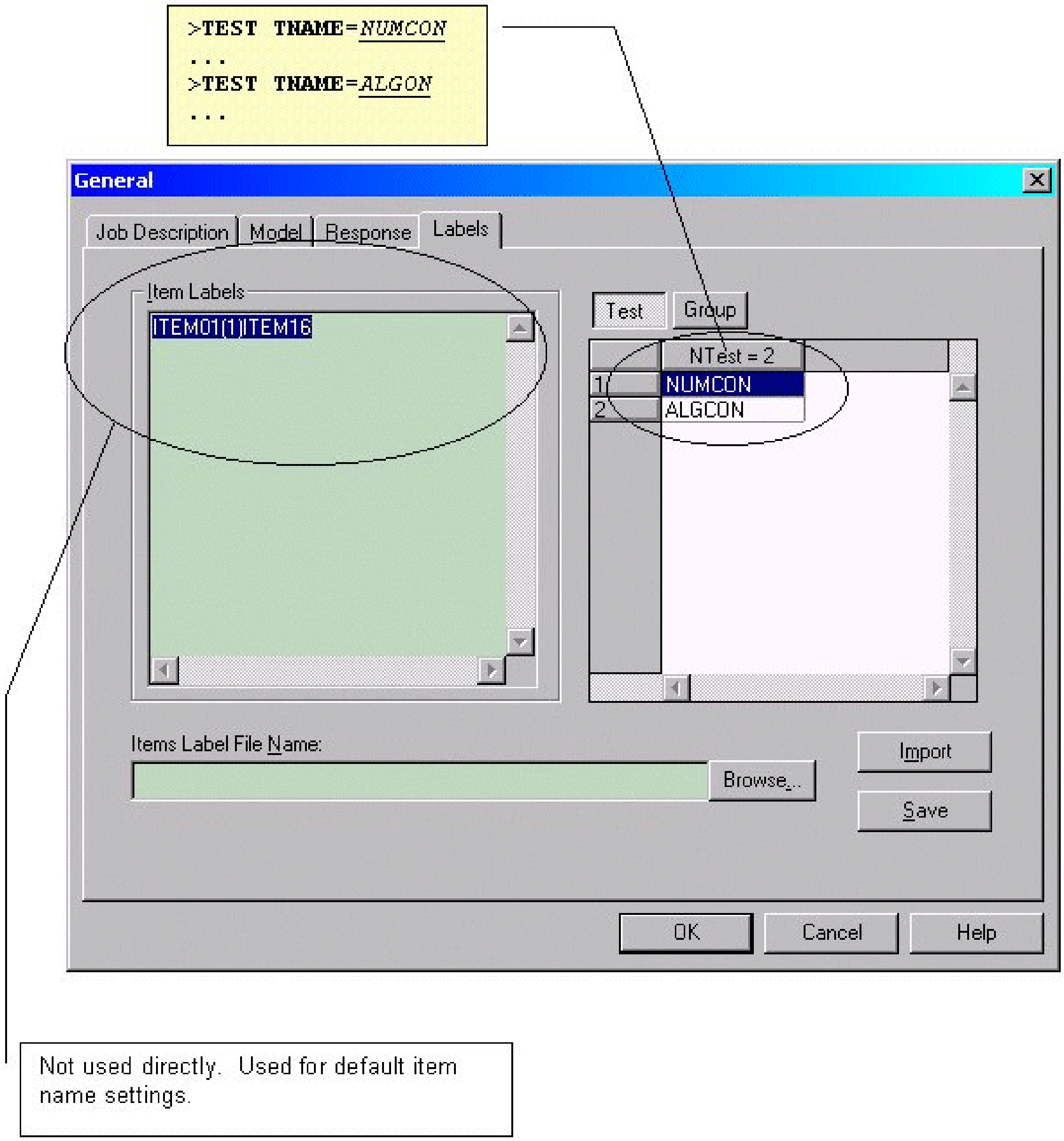
- GLOBAL command: LOGISTIC option, NPARM and NVTEST keywords (see Section 3.2.5)
- INPUT command: DIF and DRIFT options (see Section 3.2.7)
- LENGTH command: NVARIANT keyword (see Section 3.2.9)



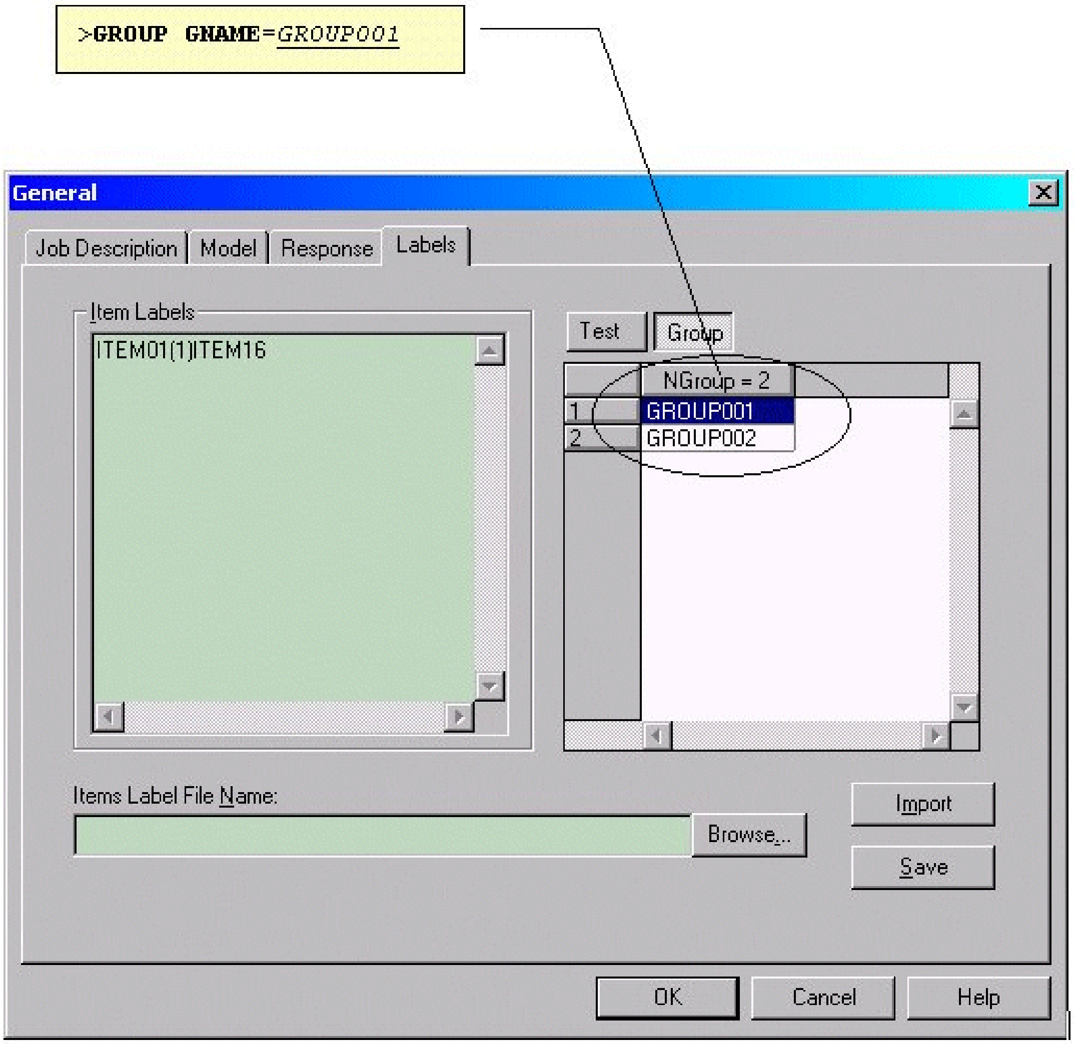
The **Response** tab allows specification of the number of response alternatives, and codes for the responses, not-presented and/or omitted items. In the case of a 3-PL model, the user may also request that omitted responses are scored fractionally correct. If the NPARM keyword on the GLOBAL command is not set to 3 to indicate a 3-PL model (see the previous tab) any instructions in the **Omits will be scored** group box will not be used.

Related topics

- GLOBAL command: OMITS option (see Section 3.2.5)
- INPUT command: NALT keyword (see Section 3.2.7)



Finally, the **Labels** tab provides the default item labels and group/test names. The user may enter names in the respective fields, or import item labels from an external file by using the **Browse** button next to the **Item Label File Name** field. After entering or selecting the file containing the item labels, click the **Import** button. Alternatively, after completion of the **Item Labels** and **Test** or **Group** fields, the user may save the labels to file using the **Save** button.



Related topics

- GROUP command: GNAME keyword (see Section 3.2.6)
- ITEMS command: INAMES keyword (see Section 3.2.8)
- TEST command: TNAME keyword (see Section 3.2.15)

Setup menu: Item Analysis dialog box

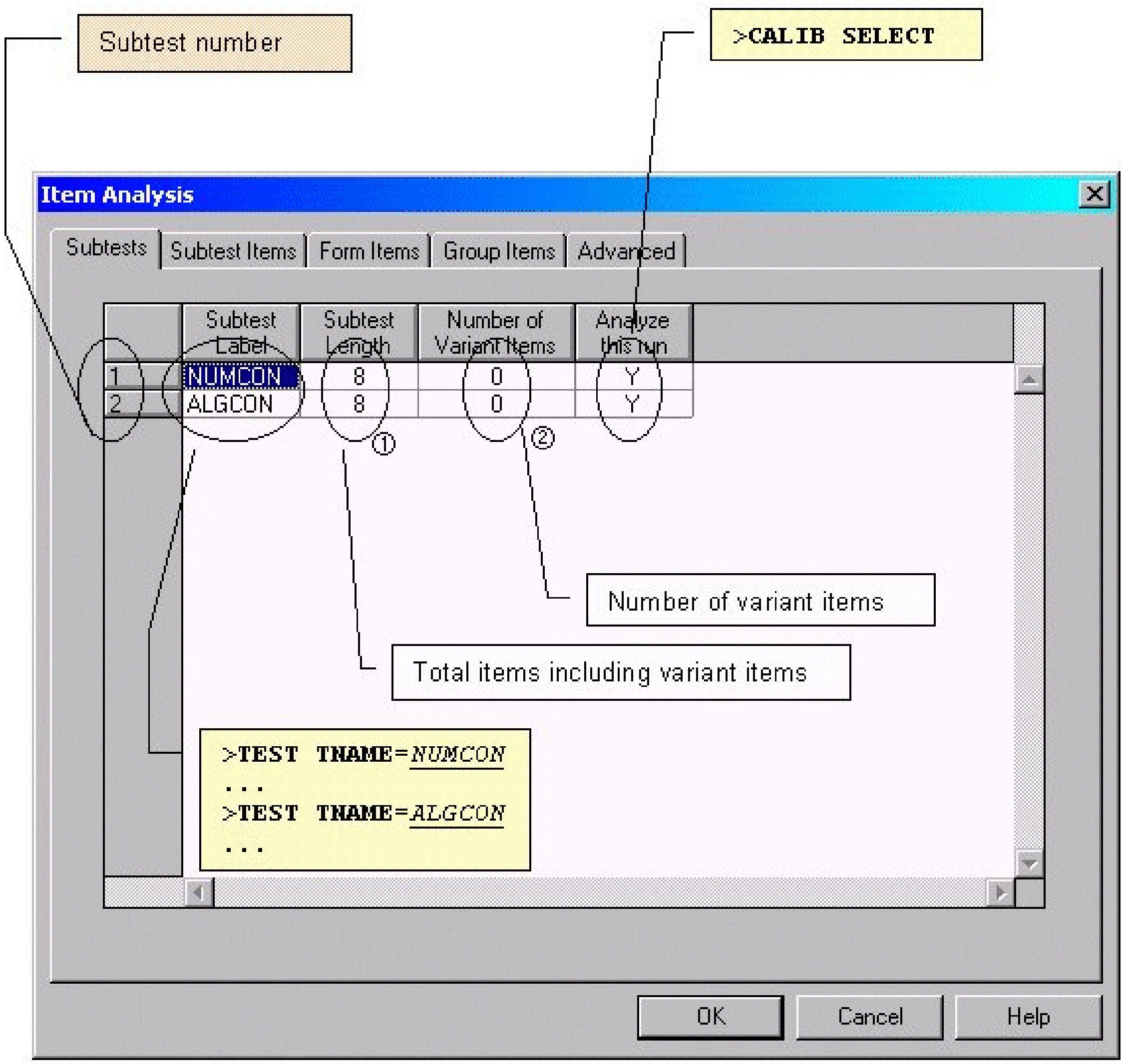
The **Item Analysis** dialog box has 5 tabs and is used to assign items to subtests, forms, and/or groups. In addition, subtests to be calibrated are selected here. Calibration specifications controlling the iterative procedure are also entered on this dialog box.

On the **Subtests** tab shown below, labels for the subtests are entered in the first fields. The next two fields are used to indicate the number of items per test. Note that variant items should also be indicated here. The final column is used to select the subtests for which item parameter estimation is required.

Related topics

- CALIB command: SELECT keyword (see Section 3.2.2)
- LENGTH command: NITEMS and NVARIANT keywords (see Section 3.2.9)

On the images below, links between the fields and the corresponding keywords are provided.



③: Number of subtest-only item = ① – ②

Table 2.2: Effect of test length

①②	Add or remove entries in: >LENGTH NITEMS Add or remove entries in the >PRIORS command: TMU， TSIGMA
②	Add or remove entries in the variant's >TEST command: INUMBER, INTERCPT, SLOPE, THRESHLD, GUESS, DISPERSN, FIX
①	Add or remove entries in the subtest's >TEST command: INUMBER, INTERCPT, SLOPE, THRESHLD, GUESS, DISPERSN, FIX Add or remove entries in the >PRIORS command: SMU, SSIGMA, ALPHA, BETA

The **Subtest Items** tab allows the user to assign specific items to the main and variant tests. Note that, if fewer items are selected here than were indicated on the **Subtests** tab, the information on the **Subtest Items** tab will be adjusted accordingly (see table above for specific information).

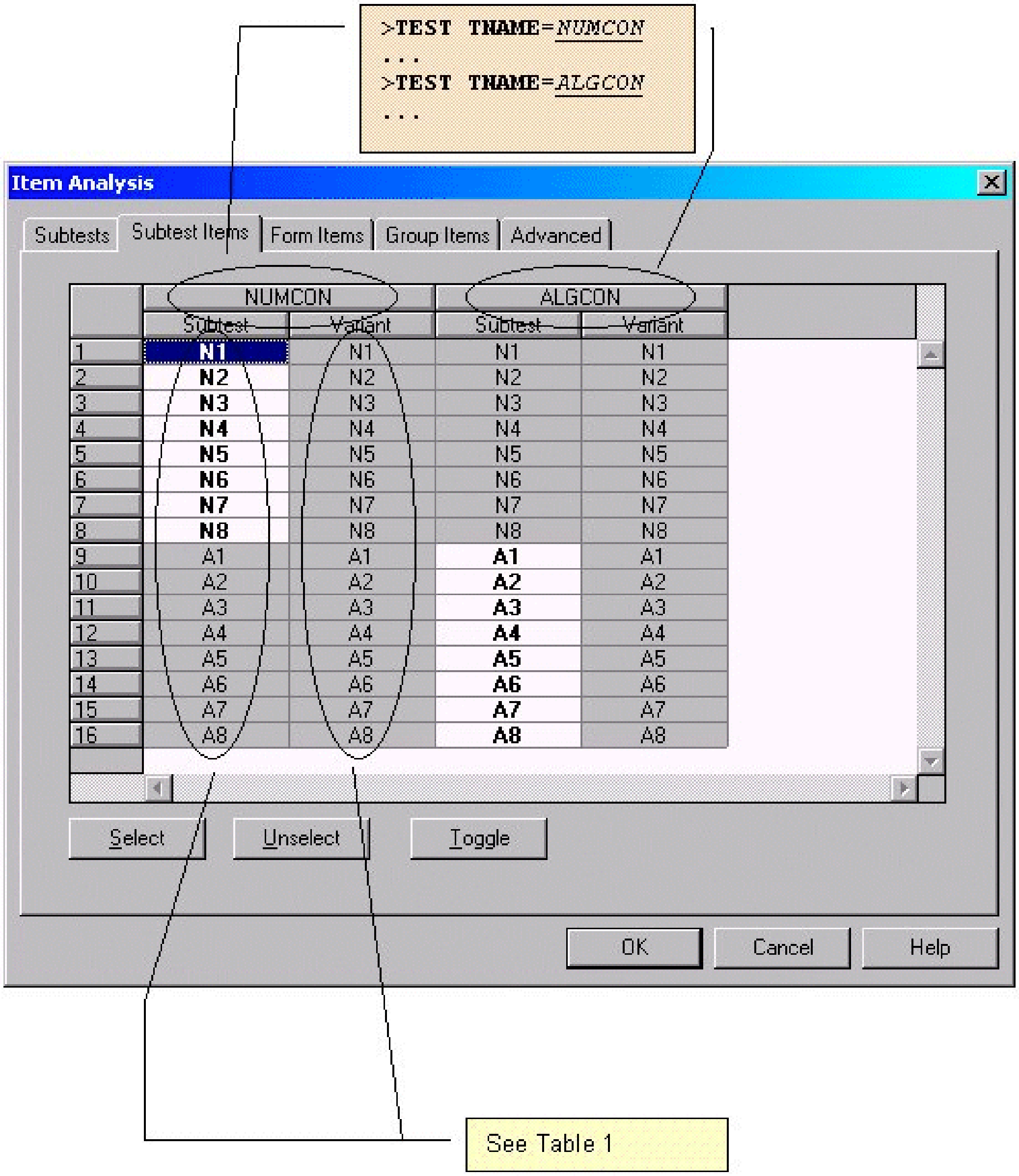
- The **Select** and **Unselect** buttons may be used to include or exclude single items or sets of items (selected by holding down the mouse button and dragging over a selection of items).
- Double-clicking a single item also reverses the state of the item.
- To reverse the state of a block of items, highlight the items and click the **Toggle** button.
- A variant item can only be selected when its corresponding subtest item is selected.
- Note that the table only supports rectangular blocks of items. There are two ways to highlight a rectangular block of items:

Click and drag: Left-click on any one corner of the block you want to highlight, hold the mouse button down and drag the mouse to the opposite corner of the block before releasing the mouse button. All items bounded by the opposite corners used will be highlighted.

Click-Shift-Click: Left-click on any corner of the block you want to highlight. Press and hold down the **Shift** key, move the mouse pointer to the opposite corner of the block and left-click. All items bounded by the opposite corners used will be highlighted.

Related topics

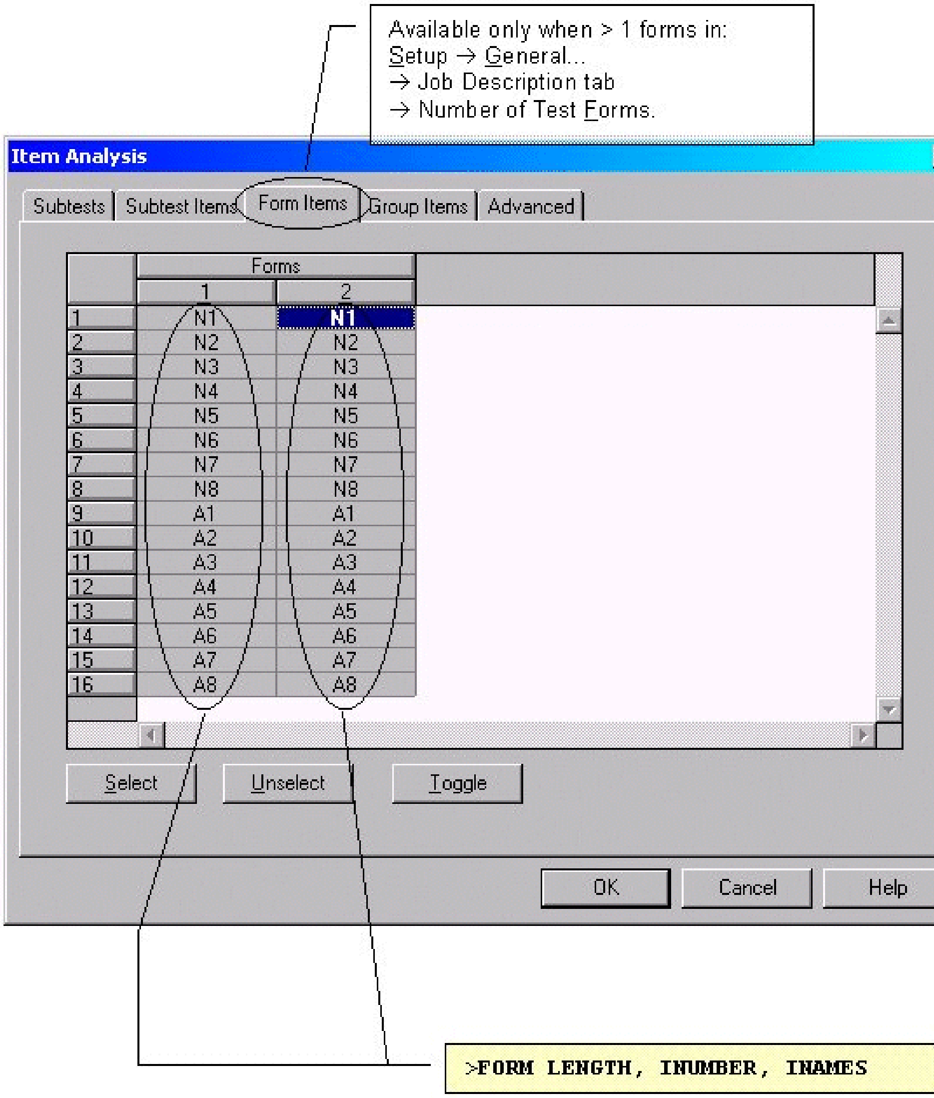
- LENGTH command: NITEMS keyword (see Section 3.2.9)
- TEST command: INUMBERS keyword (see Section 3.2.15)



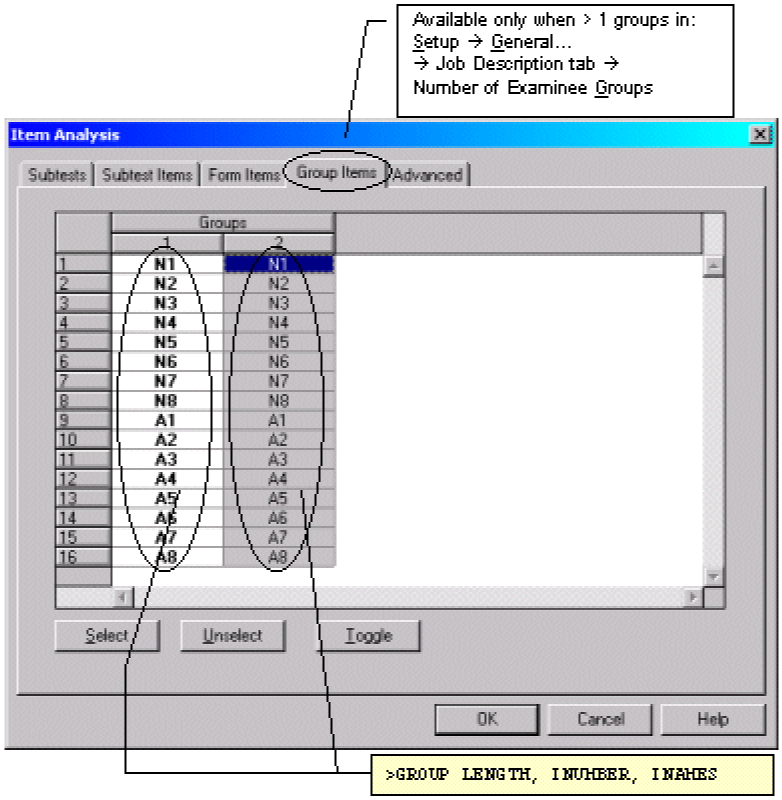
The next two tabs, **Form Items** and **Group Items**, are only available when a multiple-group or multiple-form analysis was indicated on the **Job Description** tab of the **General** dialog box. Both dialog boxes have the same form and mode of operation as the **Subtest Items** tab previously discussed, the only difference being that information entered here are recorded on the FORMS and GROUPS commands respectively. Both are used to indicate the length of and assignment of items to forms/groups.

Related topics

- FORM command: INAMES, INUMBERS, and LENGTH keywords (see Section 3.2.4)
- GROUP command: INAMES, INUMBERS, and LENGTH keywords (see Section 3.2.6)



Available only when > 1 forms in:
Setup → General...
→ Job Description tab
→ Number of Test Forms.
Item Analysis
Subtests
Subtest Items
Form Items
Group Items
Advanced
Forms
1
2
N1 N2 N3 N4 N5 N6 N7 N8 A1 A2 A3 A4 A5 A6 A7 A8
Select
Unselect
Toggle
OK
Cancel
Help
>FORM LENGTH, INUMBER, INAMES

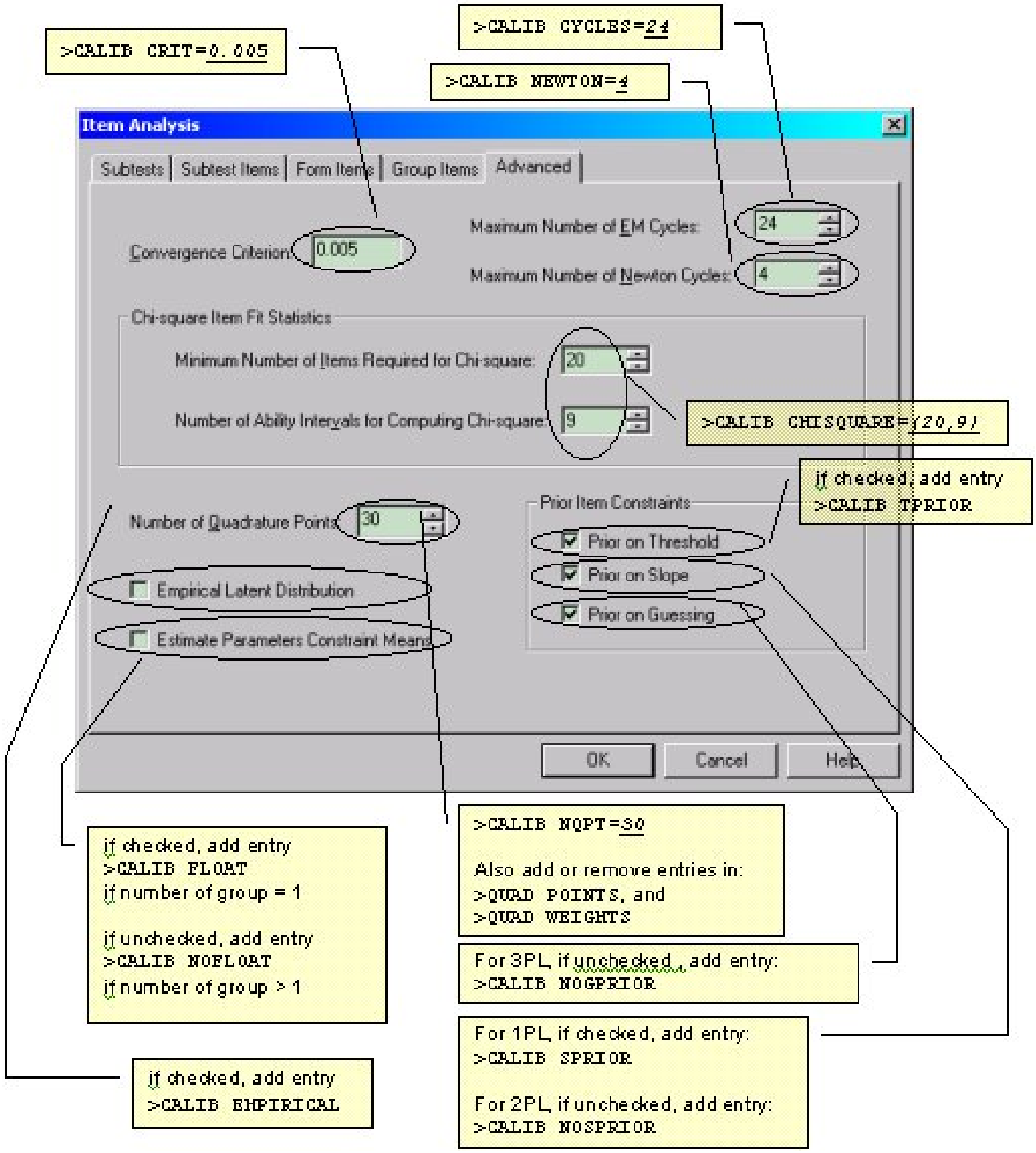


The final tab of the **Item Analysis** dialog box is the **Advanced** tab that controls the estimation of item parameters. Most of the information pertains to the CALIB command. The number of iterations and convergence criterion are set at the top of the dialog box, while the number of items and ability intervals for calculation of χ^2 item fit statistics are specified in the **Chi-square Item Fit Statistics** group box. At the bottom of the dialog box, prior item constraints may be requested and the estimation of the means of the prior distributions on the item parameters specified to be kept at a fixed value or to be estimated along with the parameters.

If a 3PL model is selected, all the prior check boxes in the **Prior Item Constraints** group box will be enabled. In the case of a 2PL model, the **Prior on Guessing** check box is disabled, while both the **Prior on Guessing** and **Prior on Slope** check boxes are disabled when a 1PL model is fitted to the data.

Related topics

- CALIB command: CHI, CRIT, CYCLES, NEWTON, NQPT, FLOAT, EMPIRICAL, GPRIOR, SPRIOR, and TPRIOR keywords (see Section 3.2.2)
- QUAD command: POINTS and WEIGHTS keywords (see Section 3.2.11)



Setup menu: Test Scoring dialog box

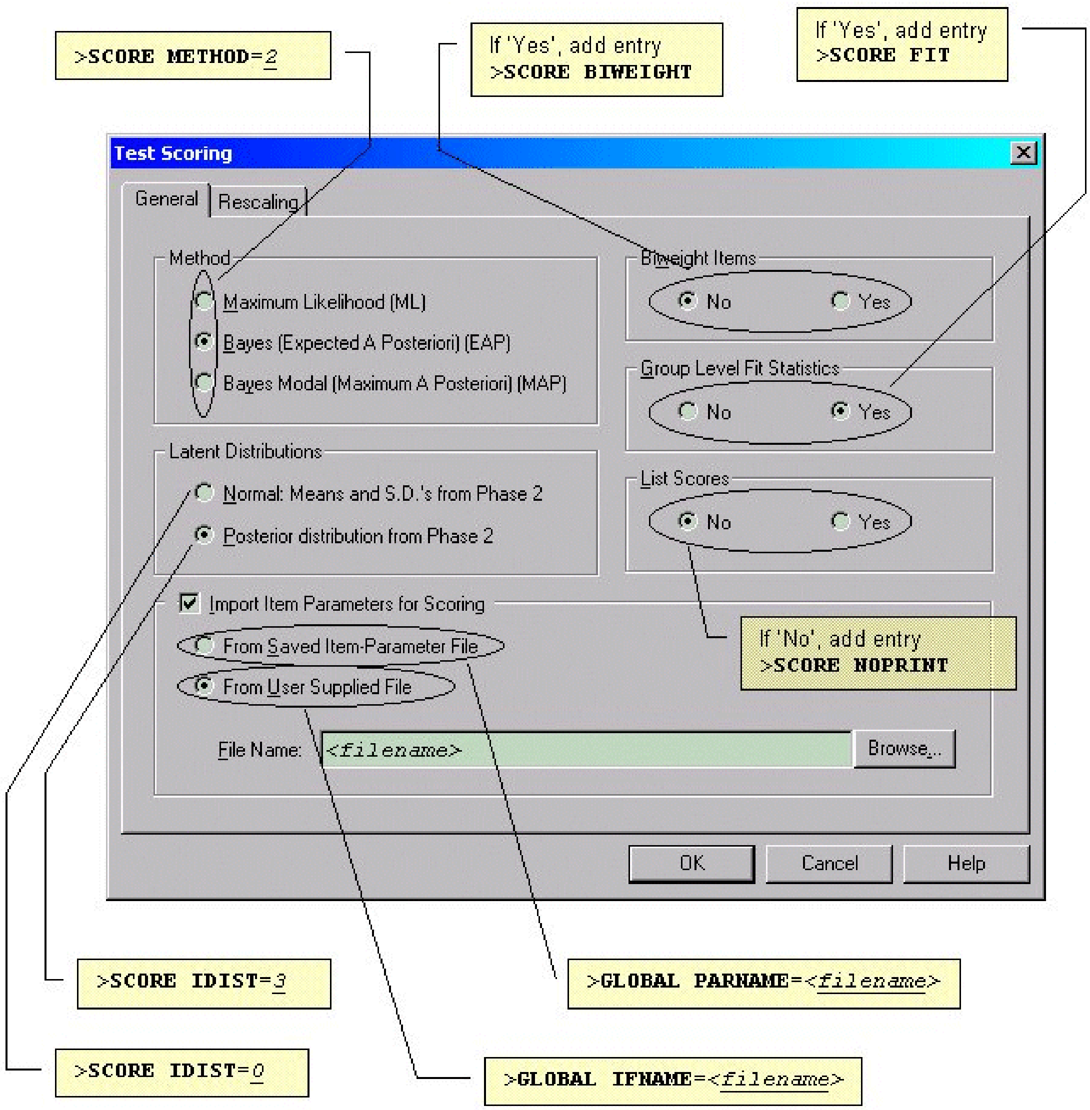
Information entered on the **Test Scoring** dialog box controls the type of scoring performed in Phase 3 of the analysis. The **General** tab of this dialog box is used to select the method of scoring and to import item parameters for scoring from previously saved files. In the latter case, the **Browse** button at the bottom of the tab can be used to locate the file containing the item parameters to be used for scoring.

Group-level fit statistics, the suppression of printing of scores to the output file when scores are saved to an external file using the SCORE keyword on the SAVE command, and bi-weighted estimates robust to isolated deviant responses are requested using the **Group Level Fit Statistics**,

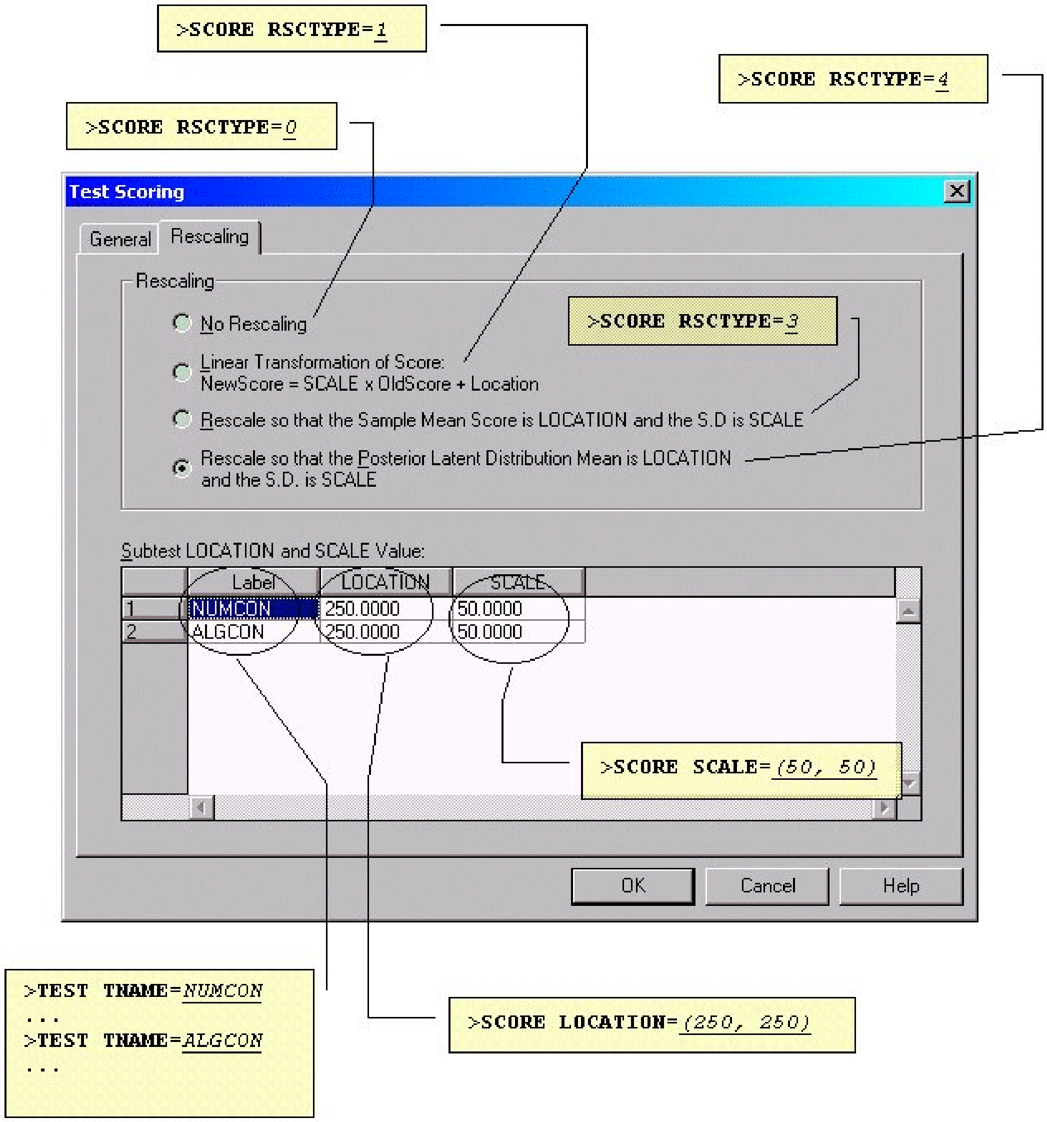
List Scores, and **Biweight Items** radio buttons. On the images below, links between the fields and the corresponding keywords are provided.

Related topics

- GLOBAL command: IFNAME and PRN keywords (see Section 3.2.5)
- SCORE command: BIWEIGHT, FIT, and NOPRINT options (see Section 3.2.14)
- SCORE command: IDIST and METHOD keywords



The **Rescaling** tab is associated with the RSCTYPE, LOCATION and SCALE keywords on the SCORE command and is used to request the scaling of the ability scores according to user-specified values. Provision is made for different scaling options for different subtests.

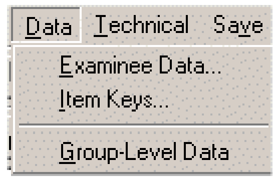


Related topics

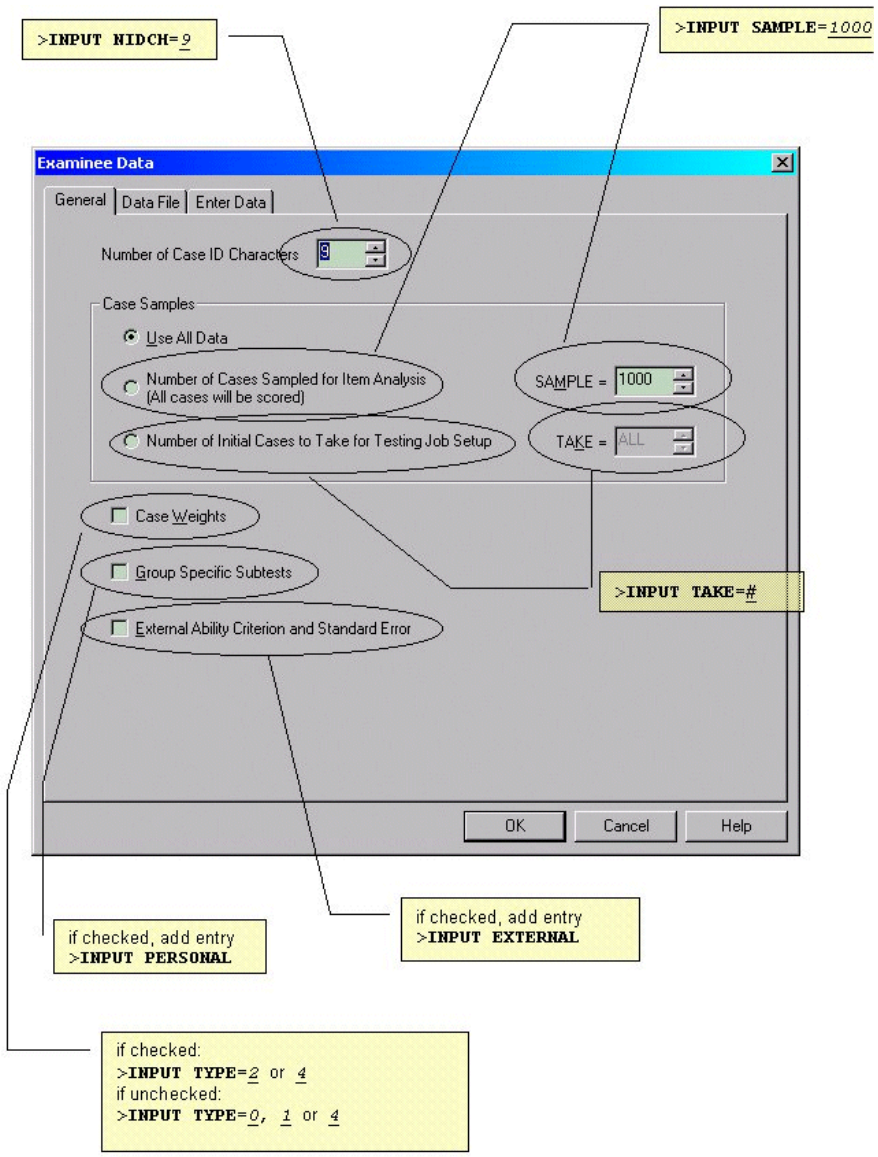
- SCORE command: LOCATION, RSCTYPE, and SCALE keywords (see Section 3.2.14)

2.4 Data menu

The **Data** menu is used to enter data or to provide information on the data file; type and number of records in the data file; and answer, omit and not-presented keys if applicable (**Item Keys** option). A distinction is made between single-subject and group-level data (**Examinee Data** and **Group-level Data** tabs respectively).



Data menu: Examinee Data dialog box

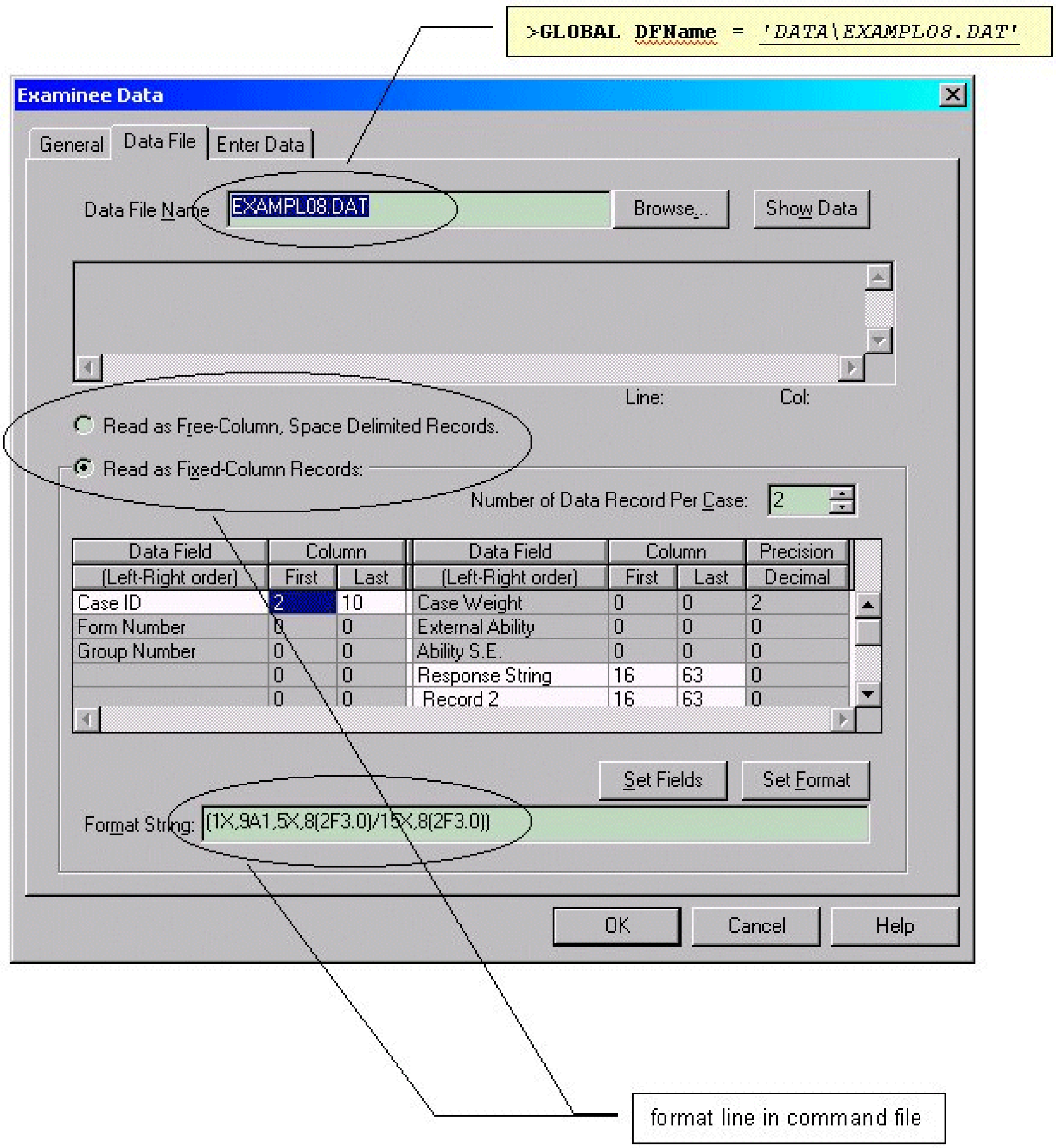


The **Examinee Data** dialog box deals with single-subject data. On the **General** tab of this dialog box, the type and number of data records to be used in the analysis are specified. All of the entries on this dialog box correspond to keyword values on the INPUT command, as indicated on the image below. Note that when the check box labeled **External Ability Criterion and Stand-**

ard Error is checked, the **External Ability** and **Ability S.E.** data fields on the **Data File** tab are enabled.

Related topics

- INPUT command: EXTERNAL, NIDCHAR, SAMPLE, TAKE, and TYPE keywords (see Section 3.2.7)
- INPUT command: PERSONAL option



The name of the raw data file and the format of the data are specified on the **Data File** tab. An external data file may be selected using the **Browse** button at the top of the tab. Data may be displayed in the window below the **Data File Name** field by clicking the **Show Data** button.

Data can be read in free- or fixed column format. For fixed-format data, a format string is required to tell the program where in the file each data element is located. To ensure the accuracy of the format information, the column locations of the various data elements can be determined directly using the spreadsheet in which the data are displayed: clicking directly in the display paces a cursor whose exact position is shown by the **Line:** and **Col:** indicators.

In the **Read as Fixed-Column Records** group box, the user can indicate the number of data records per case and then fill in the information on the positions of the case ID, the form and group numbers (if applicable), and the responses. The **Set Format** button is then clicked to create automatically a format statement in the **Format String** data field. Alternatively, the format statement may be entered directly in the **Format String** data filed. Clicking the **Set Fields** button will then automatically fill in the fields in the **Read as Fixed-Column Records** group box. Note that with either method the response string **must** be continuous, that is, there can be no spaces in the response string. Any attempt to specify non-continuous response data will result in incorrect format and/or response information, and the data will not be read correctly. For example, if the response string is "10A,1X,10A,1X,15A" clicking the **Set Fields** button does not correctly set the **Format String** data field.

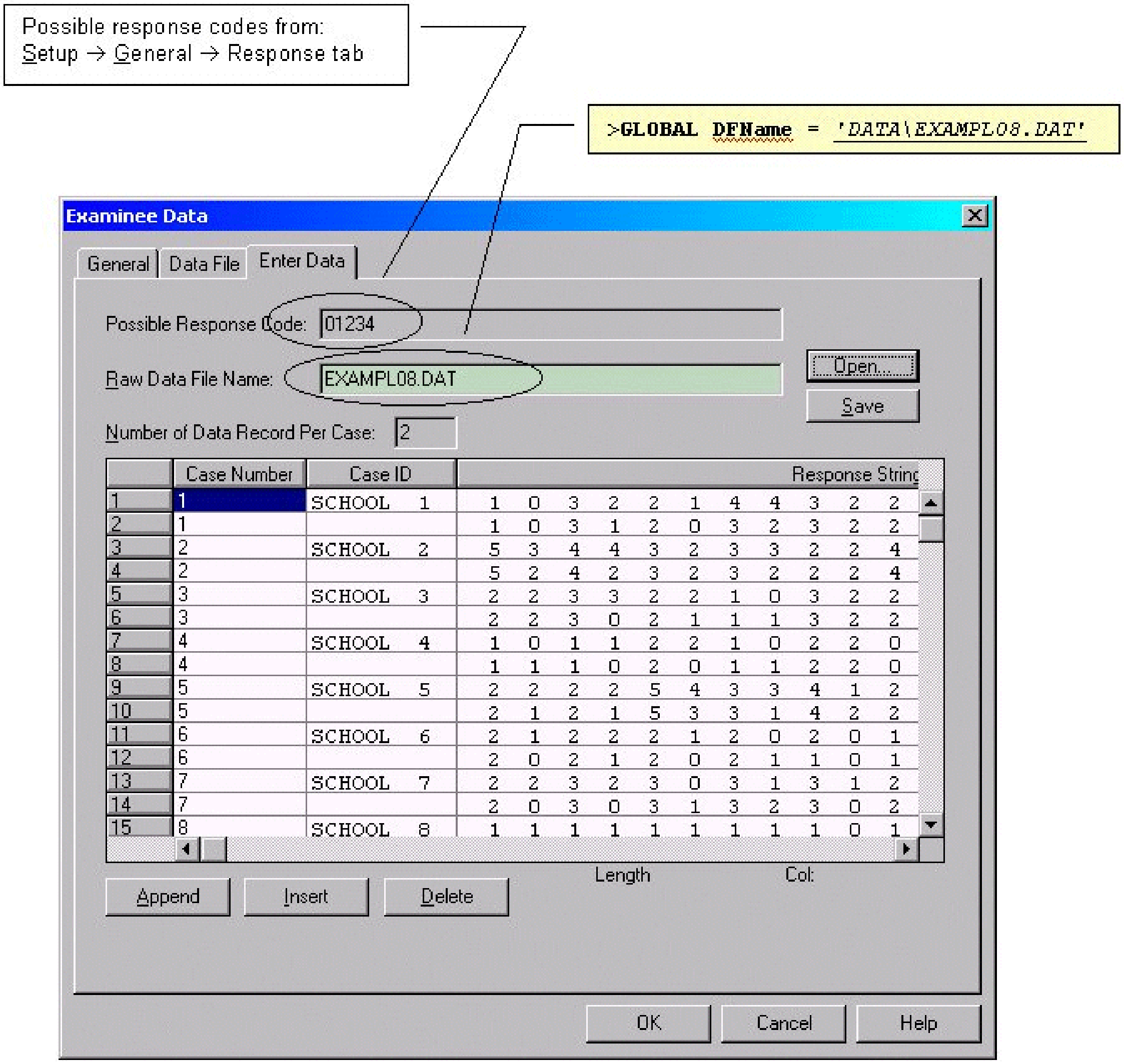
Related topics

- GLOBAL command: DFNAME keyword (see Section 3.2.5)
- Input files (see Section 3.2.18)
- INPUT command: NFMT keyword (see Section 3.2.7)
- Variable format statement (see Section 3.2.16)

Data may be entered interactively on the third tab of the **Examinee Data** dialog box. The **Append** button is used to add a new case to the end of the data file. The **Insert** button is used to insert a new case at the current cursor location, while the **Delete** button is used to delete lines of entered data. For example, if case 10 is highlighted in the table, pressing the **Insert** button will insert a new case at case 10, and all cases starting from 10 will move one row down in the table.

Related topics

- GLOBAL command: DFNAME keyword (see Section 3.2.5)
- INPUT command: NFMT keyword (see Section 3.2.7)
- Variable format statement (see Section 3.2.16)



Data menu: Item Keys dialog box

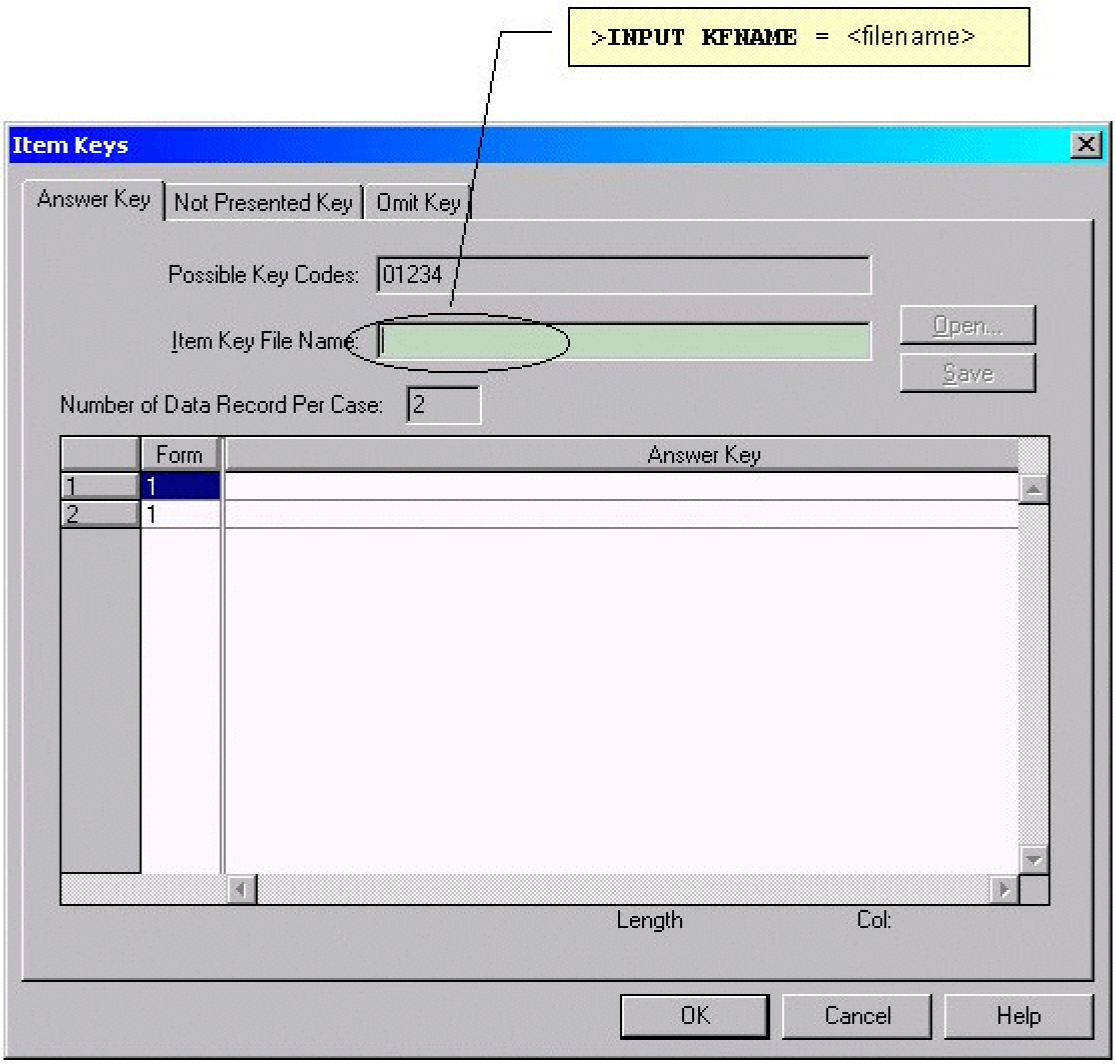
The sole purpose of the **Item Keys** option on the **Data** menu is to provide the option to use answer, not-presented or omit keys. The three tabs on the **Item Keys** dialog box are similar. Possible key codes consist of the entire list of "possible keys". The information is taken from the **Response Codes** edit box on the **Response** tab of the **General** dialog box on the **Setup** menu.

On the first tab, **Answer Key**, an answer key may be read from an external file using the **Open** button and browsing for the file containing the answer key, or entered interactively in the window towards the top of the tab.

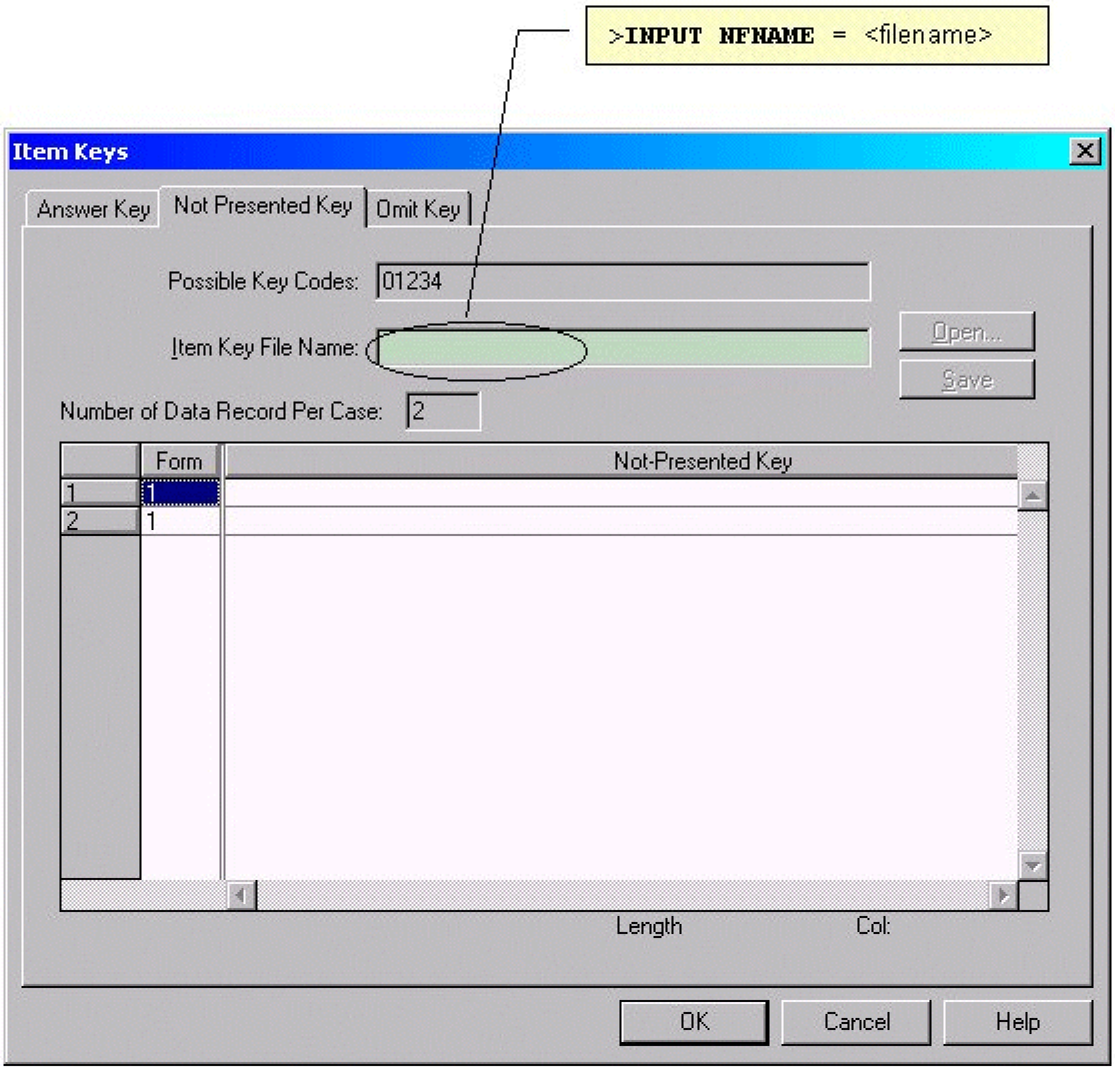
In the case of multiple forms, a separate answer key for each form should be provided. The format of the keys should be the same as that used for the raw response data. If a key is entered interactively, the **Save** button may be used to save the entered information to an external file. The file used as answer key is referenced by the KFNAME keyword on the INPUT command.

Related topics

- INPUT command: KFNAME keyword (see Section 3.2.7)



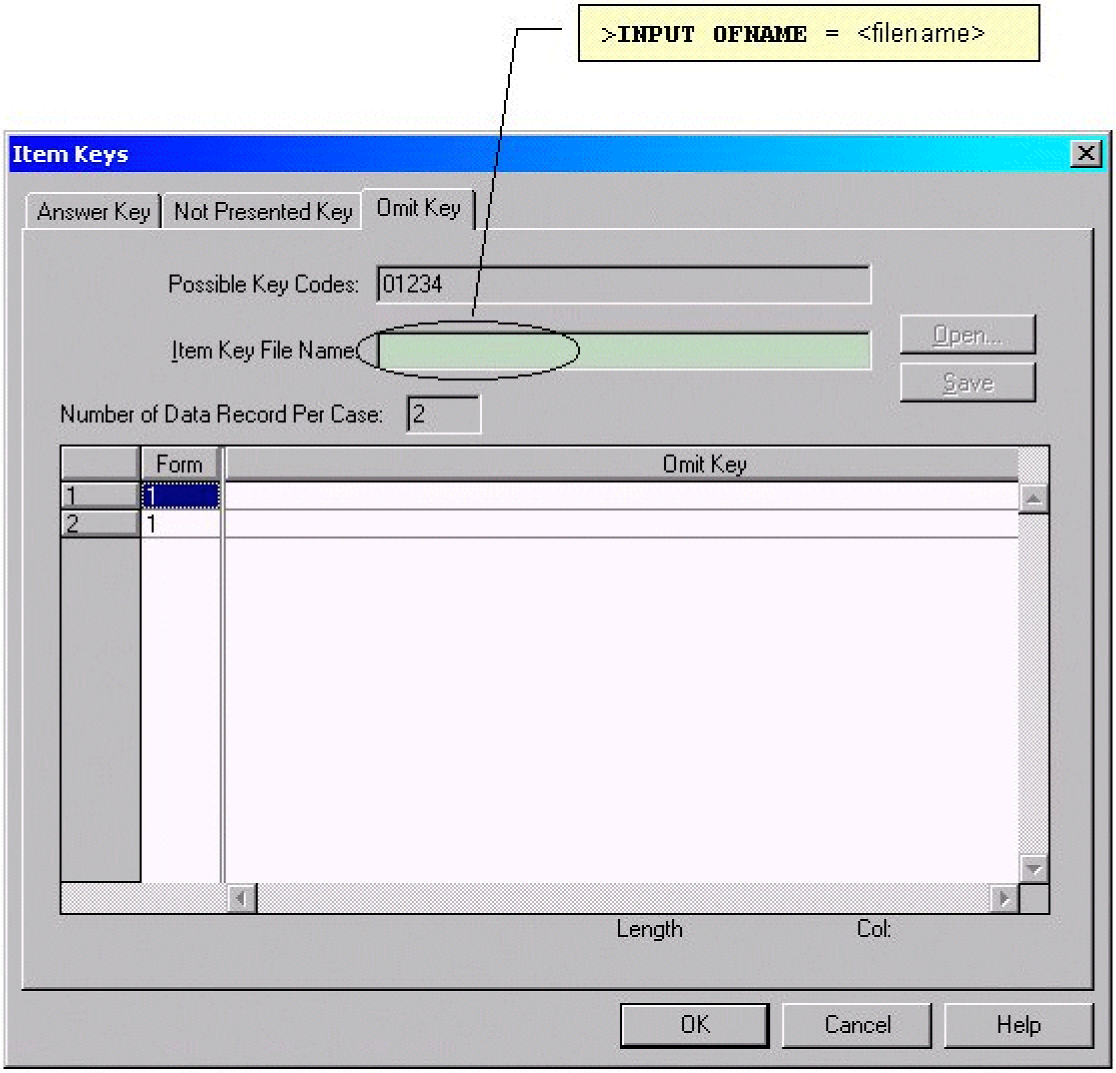
The second tab is used for the not-presented key (if any) and information entered here is echoed to the NFNAME keyword on the INPUT command.



Related topics

- INPUT command: NFNAME keyword (see Section 3.2.7)

The **Omit Key** tab is used for the omit key, if any. This tab corresponds to the OFNAME entry on the INPUT command in the completed command file.



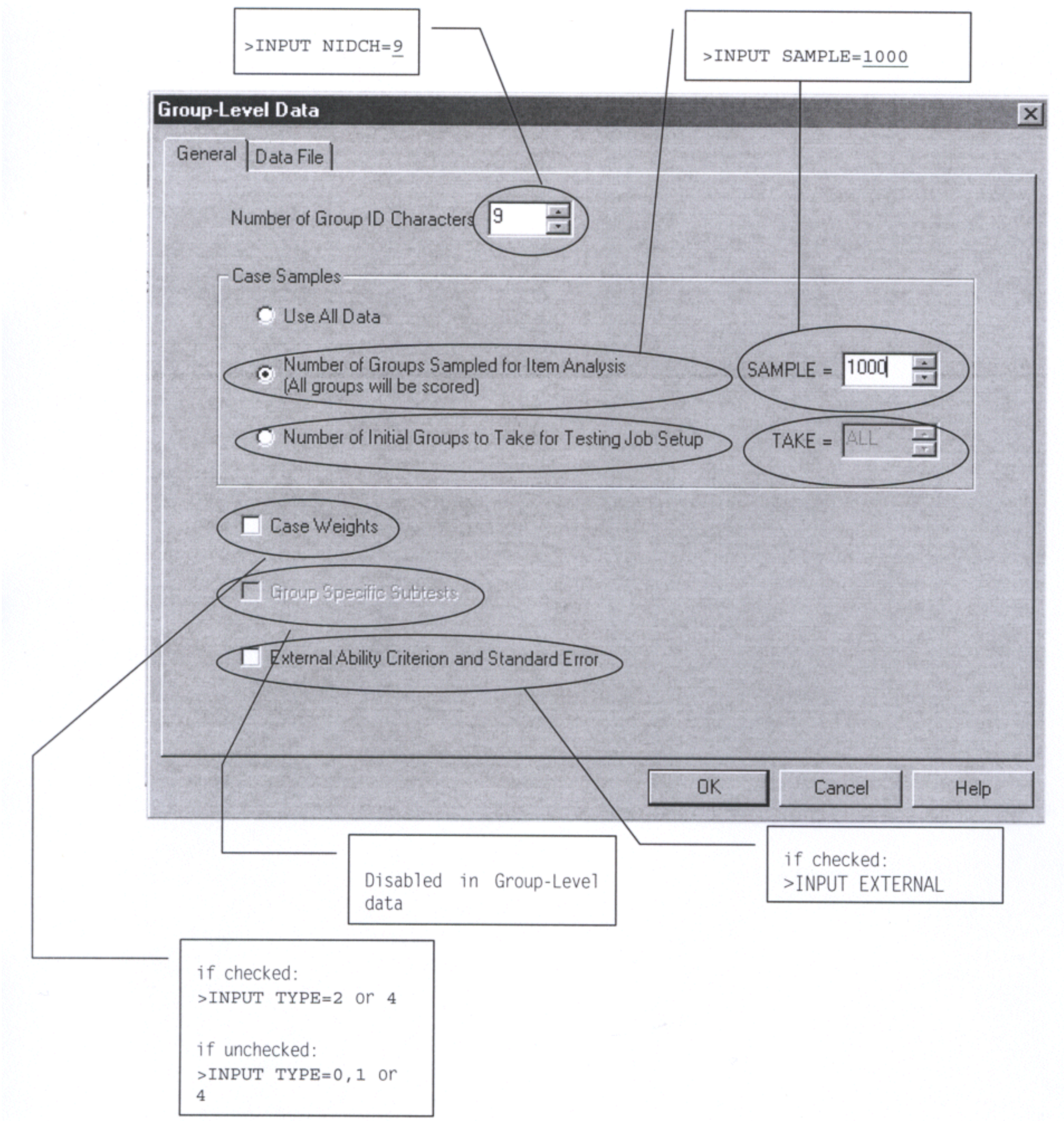
Related topics

- INPUT command: OFNAME keyword (see Section 3.2.7)

Data menu: Group-Level Data dialog box

The **Group-Level Data** dialog box is similar in purpose to the **Examinee Data** dialog box where single-subject data may be entered. On this dialog box, however, information on the structure of group-level data to be used in analysis is provided.

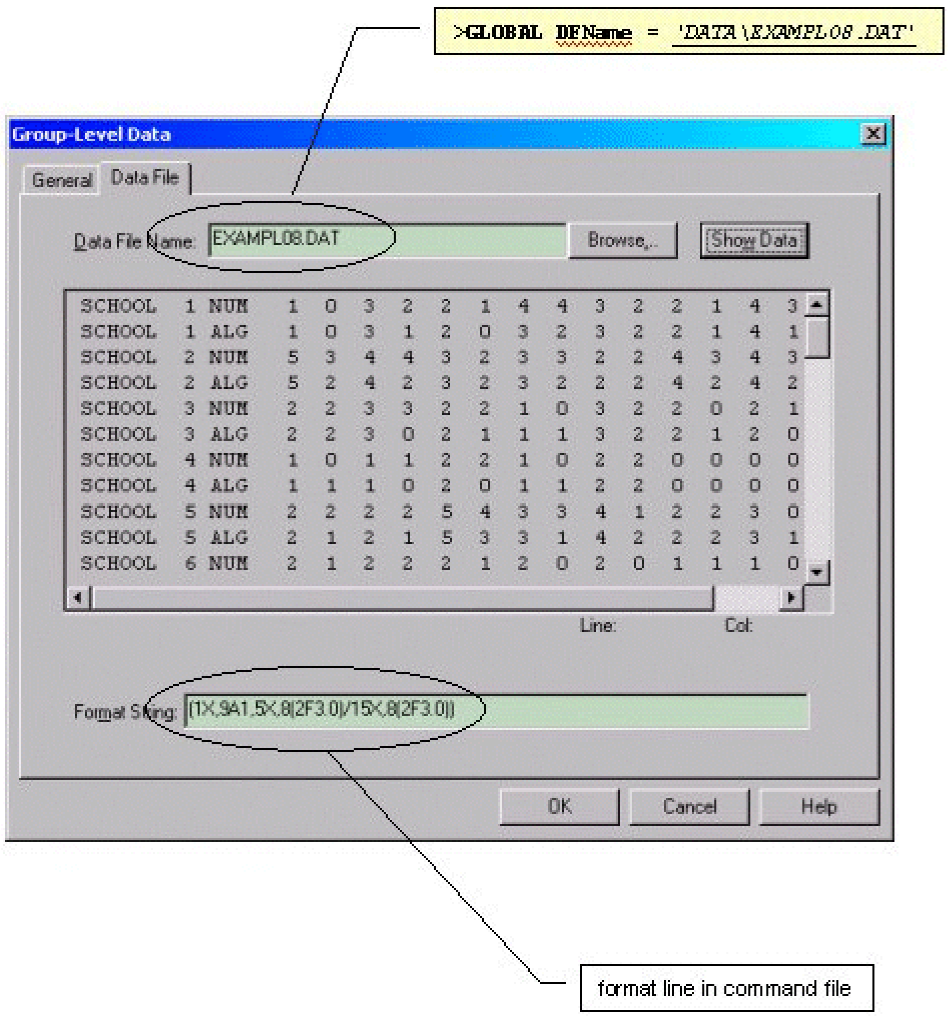
The **General** tab is used to provide information on the number of groups, group ID, and number of data records and weights, if any, to use in analysis. All entries correspond to keywords on the INPUT command.



Related topics

- INPUT command: EXTERNAL, NIDCHAR, SAMPLE, TAKE, and TYPE keywords (see Section 3.2.7)

The **Browse** button on the **Data File** tab allows the user to browse for the file containing the group-level data. After clicking the **Show Data** button the contents of the selected file are displayed in the window below these buttons. The **Format String** field should be completed according to the contents of the file. In contrast to item responses in the case of single-subject data, which are read in "A" format, the frequencies in group-level data files are read in "F" format as shown below.

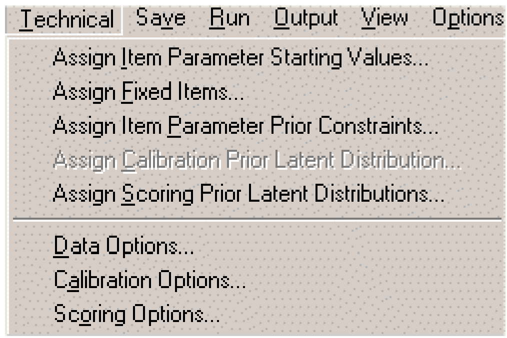


Related topics

- GLOBAL command: DFNAME keyword (see Section 3.2.5)
- INPUT command: NFMT keyword (see Section 3.2.7)
- Variable format statement (see Section 3.2.16)

2.5 Technical menu

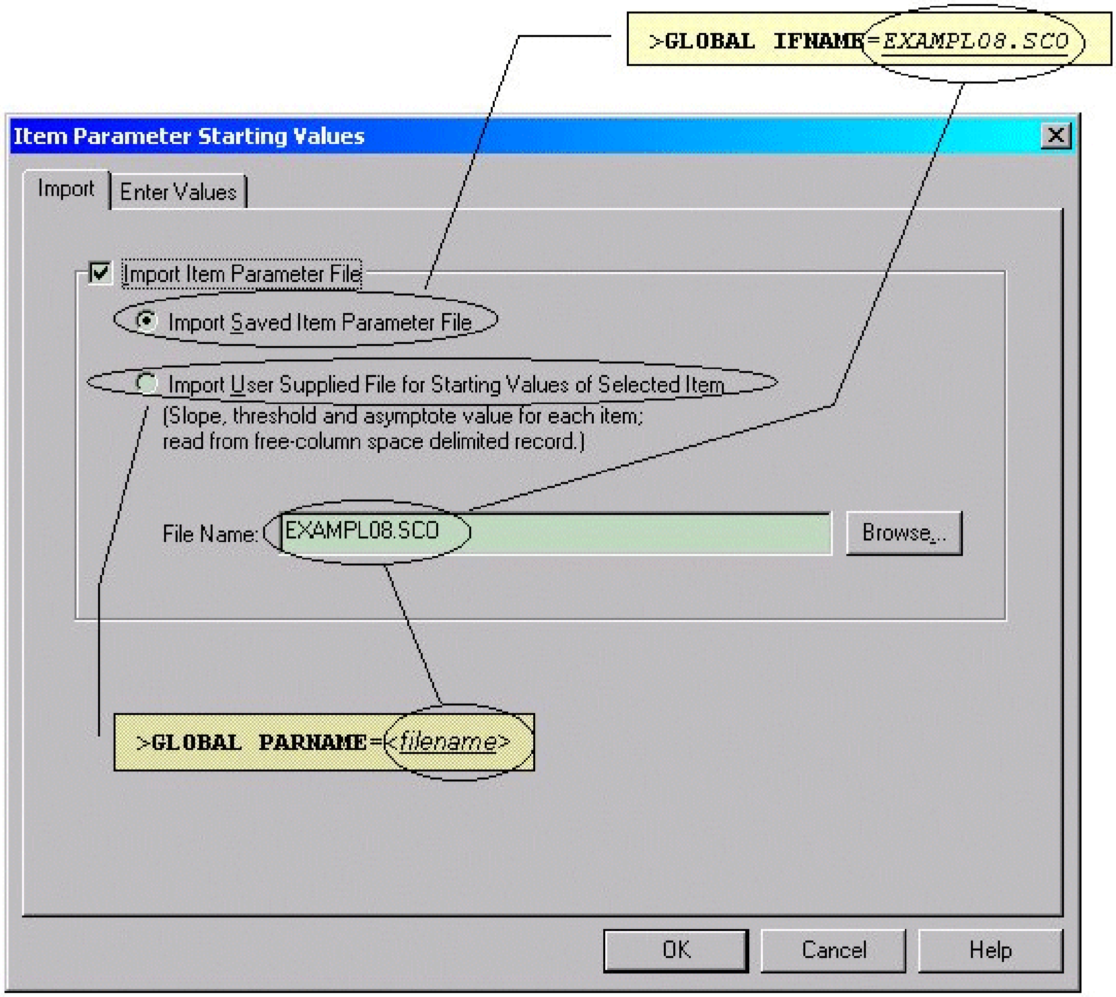
The first set of options on the **Technical** menu is used to assign starting values, prior constraints, and information on prior latent distributions for both calibration and scoring during the analysis. The last three options on this menu provide the user with the option to exercise even more control over the sampling and EM and Newton cycles (**Data Options**); request a Rasch model, plots per group or to prevent the adjustment of the latent distribution to a mean of 0 and S.D. of 1 (**Calibration Options**), and finally to calculate domain scores based on a user-supplied file containing information on previously calibrated items (**Score Options**).



Technical menu: Item Parameter Starting Values dialog box

The **Assign Item Parameter Starting Values** option on the **Technical** menu may be used to import starting values for item parameters from a saved item parameter or user-supplied file or, alternatively, to enter starting values interactively.

The first tab on the **Item Parameter Starting Values** dialog box is used to select a previously created file. To use an item parameter file created during a previous BILOG-MG analysis, check the radio button next to the **Import Saved Item Parameter File** option. If starting values are provided through a user-supplied file, check the radio button next to the **Import User Supplied File …** option. The **Browse** button is used to locate the file.



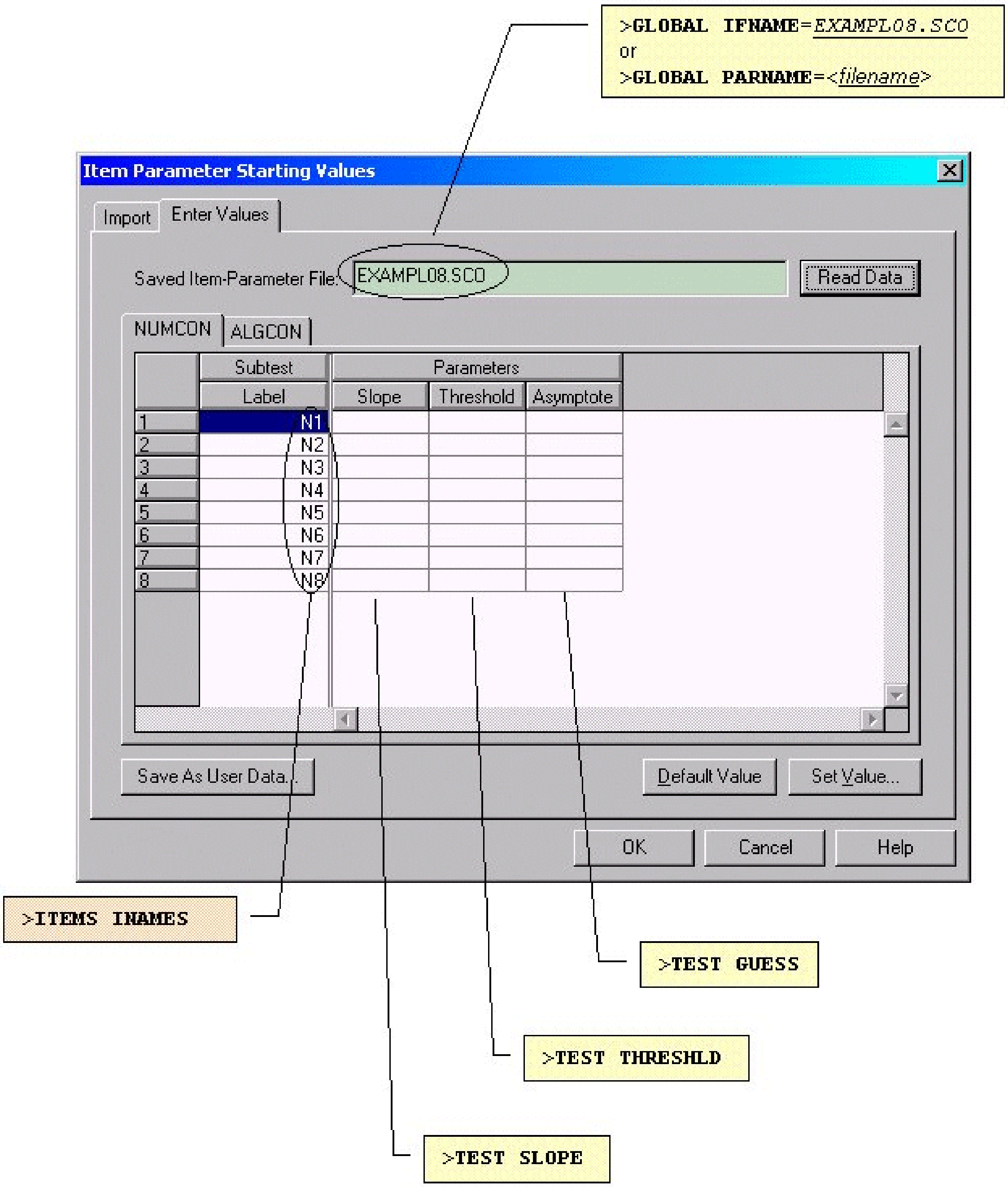
Enter starting values for the item parameters on the **Enter Values** tab to set values for the corresponding keywords on the TEST command. A subset of slope, threshold or asymptote parameters may be selected by holding the mouse button down and dragging until the selection is complete. Clicking the right mouse button will display a pop-up menu that can be used to assign values to the parameters.

All selected parameters may be set to a specific value or to the default value. In addition, the user may select one parameter and assign a value to this and all other parameters below it by selecting the appropriate option from the pop-up menu. Alternatively, the **Default Value** or **Set Value** buttons may be used to assign values to the selected parameters.

There are two ways to select the cell for parameter values:

- Select a rectangular block of cells. Use either the click-and-drag or click-shift-click method described in the discussion of the **Item Analysis** dialog box of the **Setup** menu. Note that cells selected must be one continuous rectangular block.
- Select one or more columns by clicking on the column header. The click-and-drag method

works when selecting a continuous block of columns. To select a disjoint block of columns, press and hold the **Ctrl** key down when clicking the header.



Note that when selecting a block of cells, the **Shift** key is used. When selecting a block of columns through clicking column headers, the **Ctrl** key is used.

Clicking the column header changes the selection state of the entire column. It toggles the items in the column from the "selected" state to the "unselected" state and vice versa. The **Save as Us-**

er Data option may be used to provide a name for the external file to which INPUTis saved with the file extension ***.prm**.

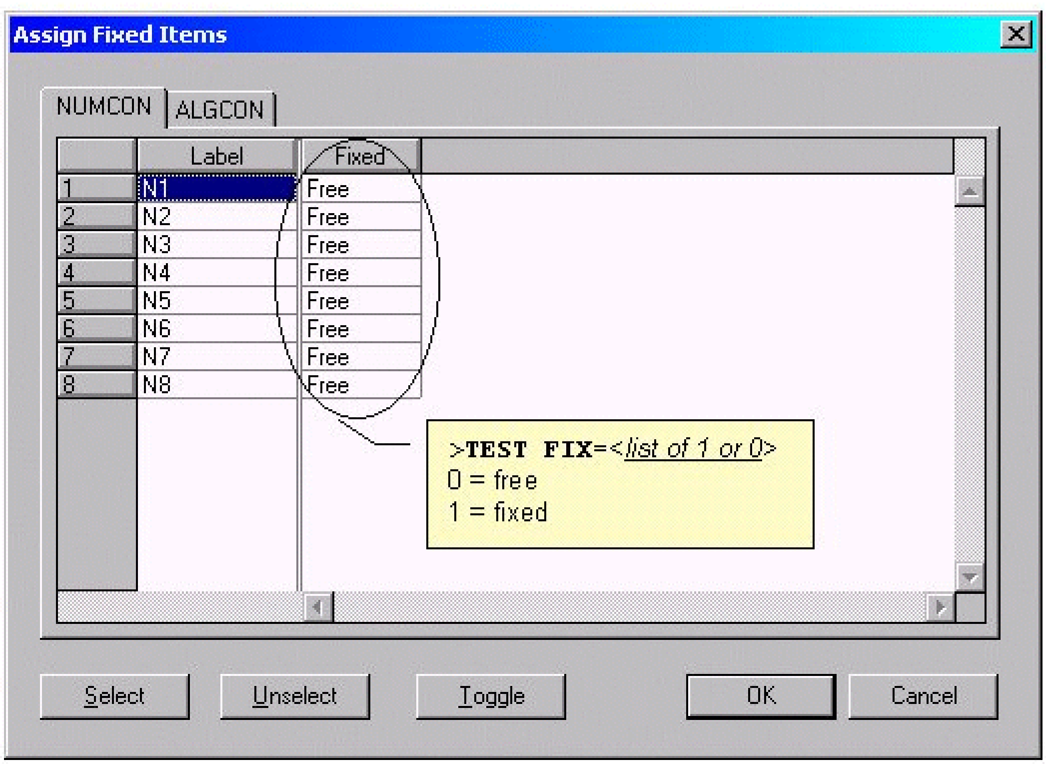
Related topics

- TEST command: DISPERSN, GUESS, INTERCPT, SLOPE, and THRESHLD keywords (see Section 3.2.15)

Technical menu: Assign Fixed Items dialog box

This dialog box is associated with the FIX keyword on the TEST command, which is used to indicate which items of a subtest are free to be estimated, and which are to be held fixed at their starting values.

As with the **Enter Values** tab on the **Item Parameter Starting Values** dialog box discussed above, cells may be selected in rectangular blocks or by columns. The same conventions for the use of the **Shift** and **Control** keys apply. Additionally, double-clicking on any one cell under the **Fixed** column also toggles the cell state: **fixed** to **free** or **free** to **fixed**.

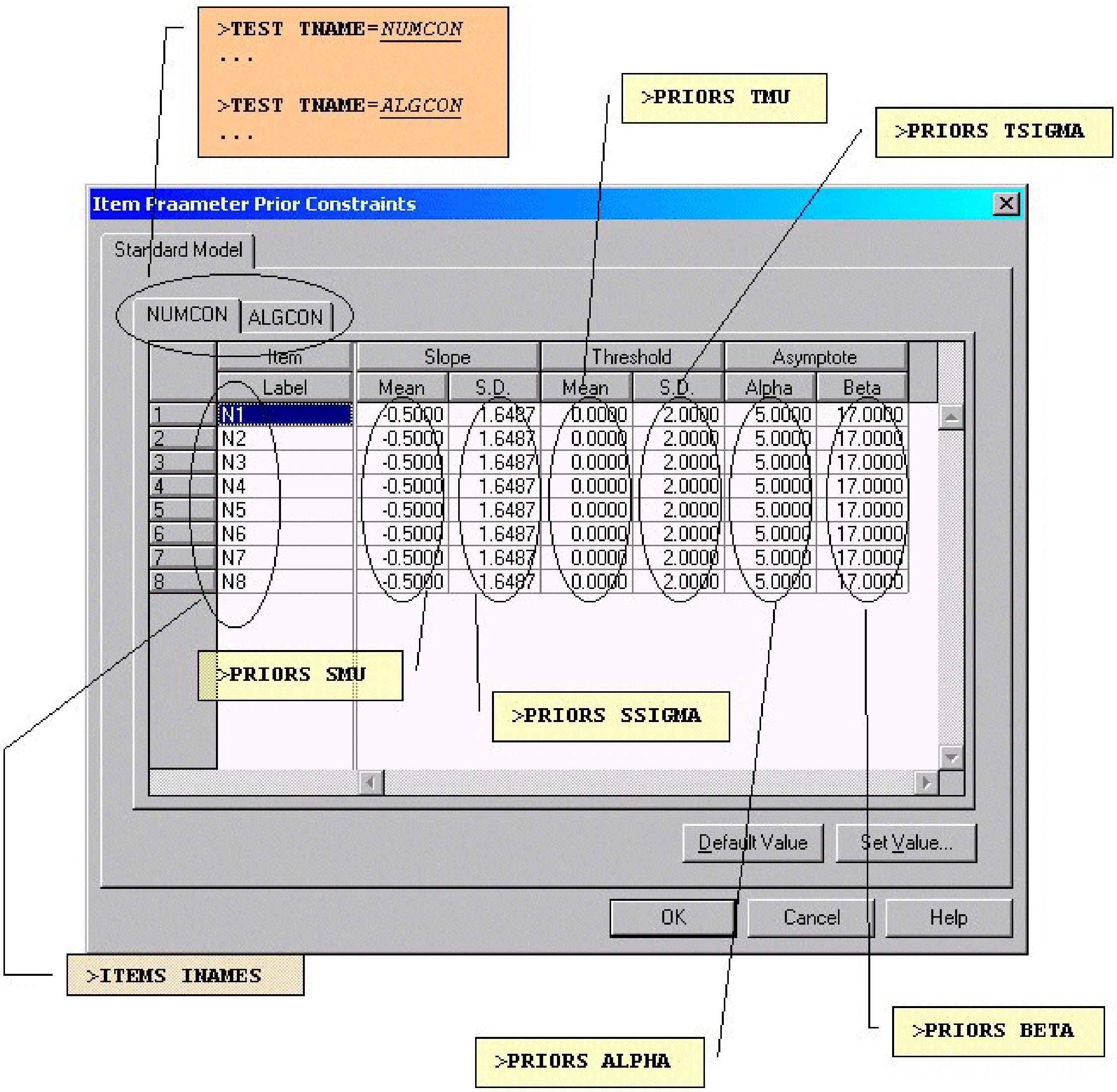


Related topics

- TEST command: FIX keyword (see Section 3.2.15)

Technical menu: Item Parameter Prior Constraints dialog box

The **Item Parameter Prior Constraints** dialog box is associated with the PRIORS command. The number of tabs on the dialog box depends on the number of subtests – priors may be entered for each subtest separately.



The user can set values by selecting an item or group of items and clicking the **Set Value** button. A subset of cells in the displayed table may be selected by holding the mouse button down and dragging until the selection is complete. Clicking the right mouse button will display a pop-up menu, which can be used to assign values to the cells. All selected cells may be set to a specific value or to the default value.

In addition, the user may select one parameter and assign a value to this and all other parameters below it by selection the appropriate option from the pop-up menu. A dialog box appears, prompting the user to enter the value to be assigned.

Alternatively, the **Default Value** or **Set Value** buttons may be used to assign values to the selected parameters. To set the priors of a selection of items to their default value, the **Default Value** button may be used. Links between the fields on this dialog box and the corresponding keywords on the PRIOR command are shown on the image below.

Related topics

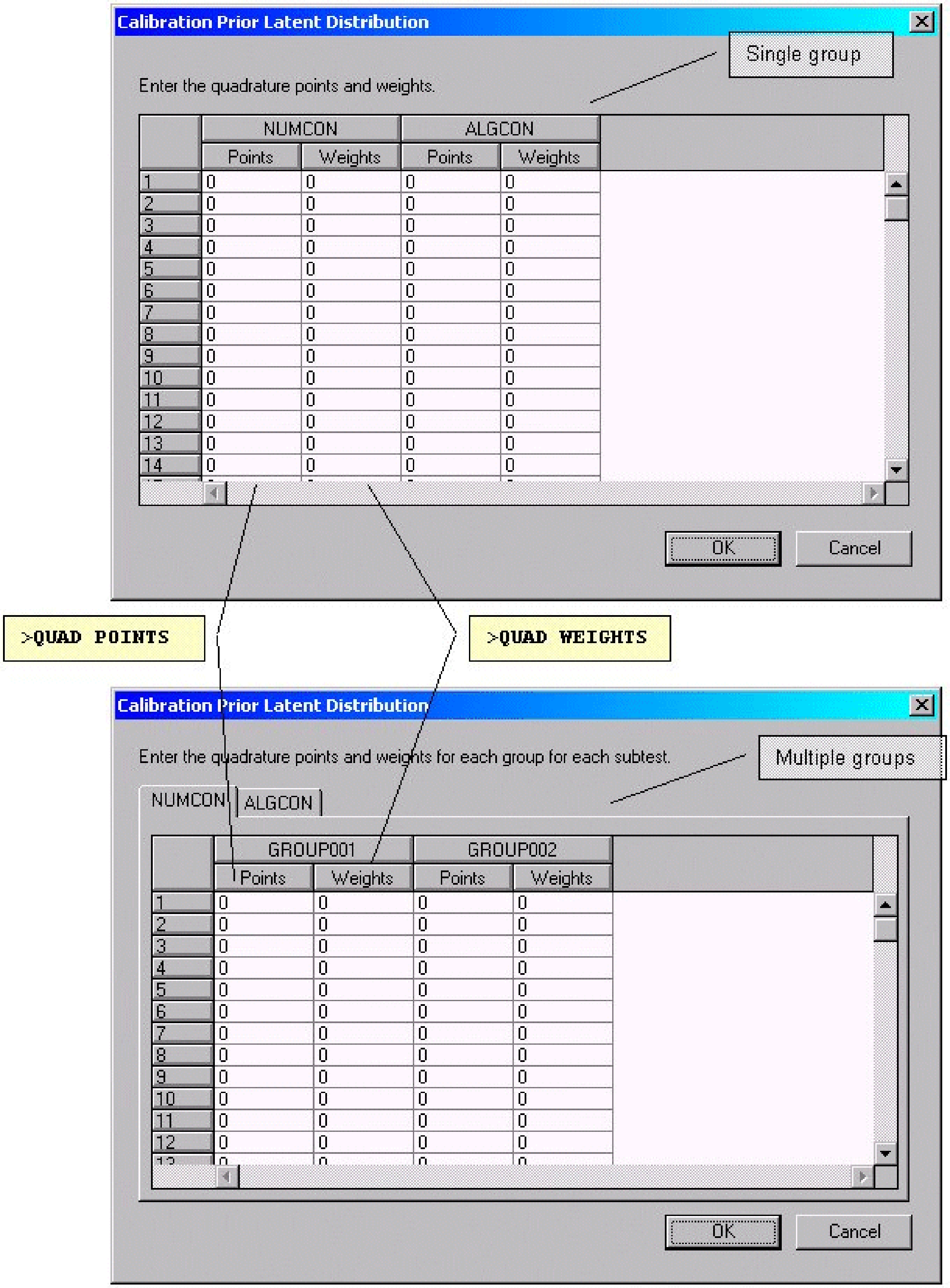
- CALIB command: READPRIOR option (see Section 0)
- PRIORS command: ALPHA, BETA, SMU, SSIGMA, TMU, and TSIGMA keywords (see Section 3.2.10)

Technical menu: Calibration Prior Latent Distribution dialog box

The **Assign Calibration Prior Latent Distribution** option provides the opportunity to assign prior latent distributions, by subtest, to be used during item parameter estimation. This dialog box is associated with the QUAD command(s). This option is only enabled when the IDIST keyword is set to 1 or 2 on the CALIB command.

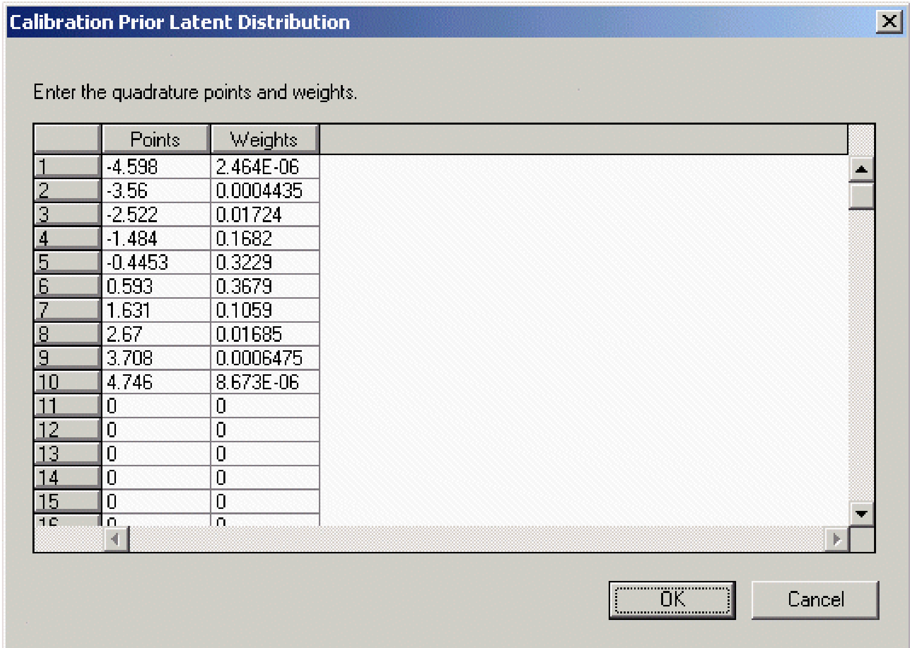
There is no user's interface to select this option. It must be set manually in the command file. For assigning prior latent distributions to be used during scoring, see the **Assign Scoring Prior Latent Distribution** dialog box.

The first image below shows the dialog box for a single group analysis. Quadrature points and weights may be provided separately for each subtest. On the second image, the **Calibration Prior Latent Distribution** dialog box for a multiple-group analysis is shown. Note that quadrature points and weights may be entered per group and subtest, as a tab for each subtest is provided in this case, and that the set of positive fractions entered as **Weights** should sum to 1.0.



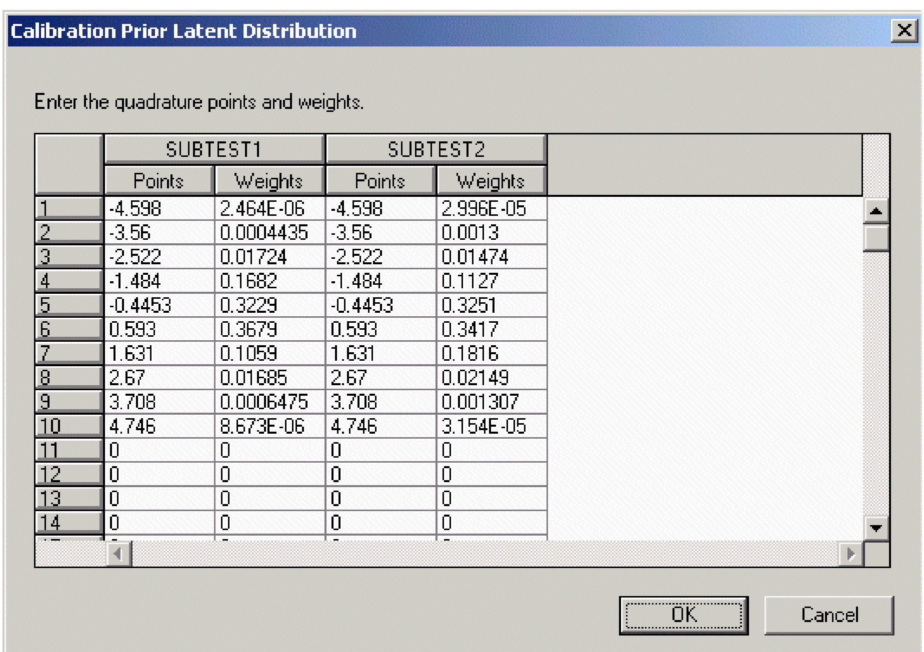
The format of the table on the **Calibration Prior Latent Distribution** dialog box depends on the values of the NTEST, NGROUP and IDIST keywords. Examples are shown below.

```
>GLOBAL NTEST=1, …
>INPUT NGROUP=1, …
>CALIB IDIST=1 or 2, …
```



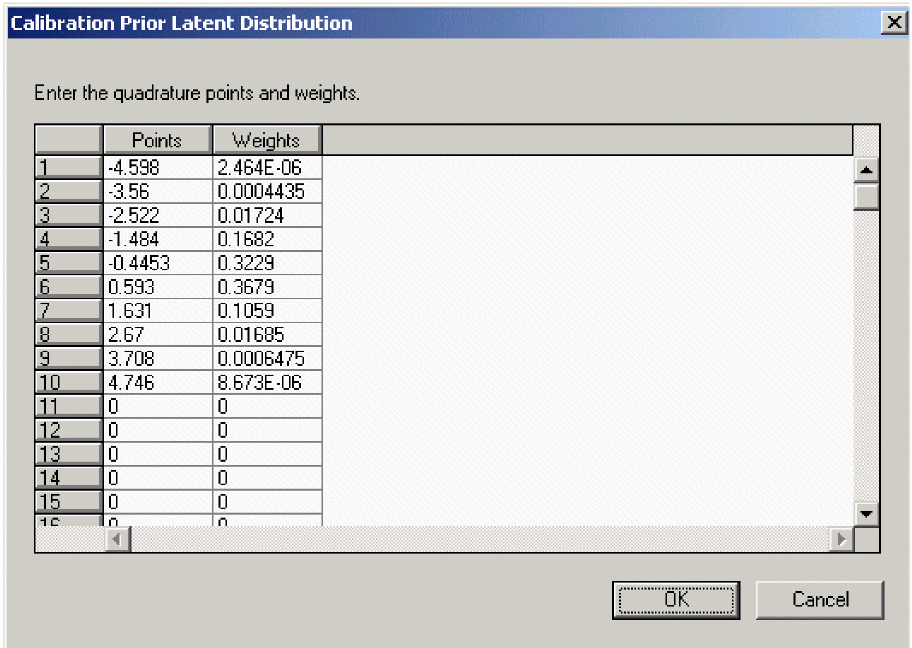
	Points	Weights
1	-4.598	2.464E-06
2	-3.56	0.0004435
3	-2.522	0.01724
4	-1.484	0.1682
5	-0.4453	0.3229
6	0.593	0.3679
7	1.631	0.1059
8	2.67	0.01685
9	3.708	0.0006475
10	4.746	8.673E-06
11	0	0
12	0	0
13	0	0
14	0	0
15	0	0

```
>GLOBAL NTEST>1, …
>INPUT NGROUP=1, …
>CALIB IDIST=1, …
```

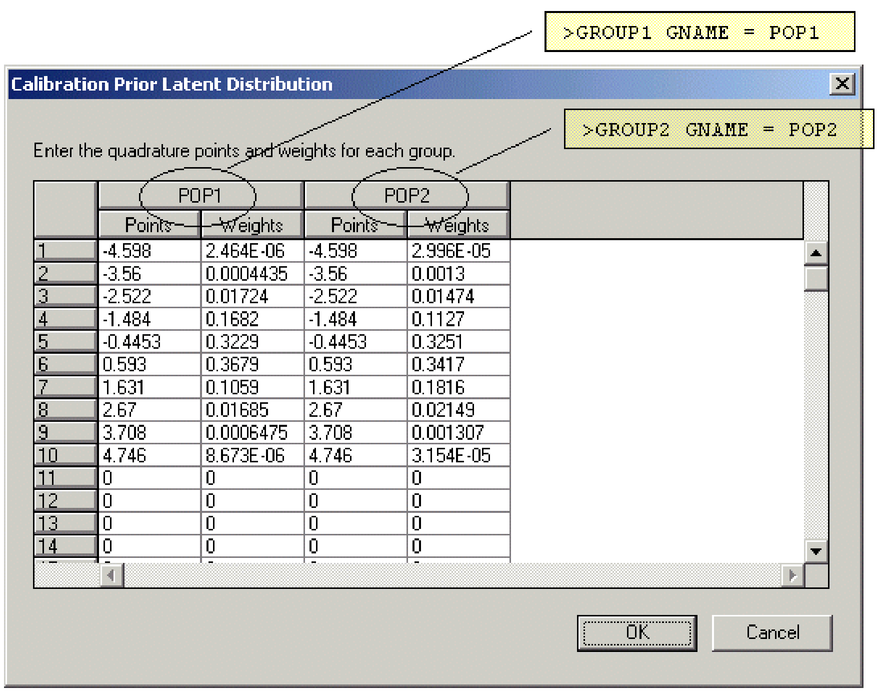


	SUBTEST1		SUBTEST2	
	Points	Weights	Points	Weights
1	-4.598	2.464E-06	-4.598	2.996E-05
2	-3.56	0.0004435	-3.56	0.0013
3	-2.522	0.01724	-2.522	0.01474
4	-1.484	0.1682	-1.484	0.1127
5	-0.4453	0.3229	-0.4453	0.3251
6	0.593	0.3679	0.593	0.3417
7	1.631	0.1059	1.631	0.1816
8	2.67	0.01685	2.67	0.02149
9	3.708	0.0006475	3.708	0.001307
10	4.746	8.673E-06	4.746	3.154E-05
11	0	0	0	0
12	0	0	0	0
13	0	0	0	0
14	0	0	0	0

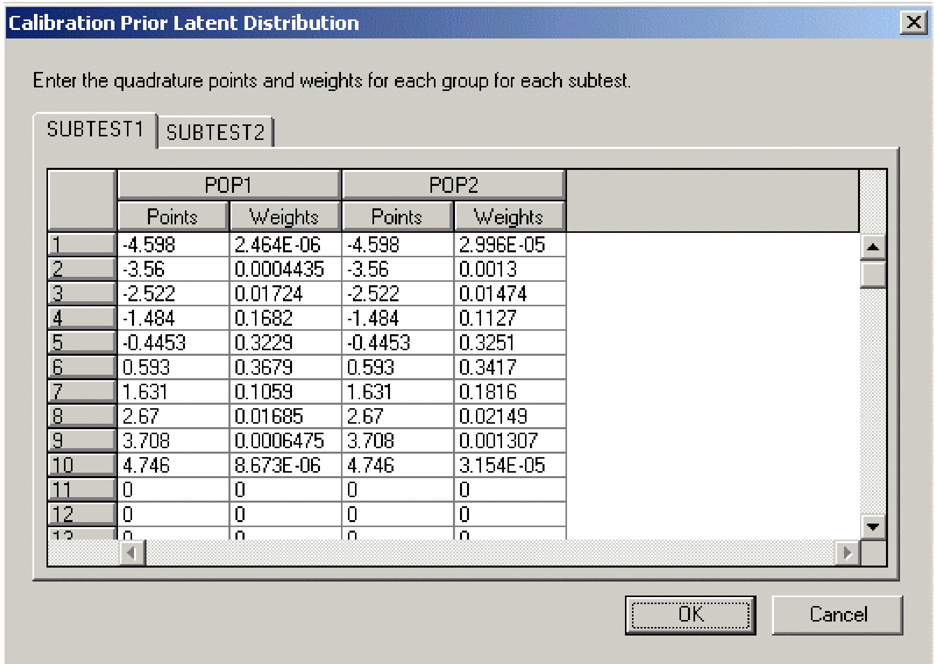
```
>GLOBAL NTEST>1, …
>INPUT NGROUP=1, …
>CALIB IDIST=2, …
```




```
>GLOBAL NTEST=1, …
>INPUT NGROUP>1, …
>CALIB IDIST=1 or 2, …
```



```
>GLOBAL NTEST>1, …
>INPUT NGROUP>1, …
>CALIB IDIST=1, …
```



Calibration Prior Latent Distribution

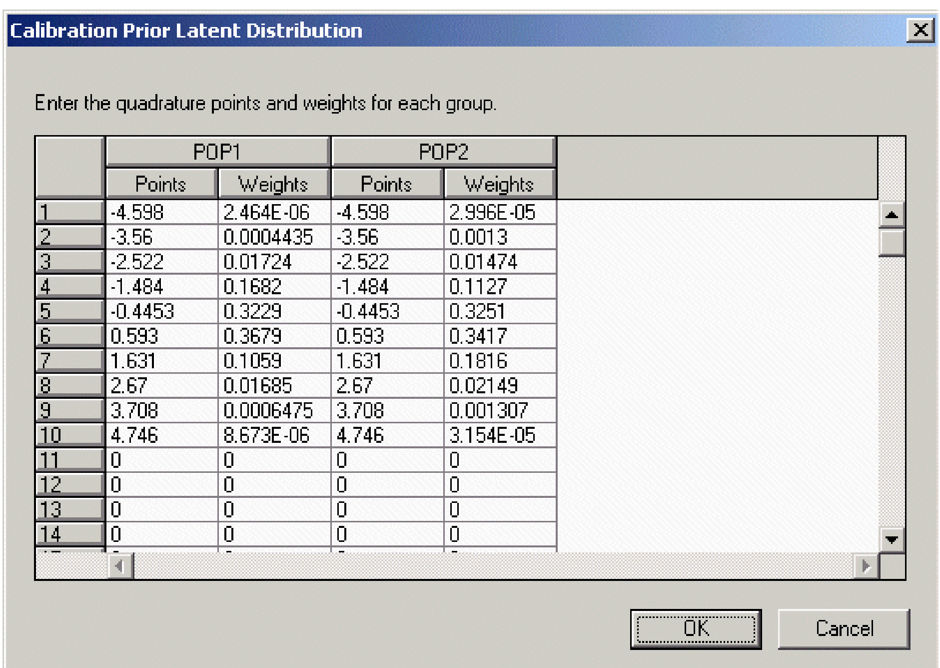
Enter the quadrature points and weights for each group for each subtest.

SUBTEST1 | SUBTEST2

	POP1		POP2	
	Points	Weights	Points	Weights
1	-4.598	2.464E-06	-4.598	2.996E-05
2	-3.56	0.0004435	-3.56	0.0013
3	-2.522	0.01724	-2.522	0.01474
4	-1.484	0.1682	-1.484	0.1127
5	-0.4453	0.3229	-0.4453	0.3251
6	0.593	0.3679	0.593	0.3417
7	1.631	0.1059	1.631	0.1816
8	2.67	0.01685	2.67	0.02149
9	3.708	0.0006475	3.708	0.001307
10	4.746	8.673E-06	4.746	3.154E-05
11	0	0	0	0
12	0	0	0	0

OK | Cancel

```
>GLOBAL NTEST>1, …
>INPUT NGROUP>1, …
>CALIB IDIST=2, …
```



Calibration Prior Latent Distribution

Enter the quadrature points and weights for each group.

	POP1		POP2	
	Points	Weights	Points	Weights
1	-4.598	2.464E-06	-4.598	2.996E-05
2	-3.56	0.0004435	-3.56	0.0013
3	-2.522	0.01724	-2.522	0.01474
4	-1.484	0.1682	-1.484	0.1127
5	-0.4453	0.3229	-0.4453	0.3251
6	0.593	0.3679	0.593	0.3417
7	1.631	0.1059	1.631	0.1816
8	2.67	0.01685	2.67	0.02149
9	3.708	0.0006475	3.708	0.001307
10	4.746	8.673E-06	4.746	3.154E-05
11	0	0	0	0
12	0	0	0	0
13	0	0	0	0
14	0	0	0	0

OK | Cancel

Related topics

- QUAD command: POINTS and WEIGHTS keywords (see Section 3.2.11)

Technical menu: Scoring Prior Latent Distribution dialog box

The **Assign Scoring Prior Latent Distribution** dialog box provides the opportunity to assign prior latent distributions, by subtest, to be used during scoring. This dialog box is associated with keywords on the SCORE and QUAD commands.

For assigning prior latent distributions to be used during the item parameter estimation phase, see the **Assign Calibration Prior Latent Distribution** dialog box.

On the **Normal** tab of this dialog box, the type of prior distribution to be used for the scale scores is the first information required. This tab is used when separate arbitrary discrete prior for each group or for each group for each subtest are to be read from a QUAD command. These options are only available when the Expected A Posteriori (EAP) method of scale score estimation is used.

When maximum likelihood (ML) or Maximum A Posteriori (MAP) estimation is selected, these options are disabled and the PMN and PSD keywords may be used to specify real-numbered values for the means and standard deviation of the normal prior distributions. The default values of these keywords for each group for each subtest, 0 and 1 respectively, are displayed.

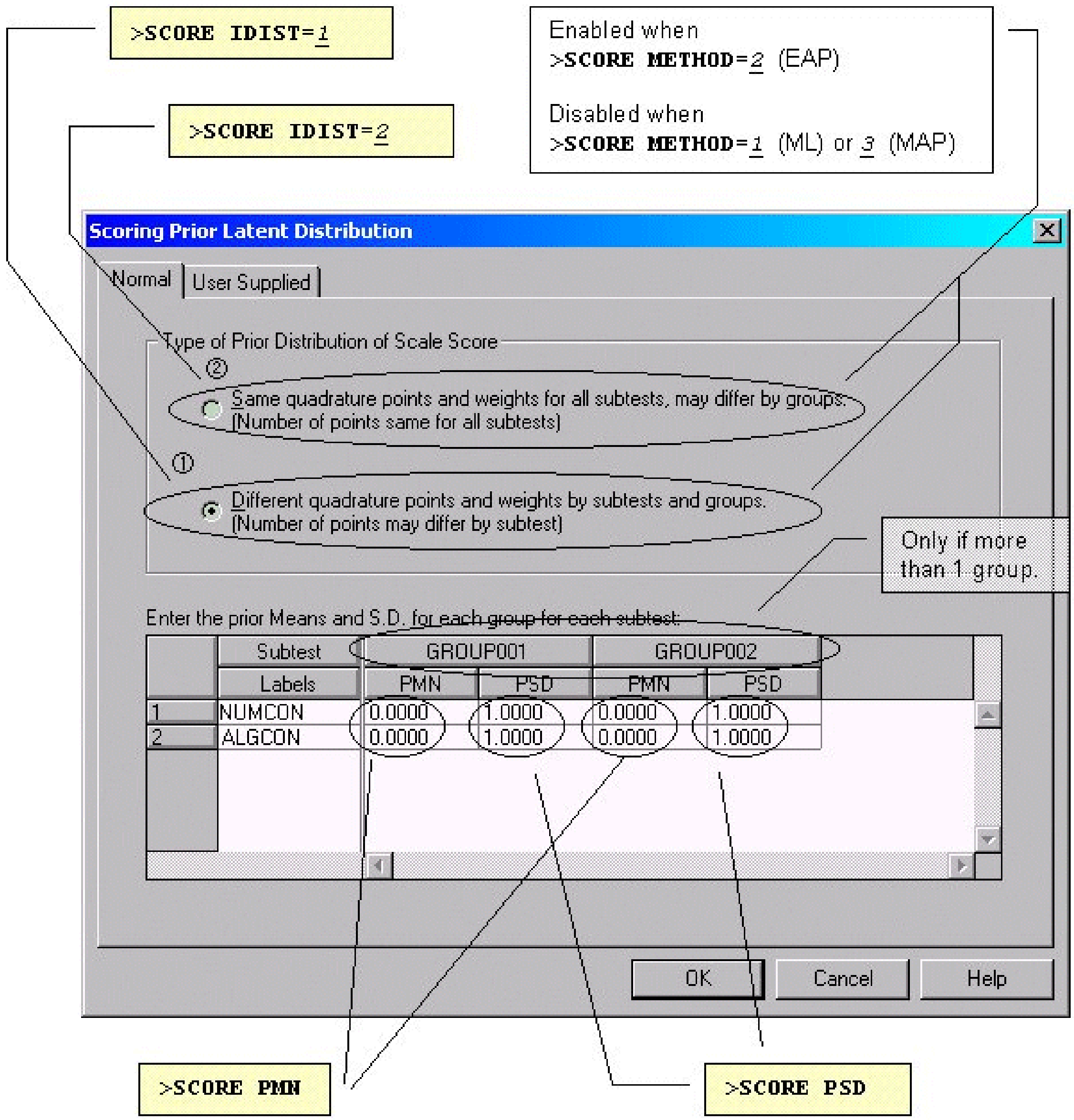
To provide alternative values for the PMN and PSD keywords, click in the fields and enter the new values.

Information in the table below corresponds to numbers on the image shown overleaf.

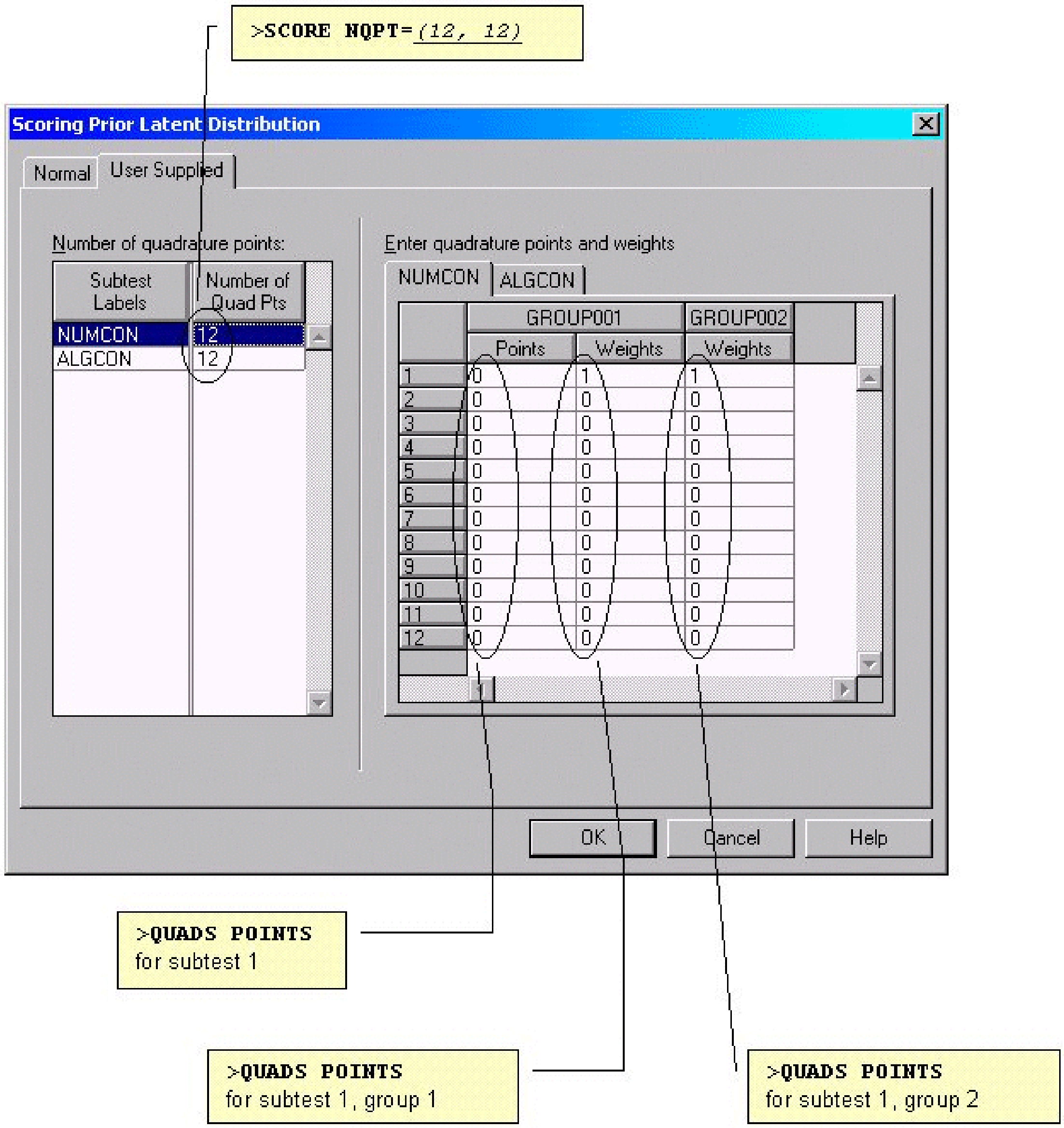
①	Set the number of entries in the >QUAD commands: POINTS， WEIGHTS to NGROUP * NTEST.
②	Set the number of entries in the >QUAD commands: POINTS, WEIGHTS to NGROUP.

Related topics

- SCORE command: IDIST, PMN and PSD keywords (see Section 3.2.14)



The **User Supplied** tab allows the user to change the number of quadrature points to be used by subtest. Different quadrature points and weights may be supplied for each group per subtest, as shown in the image below where two subtests were administered to two groups of examinees.

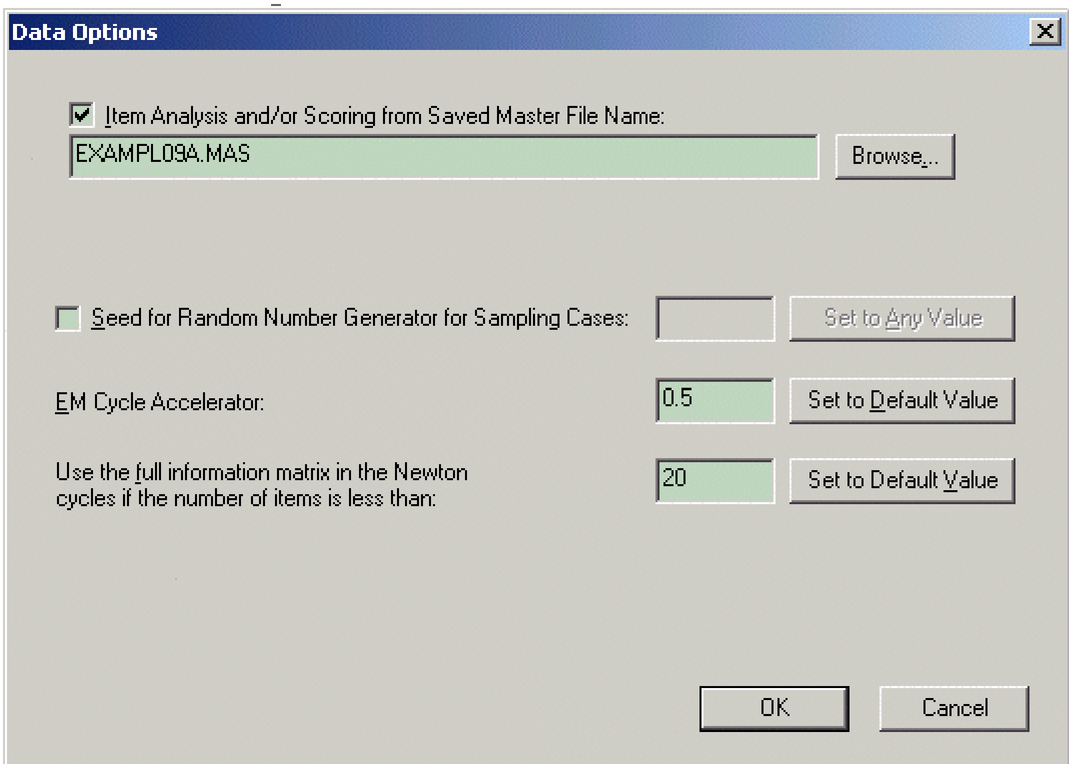


Related topics

- QUAD command: POINTS and WEIGHTS keywords (see Section 3.2.11)
- SCORE command: NQPT keyword (see Section 3.2.14)

Technical menu: Data Options dialog box

To set the values for the random number generator seed used with the SAMPLE keyword on the INPUT command, or to change the value of the acceleration constant used during the E-steps in item calibration, the **Data Options** dialog box may be used. To use default values, the **Set to Default Value** buttons may be clicked after which the program defaults will be displayed in the corresponding fields.



Note that:

- The **Item Analysis and/or Scoring from Saved Master File Name** section is the same as the **Master Data** edit box in the **Save Output to File** dialog box .
- The dialog box does not read any data from the specified master file. The filename is simply copied to the MASTER keyword on the SAVE command.

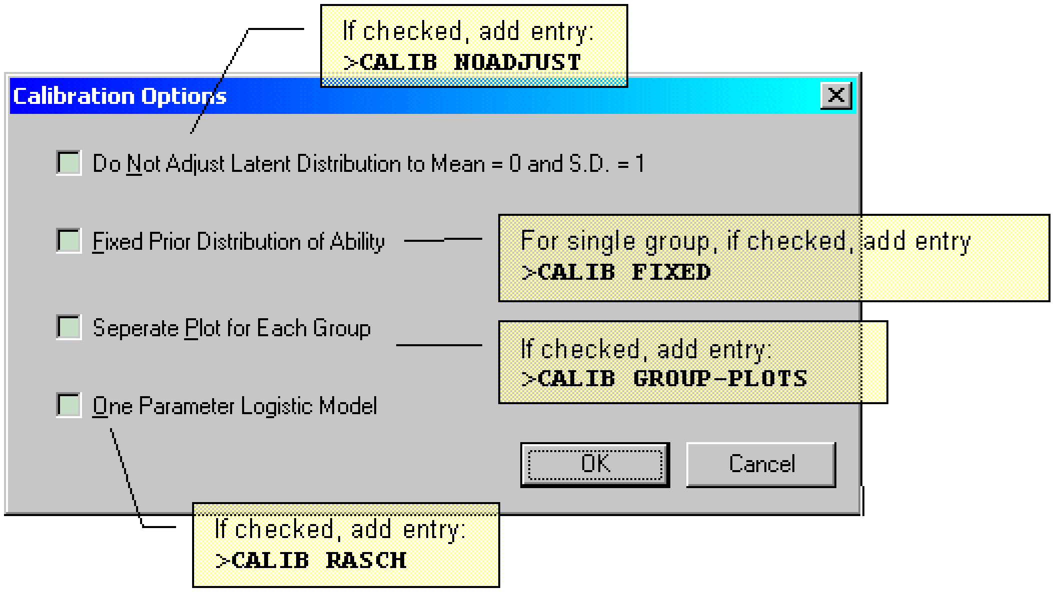
Related topics

- CALIB command: ACCEL and NFULL keywords (see Section 3.2.2)
- INPUT command: ISEED and SAMPLE keywords (see Section 3.2.7)

Technical menu: Calibration Options dialog box

The **Calibration Options** dialog box is associated with keywords on the CALIB command.

To request a Rasch model, the **One Parameter Logistic Model** option should be checked. Separate item plots for each group may be requested using the **Separate Plot for Each Group** check box while adjustment of the latent distribution to a mean of 0 and S.D. of 1 may be suppressed using the first check box. To keep the prior distributions of ability in the population of respondents fixed at the value specified in the IDIST keyword and/or the QUAD commands, the **Fixed Prior Distribution of Ability** check box should be checked. This corresponds to the FIXED option on the CALIB command.



Related topics

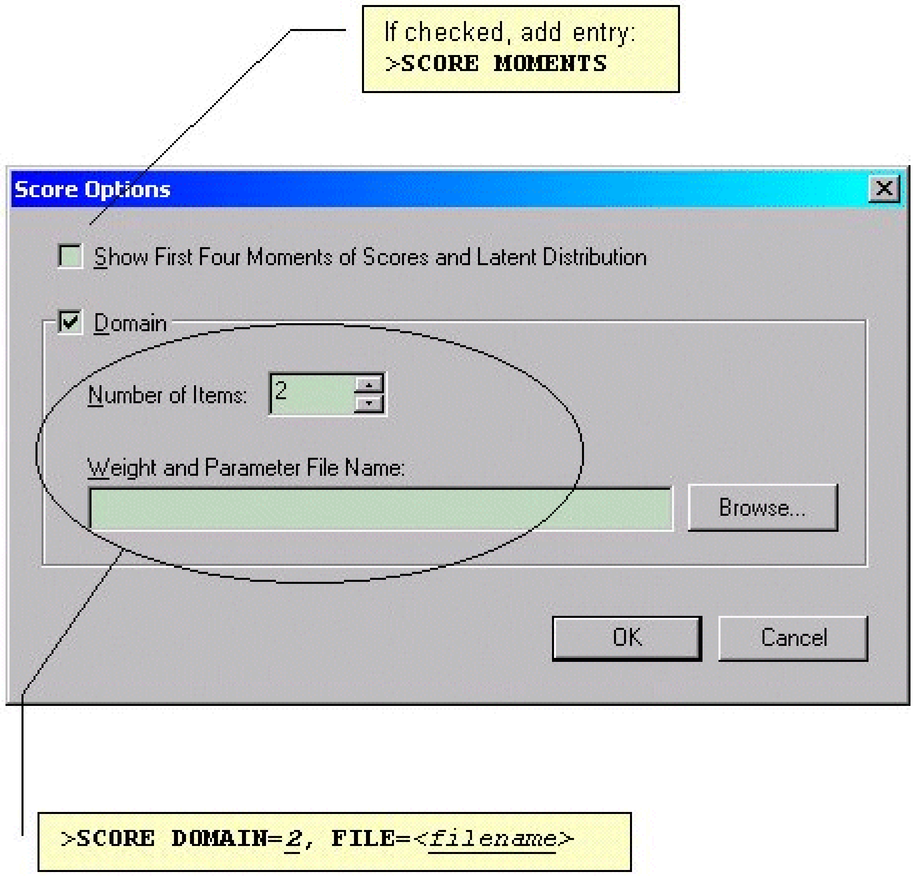
- CALIB command: FIXED, GROUP-PLOTS, NOADJUST and RASCH options (see Section 3.2.2)

Technical menu: Score Options dialog box

This dialog box allows the user to request the calculation of domain scores based on a user-supplied file containing the item parameters for a sample of previously calibrated items for a domain and to request the computation and listing of the coefficients of skewness and kurtosis of the ability estimates and of the latent distribution.

Related topics

- SCORE command: DOMAIN and FILE keywords (see Section 3.2.14)
- SCORE command: MOMENTS option

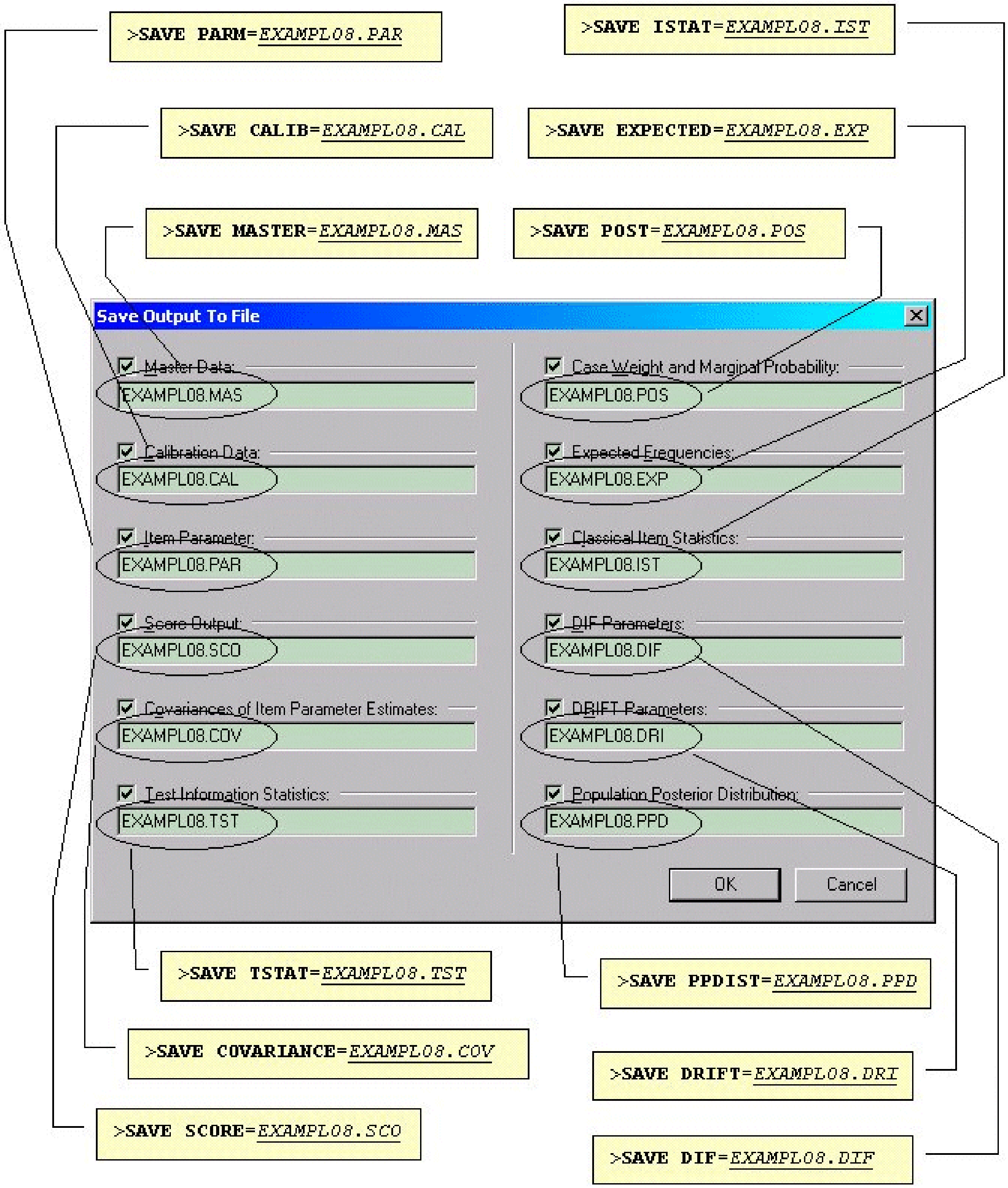


2.6 Save menu

The **Save Output to File** dialog box is accessed through the **Save** menu. Various types of data may be saved to external files using the SAVE command. On the image below, links are provided between the fields of this dialog box and the corresponding keywords on the SAVE command.

Related topics

- GLOBAL command: SAVE option (see Section 3.2.5)
- SAVE command



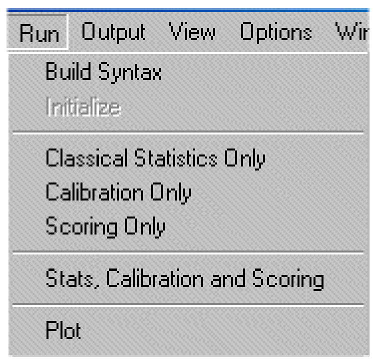
2.7 Run menu

The **Run** menu provides the necessary options to generate syntax from the input provided on the dialog boxes accessed through the **Setup**, **Data**, and **Technical** menus (**Build Syntax** option) or to run separate or all phases of the analysis. This menu is also used to access the graphics procedure described in Chapter 4 via the **Plot** option. Note that this option is only enabled after completion of the three phases of analysis.

Select the **Build Syntax** option to generate a syntax or command file based on the contents of the previous dialog boxes and menus. When the **Initialize** option is selected, changes made to an existing command file in the syntax window are transferred to the dialog boxes and menus.

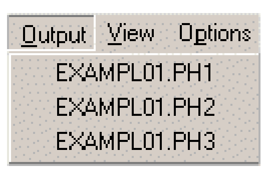
Run only the first phase of the analysis to obtain the classical statistics by selecting the **Classical Statistics Only** option. The item parameter estimation may be performed next by selecting the **Calibration Only** option, and scoring after that using the **Scoring Only** option. These options are provided to allow the user to run and verify information in the output for each phase of the analysis before continuing to the next step. When running the analysis phase by phase, the option to run the next phase will only be enabled after successful completion of the previous phase.

Alternatively, the user can request to run all three phases in succession by selecting the **Stats, Calibration and Scoring** option. A message indicating the normal or abnormal termination of each phase will appear in the main window between phases to alert the user to possible problems in a particular phase of the analysis. This message may be suppressed using the **Options** menu.



2.8 Output menu

To view the output obtained during any of the three phases of analysis, the options on the **Output** menu may be used. Options will be enabled or disabled depending on the number of completed phases of the analysis. When any of these options is selected, the relevant output file will be displayed. After inspection, the user may close this file to return to the main BILOG-MG window, where the command file on which the analysis was based will be displayed.



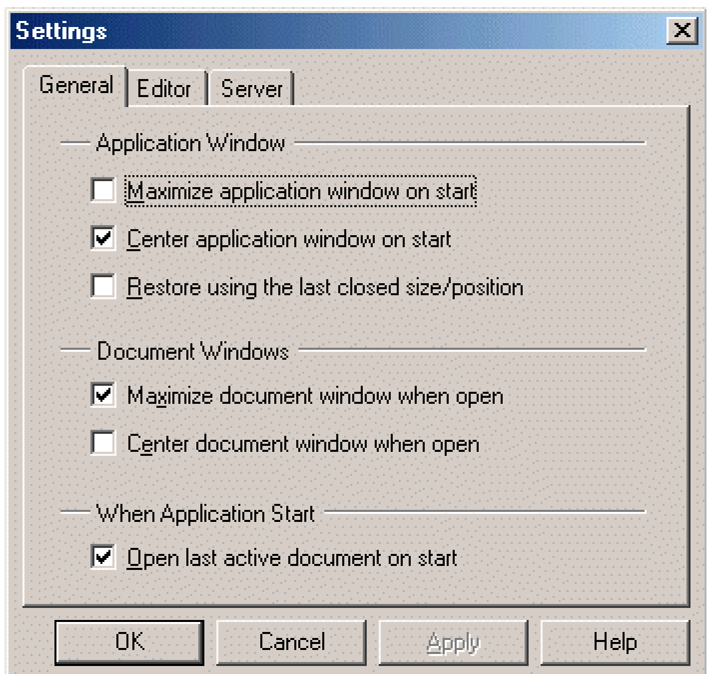
2.9 View menu

The **View** menu allows the user to add or remove the status bar displayed at the bottom of the main BILOG-MG window. The toolbar, allowing the standard Windows editing functions, is displayed by default.

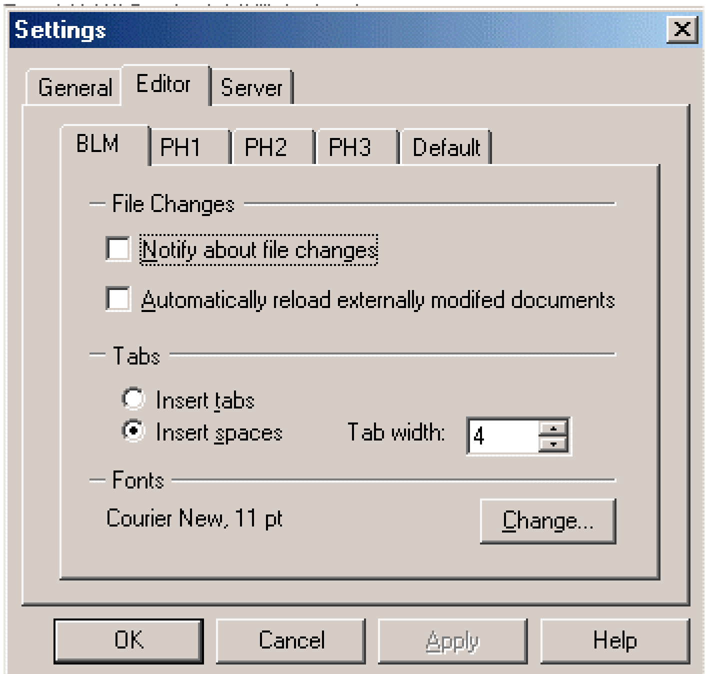
2.10 Options menu

The **Options** menu provides access to the **Settings** dialog box. This dialog box has three tabs: **General**, **Editor,** and **Server**.

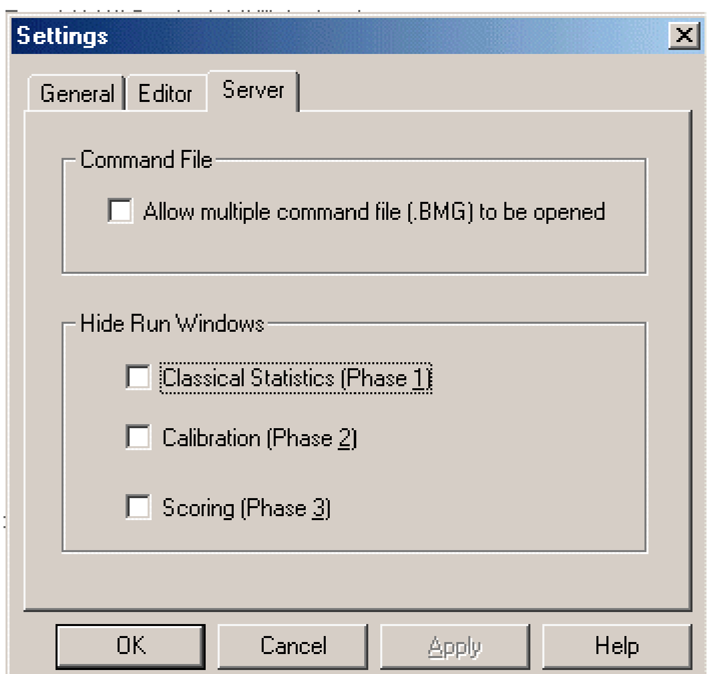
On the **General** tab as shown below, the size of the application window and document window can be set. The user may opt to always open the last active document when opening BILOG-MG (default) or to start with a blank screen instead by unchecking the **Open last active document on start** check box.



To change the font in which the contents of the editor window are displayed, or to use tabs, the **Editor** tab of the **Settings** dialog box may be used. Reminders of file changes and automatic re-loading of externally modified documents may also be requested.

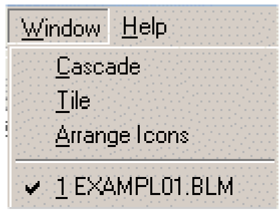


The **Server** tab of the **Settings** dialog box may be used to show or hide the windows in which details of the analysis are displayed during the run. To open multiple command files, which can then be accessed using the **Windows** menu, check the box next to the **Allow multiple command file…** option on this tab.



2.11 Window menu

The **Window** menu allows the user to arrange multiple windows or to switch between open files. To open multiple command files simultaneously that may be accessed through this menu, use the **Server** tab on the **Settings** dialog box accessed through the **Options** menu.



2.12 Help menu

The **Help** menu provides access to the BILOG-MG help file (**Help Topics** option) and to the **About BILOG-MG for Windows** dialog box in which the version and build number of the application are displayed. This box may also be used to directly e-mail SSI for technical support or product information or to link to the SSI website.

2.13 Location of keywords in interface

Command and keyword	**Menu and option**	**Tab on dialog box**
TITLE command	Setup, General	Job Description
COMMENT command	Setup, General	Job Description
GLOBAL command:		
`DFNAME`	Data, Examinee Data/Data, Group-Level Data	Data File/Enter Data
`MFNAME`	-	-
`CFNAME`	-	-
`IFNAME`	Setup, Test Scoring	General
`NPARM`	Setup, General	Model
`NWGHT`	-	-
`NTEST`	Setup, General	Job Description
`NVTEST`		
`LOGISTIC`	Setup, General	Model

Command and keyword	**Menu and option**	**Tab on dialog box**
OMITS	Setup, General	Response
SAVE	Save	-
PRNAME	Setup, Test Scoring	General
SAVE command:		
MASTER	Save	-
CALIB	Save	-
PARM	Save	-
SCORE	Save	-
COVARIANCE	Save	-
TSTAT	Save	-
POST	Save	-
EXPECTED	Save	-
ISTAT	Save	-
DIF	Save	-
DRIFT	Save	-
PDISTRIB	Save	-
LENGTH command:		
NITEMS	Setup, Item Analysis	Subtests
NVARIANT	Setup, Item Analysis	Subtests
INPUT command:		
NTOTAL	Setup, General	Job Description
NFMT	Data, Examinee Data/Data, Group-Level Data	Data File
TYPE	Data, Examinee Data/Data, Group-Level Data	General
SAMPLE	Data, Examinee Data/Data, Group-Level Data	General
NALT	Setup, General	Response

Command and keyword	**Menu and option**	**Tab on dialog box**
NIDCHAR	Data, Examinee Data/Data, Group-Level Data	General
TAKE	Data, Examinee Data/Data, Group-Level Data	General
NGROUP	Setup, General	Job Description
NFORM	Setup, General	Job Description
DIAGNOSE		
KFNAME	Data, Item Keys	Answer Key
NFNAME	Data, Item Keys	Not Presented Key
OFNAME	Data, Item Keys	Omit Key
DRIFT	Setup, General	Model
DIF	Setup, General	Model
PERSONAL	Data, Examinee Data	General
EXTERNAL	Data, Examinee Data/Data, Group-Level Data	General
ISEED	Technical, Data Options	-
ITEMS command:		
INUMBERS	-	-
INAMES	Setup, General	Labels
TEST command:		
TNAME	Setup, General	Labels
INUMBERS	Setup, Item Analysis	Subtest Items
INAMES	Setup, General	Labels
INTERCPT	Technical, Assign Item Parameter Starting Values	Import/Enter Values
SLOPE	Technical, Assign Item Parameter Starting Values	Import/Enter Values
THRESHLD	Technical, Assign Item Parameter Starting Values	Import/Enter Values
GUESS	Technical, Assign Item Parameter Starting Values	Import/Enter Values

Command and keyword	Menu and option	Tab on dialog box
DISPERSN	Technical, Assign Item Parameter Starting Values	Import/Enter Values
FIX	Technical, Assign Fixed Items	-
FORM command:		
LENGTH	Setup, Item Analysis	Form Items
INUMBERS	Setup, Item Analysis	Form Items
INAMES	Setup, General	Labels
GROUP command:		
GNAME	Setup, General	Labels
LENGTH	Setup, Item Analysis	Group Items
INUMBERS	Setup, Item Analysis	Group Items
INAMES	Setup, General	Labels
DRIFT command:		
MAXPOWER	-	-
MIDPOINT	-	-
Variable format statement	Data, Examinee Data / Data, Group-Level Data	Data File
CALIB command:		
CHI	Setup, Item Analysis	Advanced
NQPT	Setup, Item Analysis	Advanced
CYCLES	Setup, Item Analysis	Advanced
NEWTON	Setup, Item Analysis	Advanced
PRINT	-	-
CRIT	Setup, Item Analysis	Advanced
IDIST	-	-
PLOT	-	-

Command and keyword	Menu and option	Tab on dialog box
DIAGNOSIS	-	-
REFERENCE	Setup, General	Job Description
SELECT	Setup, Item Analysis	Subtests
RIDGE	-	-
ACCEL	Technical, Data Options	-
NSD	-	-
COMMON	-	-
EMPIRICAL	Setup, Item Analysis	Advanced
NORMAL	-	-
FIXED	Technical, Calibration Options	-
TPRIOR	Setup, Item Analysis	Advanced
SPRIOR	Setup, Item Analysis	Advanced
GPRIOR	Setup, Item Analysis	Advanced
NOTPRIOR	Setup, Item Analysis	Advanced
NOSPRIOR	Setup, Item Analysis	Advanced
NOGPRIOR	Setup, Item Analysis	Advanced
READPRIOR	Technical menu: Item Parameter Prior Constraints dialog box	-
FLOAT	Setup, Item Analysis	Advanced
NOFLOAT	Setup, Item Analysis	Advanced
GROUP-PLOTS	Technical, Calibration Options	-
NOADJUST	Technical, Calibration Options	-
RASCH	Technical, Calibration Options	-
NFULL	Technical, Data Options	-

Command and keyword	**Menu and option**	**Tab on dialog box**
QUAD command:		
POINTS	Technical, Assign Calibration Prior Latent Distributions	-
WEIGHTS	Technical, Assign Calibration Prior Latent Distributions	-
PRIORS command:		
TMU	Technical, Assign Item Parameter Prior Constraints	-
TSIGMA	Technical, Assign Item Parameter Prior Constraints	-
SMU	Technical, Assign Item Parameter Prior Constraints	-
SSIGMA	Technical, Assign Item Parameter Prior Constraints	-
ALPHA	Technical, Assign Item Parameter Prior Constraints	-
BETA	Technical, Assign Item Parameter Prior Constraints	-
SCORE command:		
METHOD	Setup, Test Scoring	General
NQPT	Technical, Assign Scoring Prior Latent Distribution	User-supplied
IDIST (values=0,3)	Setup, Test Scoring	General
IDIST (values=1,2)	Technical, Assign Scoring Prior Latent Distribution	Normal
PMN	Technical, Assign Scoring Prior Latent Distribution	Normal
PSD	Technical, Assign Scoring Prior Latent Distribution	Normal
RSCTYPE	Setup, Test Scoring	Rescaling
LOCATION	Setup, Test Scoring	Rescaling
SCALE	Setup, Test Scoring	Rescaling
INFO	-	-

Command and keyword	Menu and option	Tab on dialog box
BIWEIGHT	Setup, Test Scoring	General
FIT	Setup, Test Scoring	General
NOPRINT	Setup, Test Scoring	General
YCOMMON	-	-
POP	-	-
REFERENCE	-	-
READF	-	-
NFORMS	-	-
MOMENTS	Technical, Score Options	-
DOMAIN	Technical, Score Options	-
FILE	Technical, Score Options	-
QUADS command:		
POINTS	Technical, Assign Scoring Prior Latent Distribution	User-Supplied
WEIGHTS	Technical, Assign Scoring Prior Latent Distribution	User-Supplied

Related topics

- Overview of required and optional commands.

2.14 Getting started with BILOG-MG

To illustrate the use of the interface in creating syntax files, the data file **exampl01.dat** in the **examples** subfolder of the BILOG-MG installation folder is used. This problem is based on an example in Thissen, Steinberg & Wainer (1993). Other examples based on the same data (see complete description below) can be found in Chapter 6.

In the late 1980s, R. Darrell Bock created a "College Level Spelling Test" comprising a sample of 100 words drawn from a large source list by simple random sampling. Data collected using that test are the basis for the empirical example in the paper "IRT Estimation of Domain Scores" (R.D. Bock, M.F. Zimowski, & D. Thissen, *Journal of Educational Measurement*, 1997, 34, 197-211). Parameter estimates for the 2PL IRT model for the 100-item test are tabulated in that paper. Bock created the script for conventional oral presentation of the test, and recorded the original reading of the script (by Monica Marie Bock) on reel-to-reel magnetic tape. Subsequent copies onto cassette tape were used by Jo Ann Mooney in the collection of data from around 1000 University of Kansas Undergraduates. We are using the file with 100 words (items) and 1000 records (examinees).

The words for the test were randomly selected from a popular wordbook for secretaries. Students were asked to write the words as used in a sentence on the tape recording. Responses were scored 1 if spelled correctly and 0 if spelled incorrectly. Because the items are scored 1,0, according to the defaults assumed by the program, an answer key is not required.

The purpose of this section is to give the new user a quick overview of the interface and the absolute minimum INPUT needed to run the program.

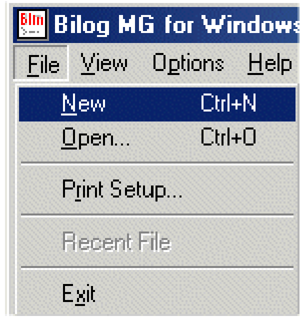
A few lines of the data in **exampl01.dat** are shown below:

```
 11 0000
 21 0001
 31 1000
 41 1001
…
162 1111
```

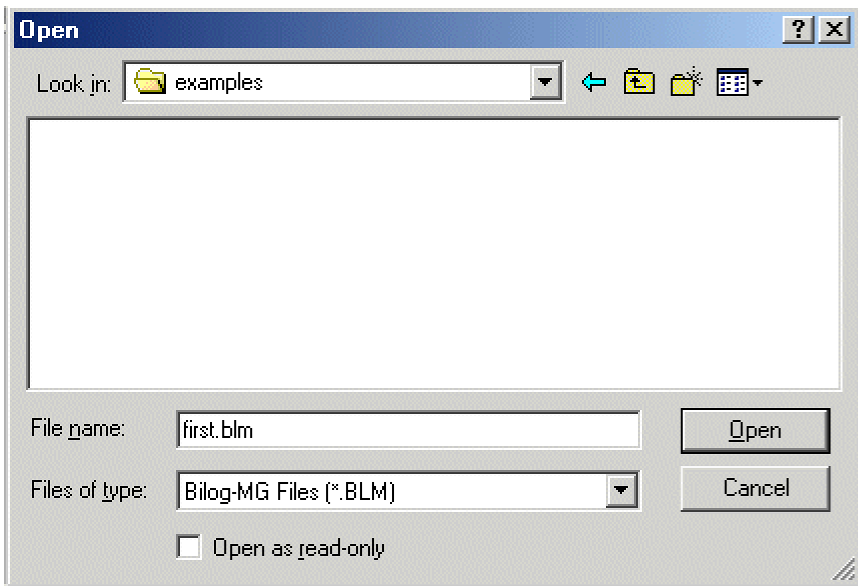
The first three characters in each line represent the examinee identification field. This is followed by the responses to the four items in the test.

2.14.1 A first model: 2PL model for spelling data

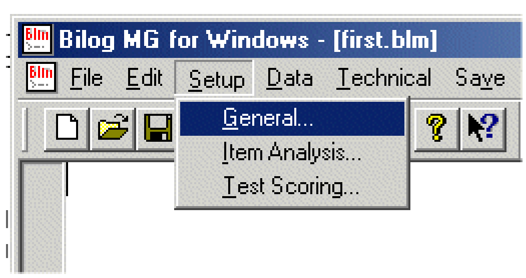
As a first example, we wish to set up a simple 2-PL model for this data. To construct a command file, begin by selecting the **New** option from the **File** menu. The **Open** dialog box is now activated.



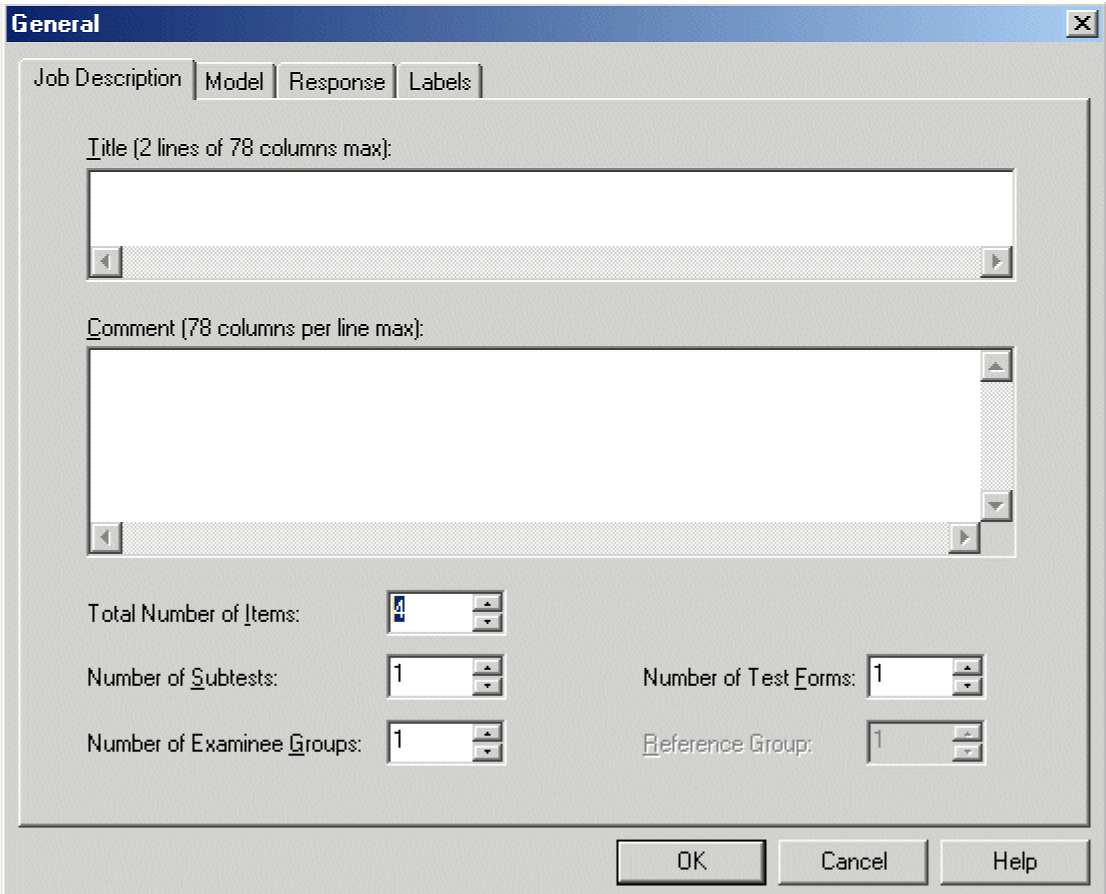
Assign a name, with the ***.blm** file extension, to the command file. In this case, the command file **first.blm** is created in the **examples** folder as shown below. Click **Open** when done to return to the main BILOG-MG window.



Note that a number of options have been added to the main menu bar of the BILOG-MG window. Of interest for this example are the **Setup**, **Data**, **Run** and **Output** options. The **Setup** menu is used to describe the model to be fitted to the data. As a first step, select the **General** option from this menu to access the **General** dialog box.



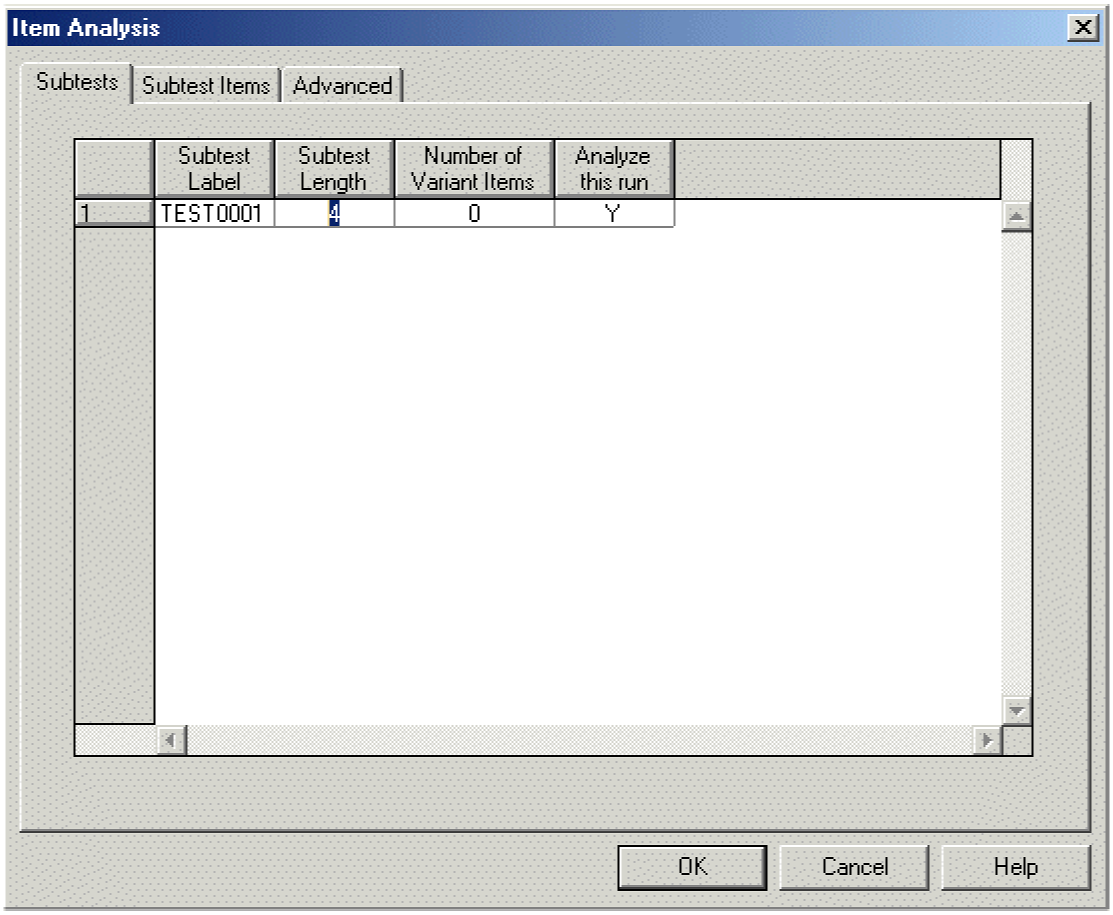
The **General** dialog box has four tabs, on which both required and optional keywords may be set. On the **Job Description** tab below, the number of items in the test is indicated as 4. The type of model is selected on the **Model** tab. As the default model fitted by BILOG-MG is a 2PL model, this tab is not used now. Click **OK** to return to the main window.



The next step in specifying the analysis is to assign the items to be calibrated to the test. To do this, select the **Item Analysis** option on the **Setup** menu to access the **Item Analysis** dialog box.

Change the default value of 1 under **Subtest Length** to 4 by clicking in this field and typing in "4". By default, all items will be analyzed, as indicated under the **Analyze this run** header. Click **OK** when done.

This completes the model specification. All that remains to be done is to provide information on the data. To do so, the **Data** menu is used. In this case, we have examinee data and thus the **Examinee Data** option is selected from the **Data** menu.

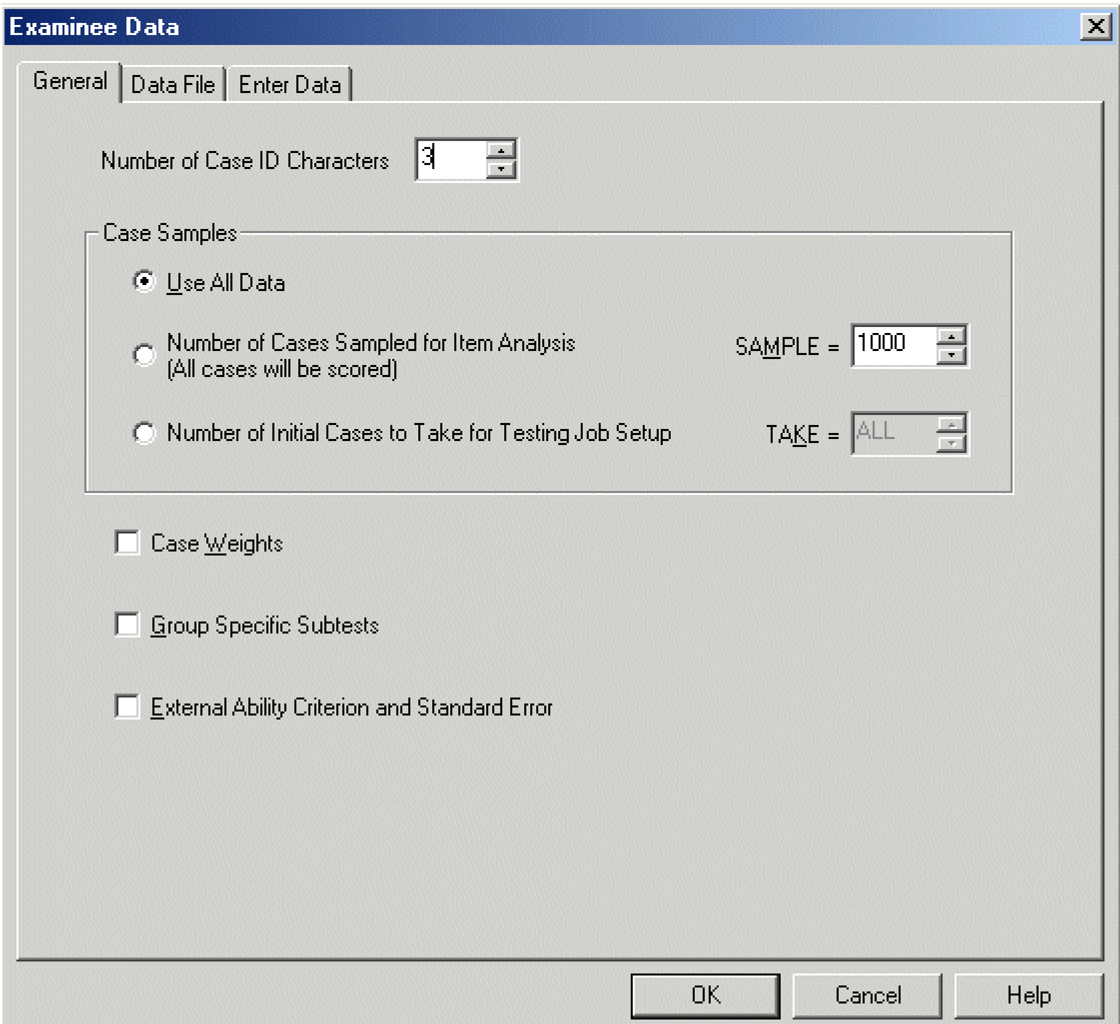


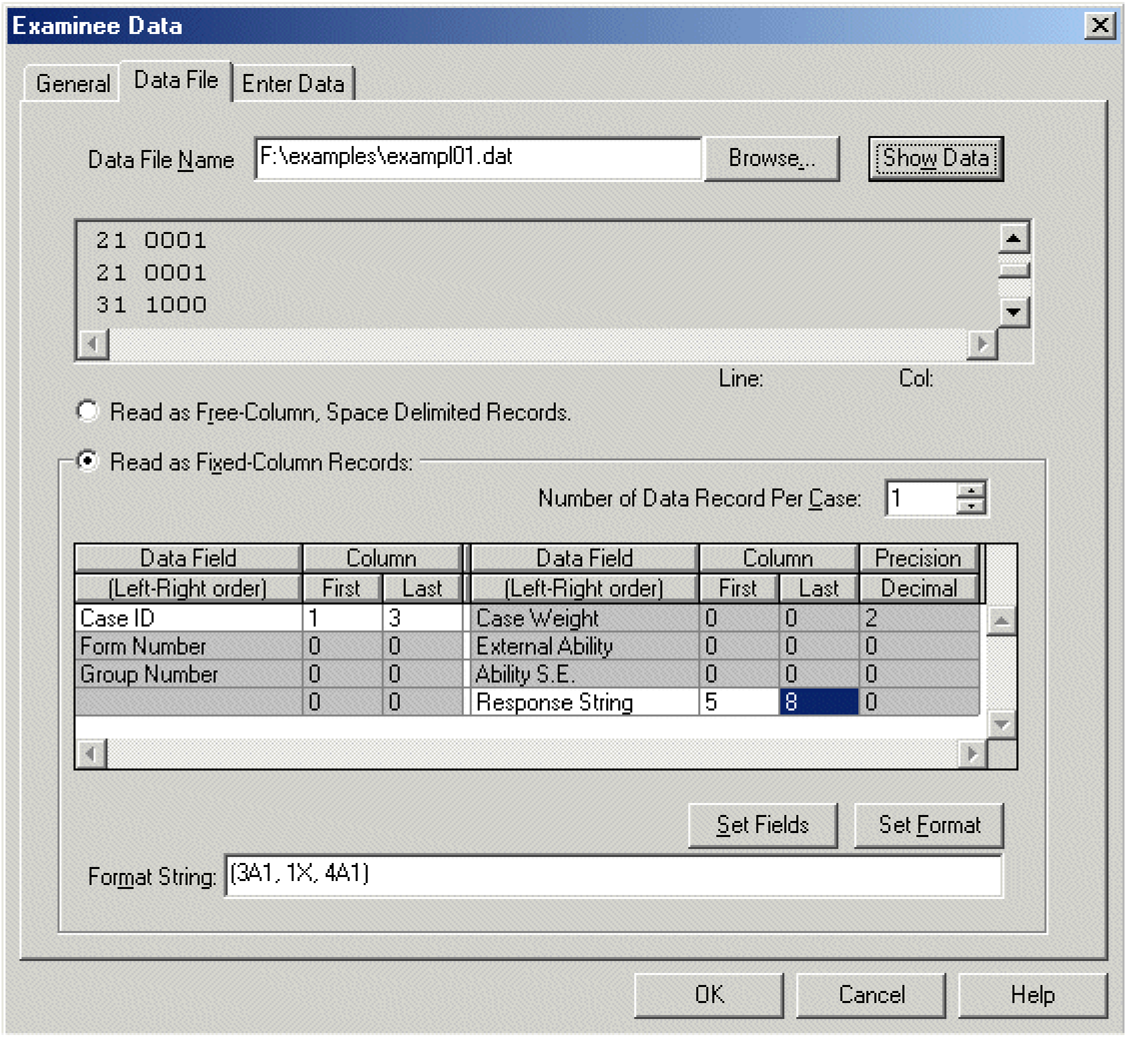
On the **Examinee Data** dialog box, enter the number of characters representing the examinee identification (in this case 3) in the **Number of Case ID Characters** field. By default, all data are used as shown below.

To provide information on the name and format of the data file, click the **Data File** tab.

- Use the **Browse** button to browse for the data file.
- Next, indicate that the data are in fixed format by clicking the **Read as Fixed-Column Records** radio button.
- Complete the table in the **Read as Fixed-Column Records** group box by clicking in the cell next to **Case ID** under the **First** header. Enter a "1" here to indicate that the examinee identification starts in column 1 of the data. Next, enter a "3" under the **Last** header to indicate the end of the examinee identification.
- Note that **Form Number**, **Group Number**, etc is disabled due to the information we entered from the **Setup** menu. The only other fields to complete are the **Response String** fields.
- The response to the first item is in column 5 of the data, so a "5" is entered in the cell next to **Response String** under the **First** header. The response to the last item is in column 8, and an "8" is thus entered in the **Last** column.
- By default, BILOG-MG assumes that there is one line of data for each examinee, as indicated by the **Number of Data Record per Case** field. As this is the case for data in **ex-**

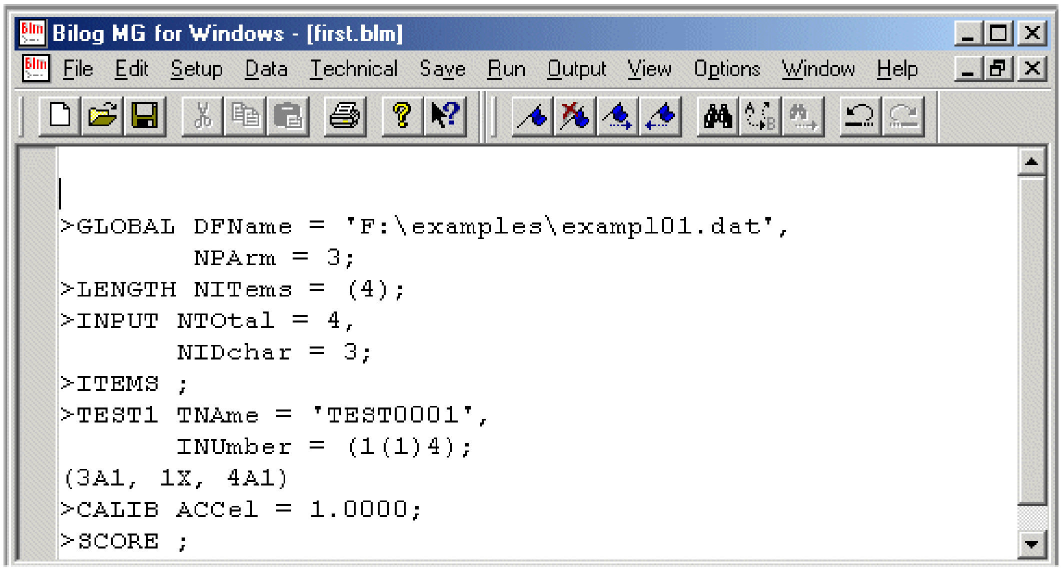
ampl01.dat, no further information is required. Click the **Set Format** button to write the format statement (3A1,1X,4A1) to the **Format String** field. Click **OK** to return to the main BILOG-MG window.



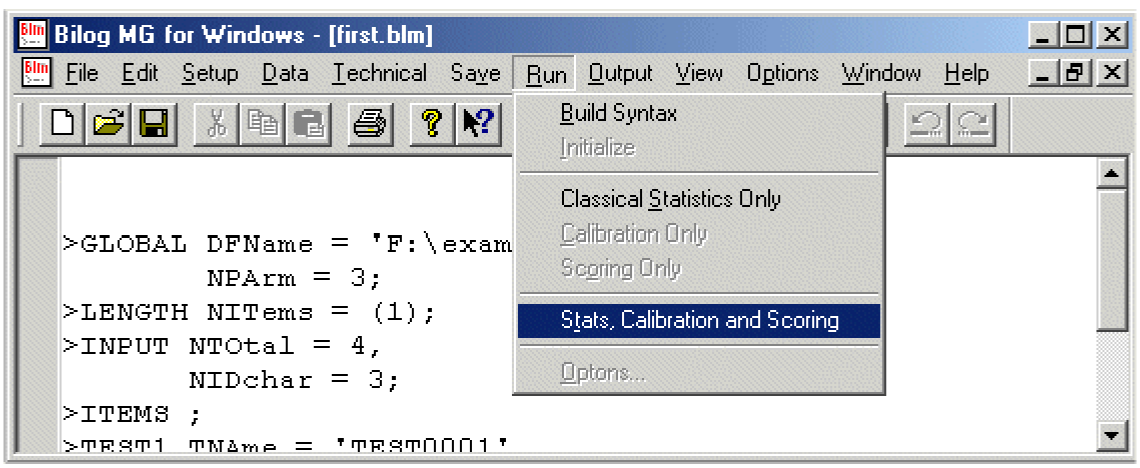


Having completed the specification in terms of model and data, the command file is created by selecting the **Build Syntax** option from the **Run** menu. The syntax created by the program is now displayed in the main window, as shown below. Note that no options are given on the ITEMS and SCORE commands in this file, indicating that all program defaults will be used.

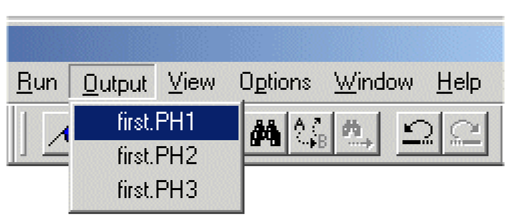
Save the completed syntax to file by selecting the **Save** option on the **File** menu.



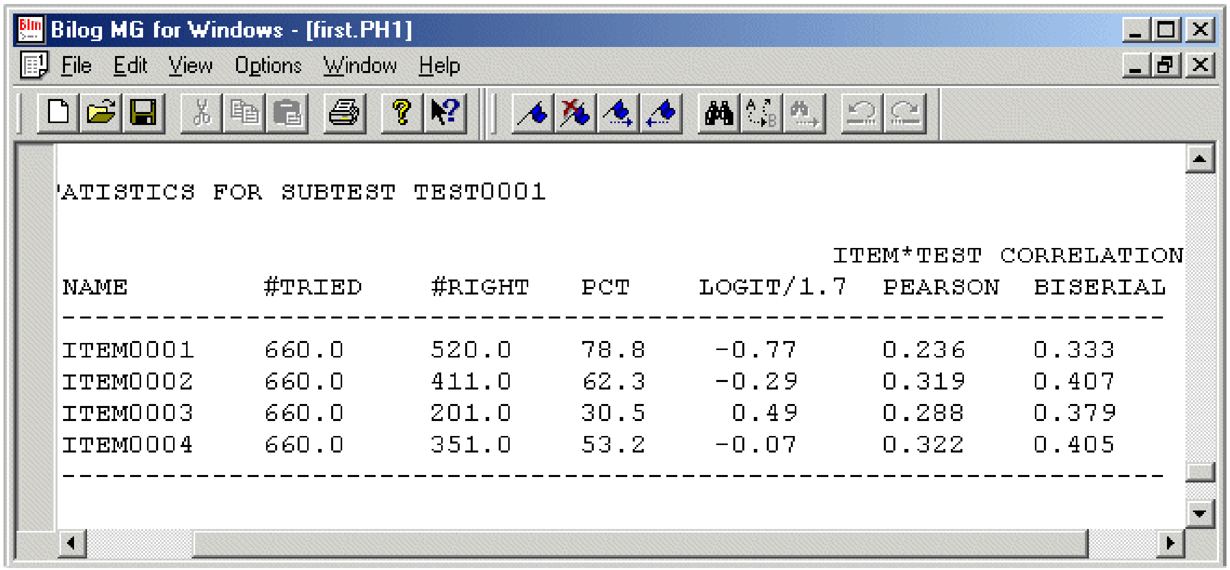
The analysis is now performed by using some of the other options on the **Run** menu. Although the analysis can be done phase by phase (using the **Classical Statistics Only**, **Calibration Only**, and **Scoring Only** options) all three phases can be run in sequence by selecting the **Stats, Calibration and Scoring** option from this menu.



After successful completion of all three phases of the analysis, a message to this effect is displayed on the screen. If a problem was encountered during analysis, this message box will indicate that all phases were not completed successfully. Access the output from the analysis through the **Output** menu. Classical statistics are given in the ***.ph1** file.



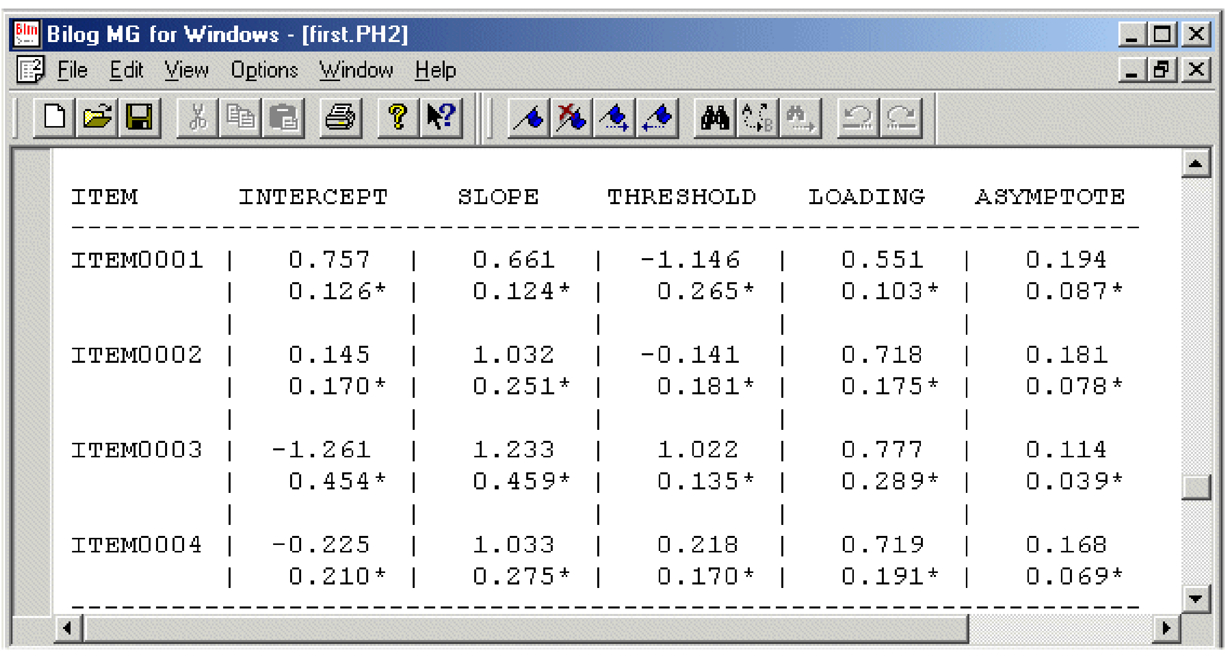
A section of this output file is shown below.



'ATISTICS FOR SUBTEST TEST0001

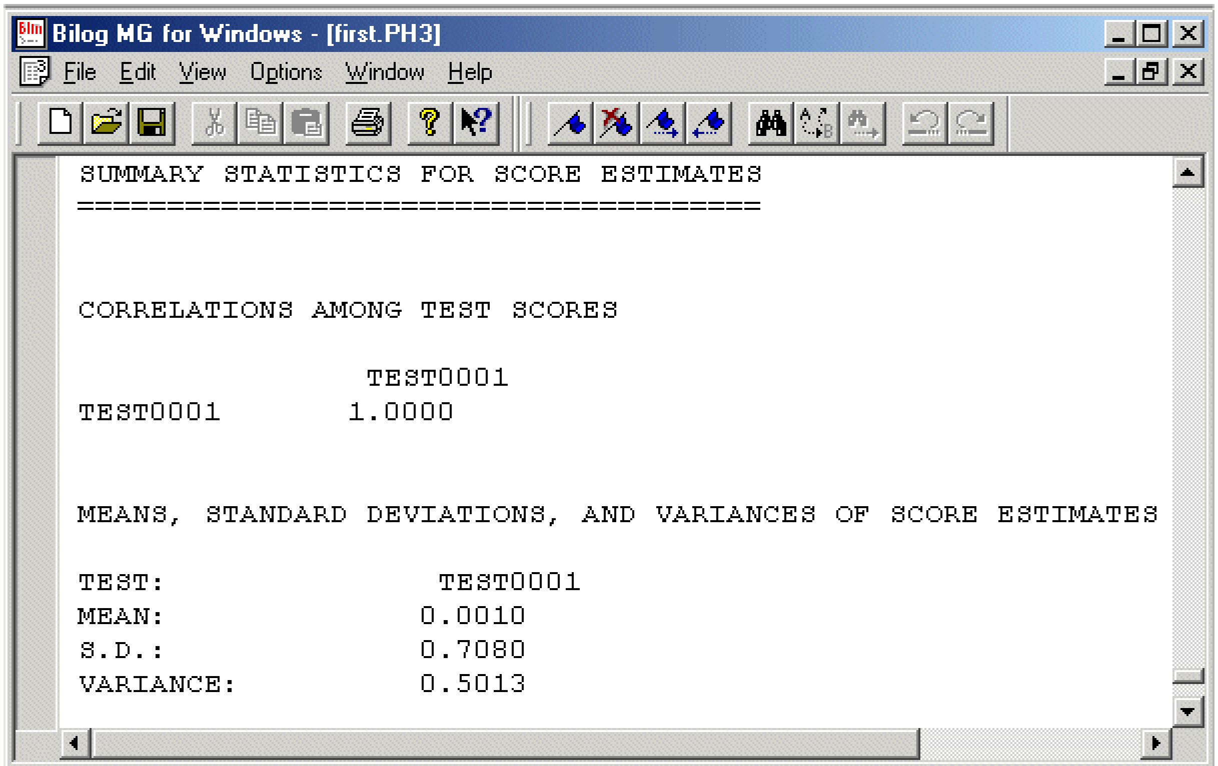
NAME	#TRIED	#RIGHT	PCT	LOGIT/1.7	ITEM*TEST CORRELATION PEARSON	BISERIAL
ITEM0001	660.0	520.0	78.8	-0.77	0.236	0.333
ITEM0002	660.0	411.0	62.3	-0.29	0.319	0.407
ITEM0003	660.0	201.0	30.5	0.49	0.288	0.379
ITEM0004	660.0	351.0	53.2	-0.07	0.322	0.405

In the **first.ph2** file, the results of the item calibrations are given. The item parameter estimates for the four items in the test are shown below.



ITEM	INTERCEPT	SLOPE	THRESHOLD	LOADING	ASYMPTOTE
ITEM0001	0.757	0.661	-1.146	0.551	0.194
	0.126*	0.124*	0.265*	0.103*	0.087*
ITEM0002	0.145	1.032	-0.141	0.718	0.181
	0.170*	0.251*	0.181*	0.175*	0.078*
ITEM0003	-1.261	1.233	1.022	0.777	0.114
	0.454*	0.459*	0.135*	0.289*	0.039*
ITEM0004	-0.225	1.033	0.218	0.719	0.168
	0.210*	0.275*	0.170*	0.191*	0.069*

Scoring results are given in the **first.ph3** file. The complete list of scores is printed to this file by default. A section of this output, showing summary statistics for the score estimates, is shown below.



2.14.2 A second model: DIF model for spelling data

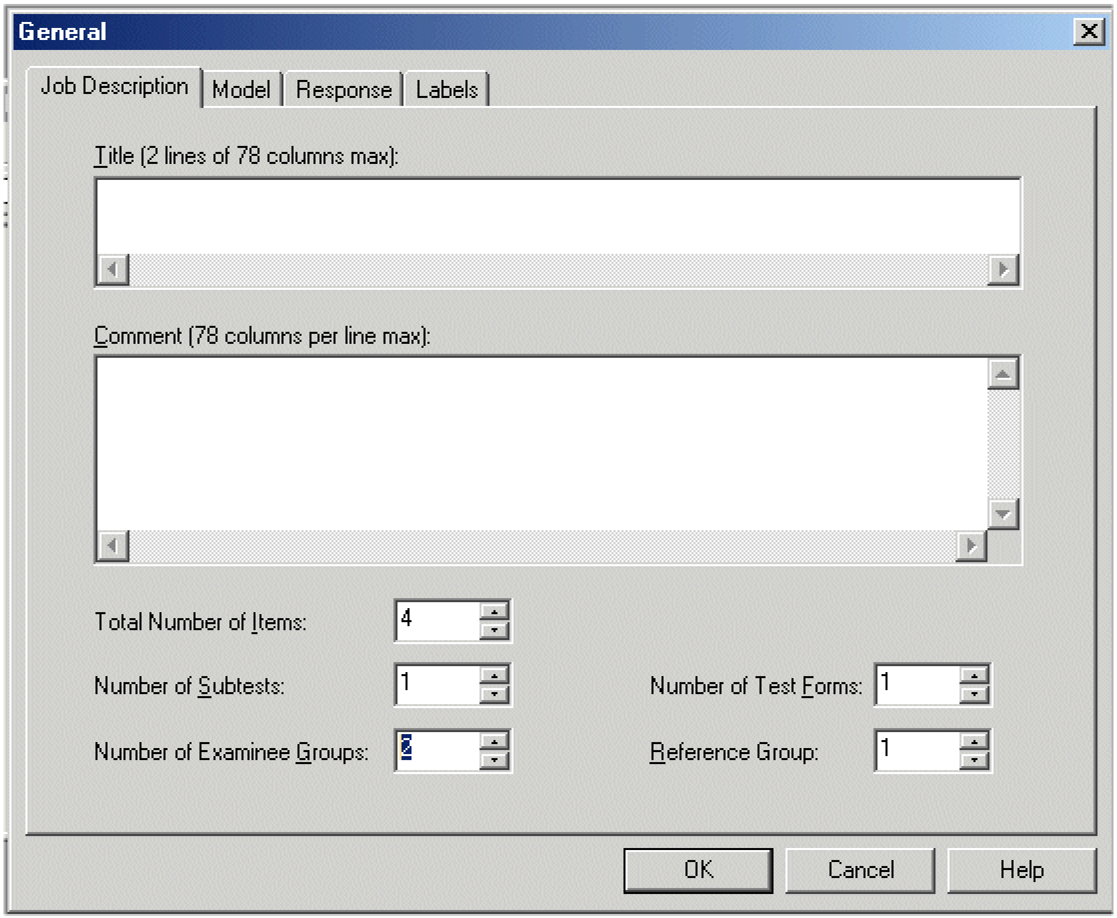
The data analyzed in the previous example actually came from two groups of respondents. The groups in this example are the two sexes. The same four items are presented to both groups on a single test form. The group indicator is found in column 3 of the data records.

```
 11 0000
 21 0001
 31 1000
 41 1001
…
162 1111
```

The third column of the data contains either a 1 or a 2, indicating whether an examinee belonged to group 1 (male) or group 2 (female).

The previous single-group analysis for this group, contained in the command file **first.blm**, is modified to perform a DIF analysis for the two groups. As a first step, the **General** option on the **Setup** menu is used to indicate the presence of multiple groups.

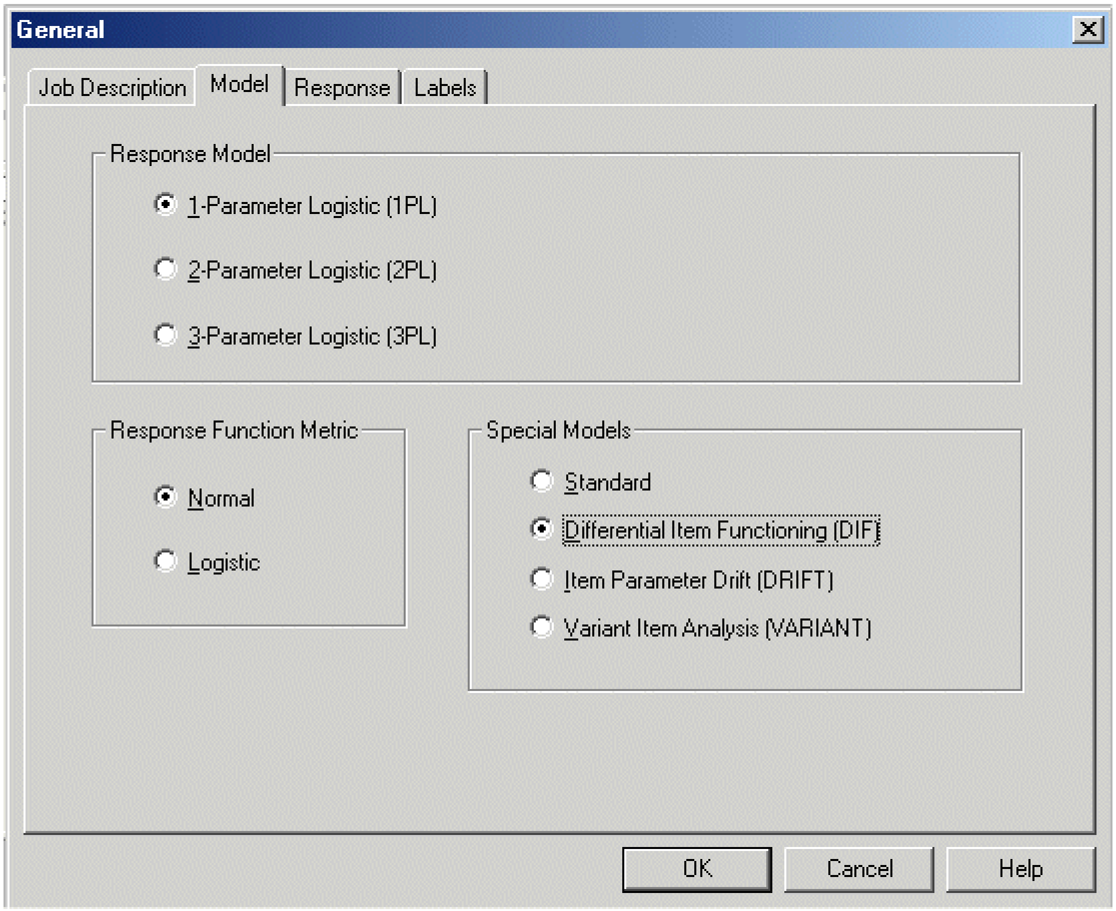
On the **Job Description** tab of the **General** dialog box, change the **Number of Examinee Groups** from the default value of 1 to 2, as shown below.



In the case of a DIF model, a 1PL model is required. To change the model from the default 2PL model previously used, click the **Model** tab and check the **1-Parameter Logistic (1PL)** radio button in the **Response Model** group box.

To request a DIF model, click the **Differential Item Functioning (DIF)** radio button in the **Special Models** group box. By default, the first group will be used as reference group as indicated in the **Reference Group** field.

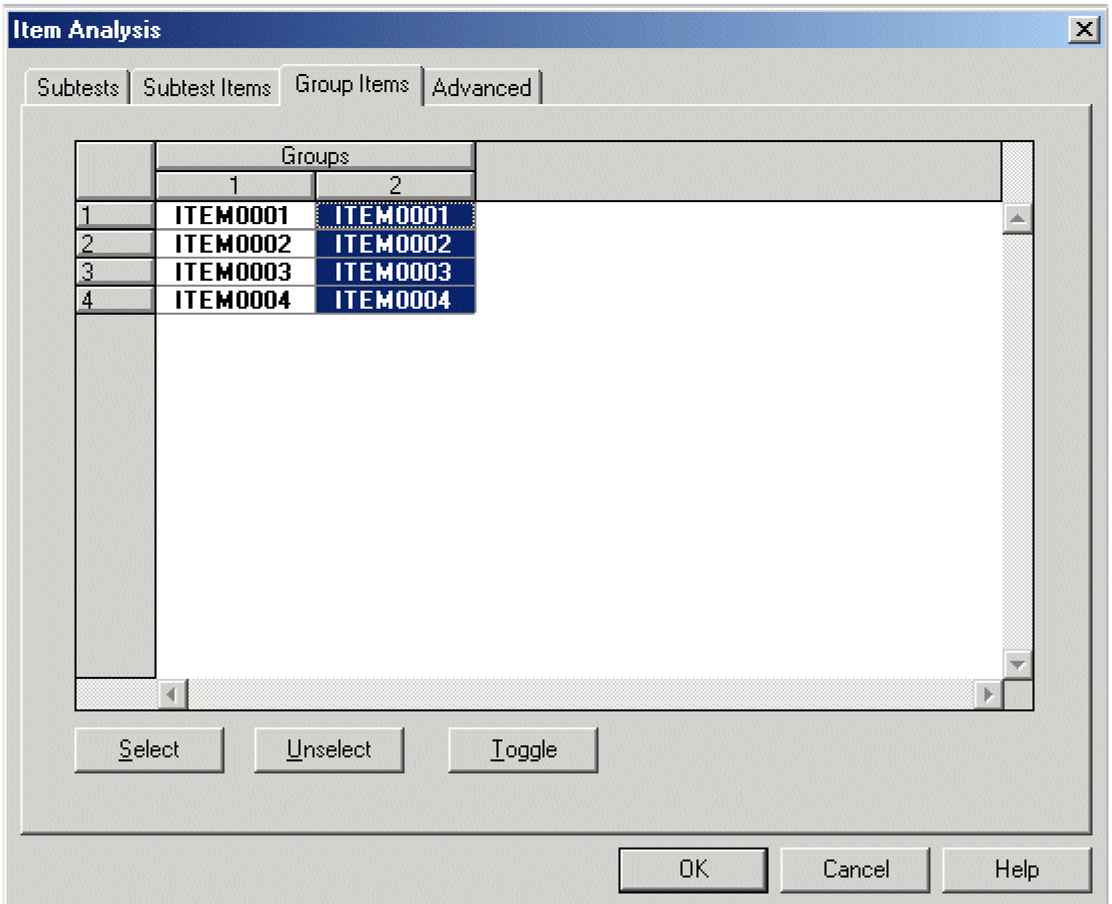
Once this is done, all necessary changes to the **General** dialog box have been made. Click **OK** to return to the main BILOG-MG window.



The allocation of items to be calibrated for each of the two groups is specified using the **Item Analysis** option of the **Setup** menu. Once this option is selected, the **Item Analysis** dialog box is displayed.

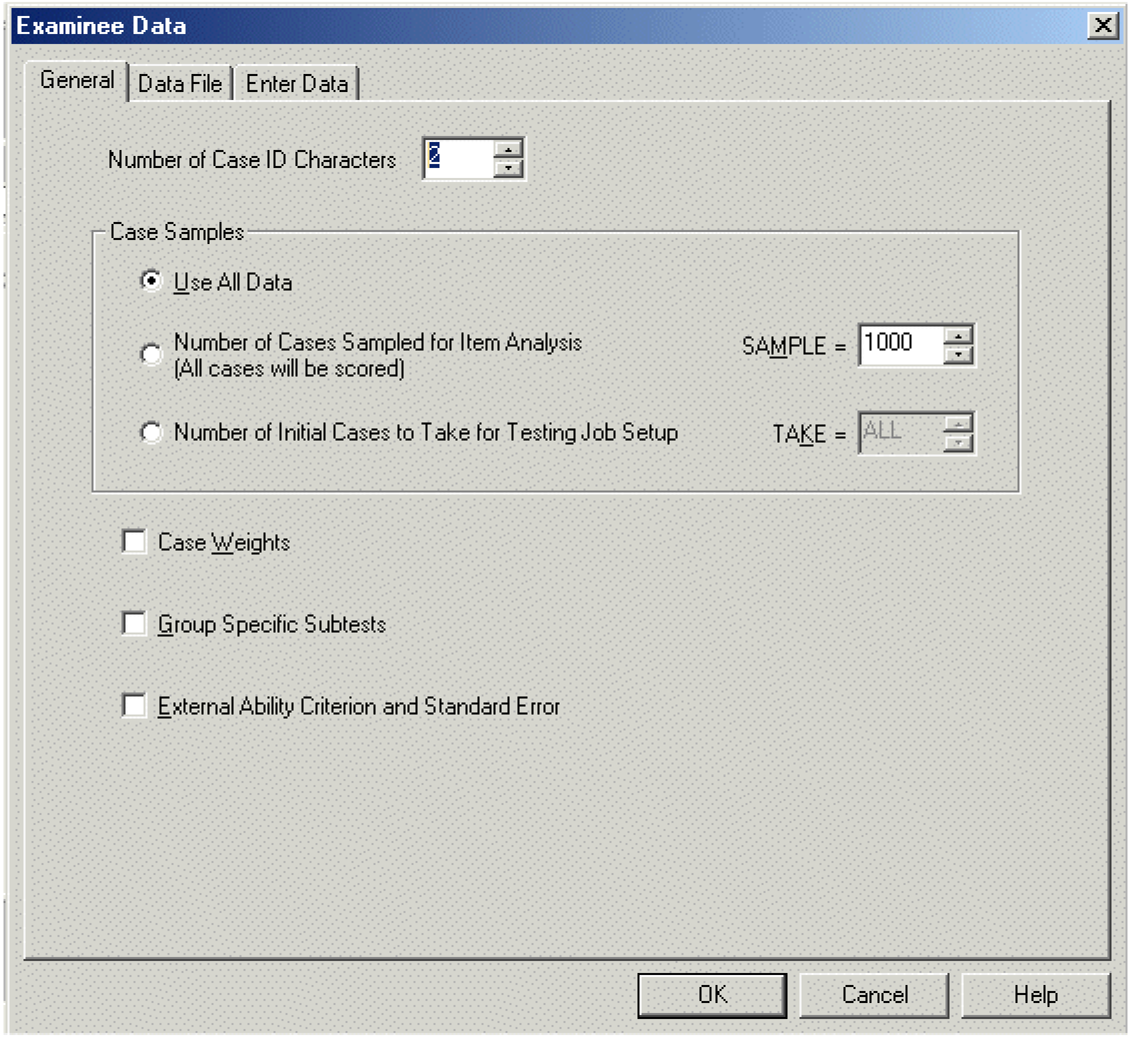
Leaving the **Subtests** tab as previously completed, click the **Group Items** tab. By default, all items will be selected for the first group. This is indicated by the display of the item names in a bold font in the first column of the table. To select all four items for the second group, click on ITEM0001 in the second column. While holding the Shift button on the keyboard down, click on ITEM0004. All four items are now highlighted. Click the **Select** button at the bottom left of the dialog box to select all items.

This completes the model specification. Click **OK** to return to the main window.



The only remaining task is to revise the reading of the data file so that the group identification field can be recognized and processed by the program. To do so, select the **Examinee Data** option from the **Data** menu.

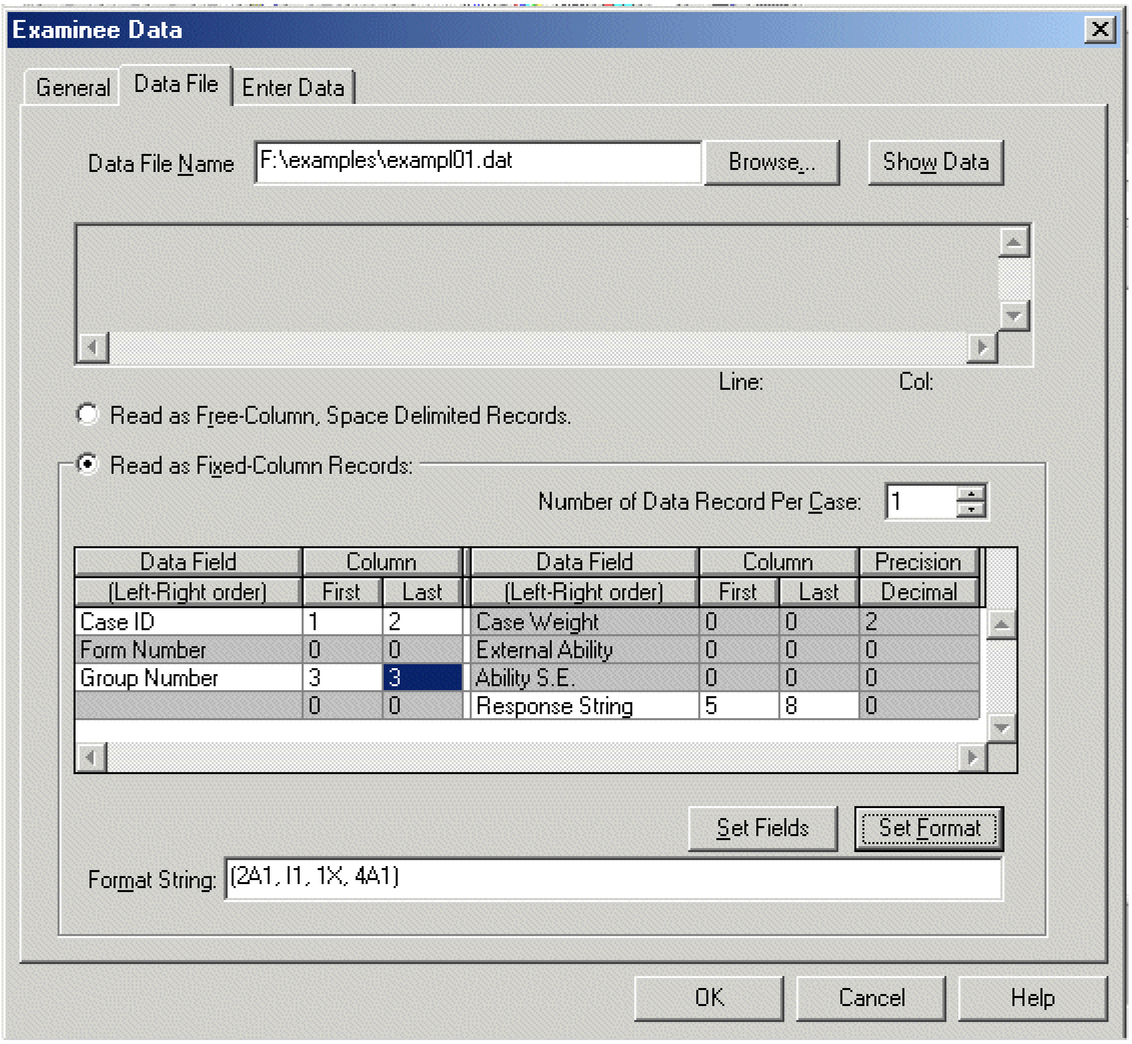
On the **General** tab of the **Examinee Data** dialog box, the number of case identification characters is now decreased to 2, as shown below. (Recall that previously this field was set to 3: in effect, a combination of actual examinee ID and group ID was used to identify the cases in the previous example.)



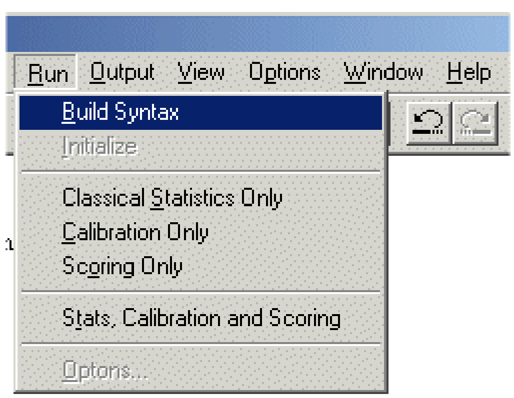
The format statement is now adjusted accordingly by changing the entries in the **Read as Fixed-Column Records** group box:

- The **Last** value for the **Case ID** is set to 2.
- The **Group Number** field, now enabled due to our selection of a multiple-group analysis from the **Setup** menu, is set to 3 under both the **First** and **Last** headers, as the group identification, given in column 3 of the data, is one character in length.
- Finally, these changes are made to the format statement by clicking the **Set Format** button. Note that the format statement now contains an additional "I1" indicating the group number in integer format.

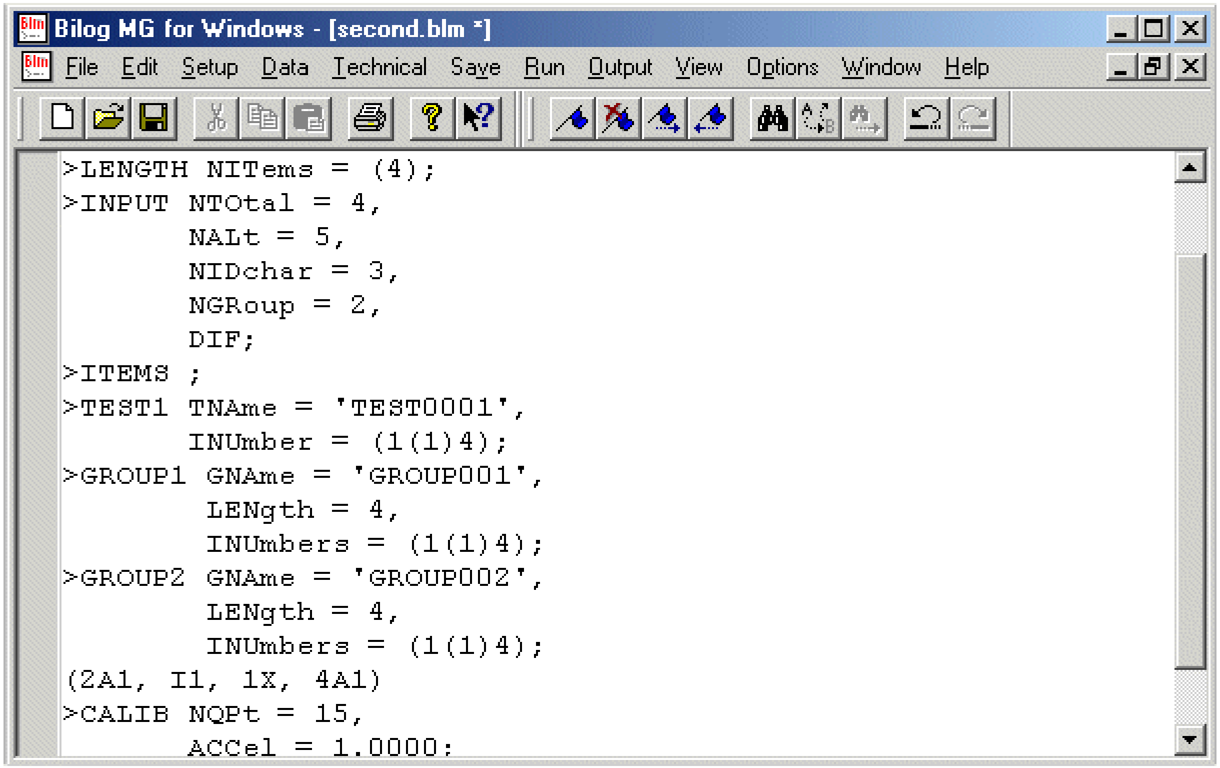
This completes the syntax specification. Return to the main window by clicking the **OK** button.



The revised command file is generated by selecting the **Build Syntax** option from the **Run** menu.



After generating the syntax, it is saved to file using the **Save As** option on the **File** menu. The revised syntax is saved in the file **second.blm** in the **examples** folder. Click the **Save** button after specifying a name for and path to the new command file.

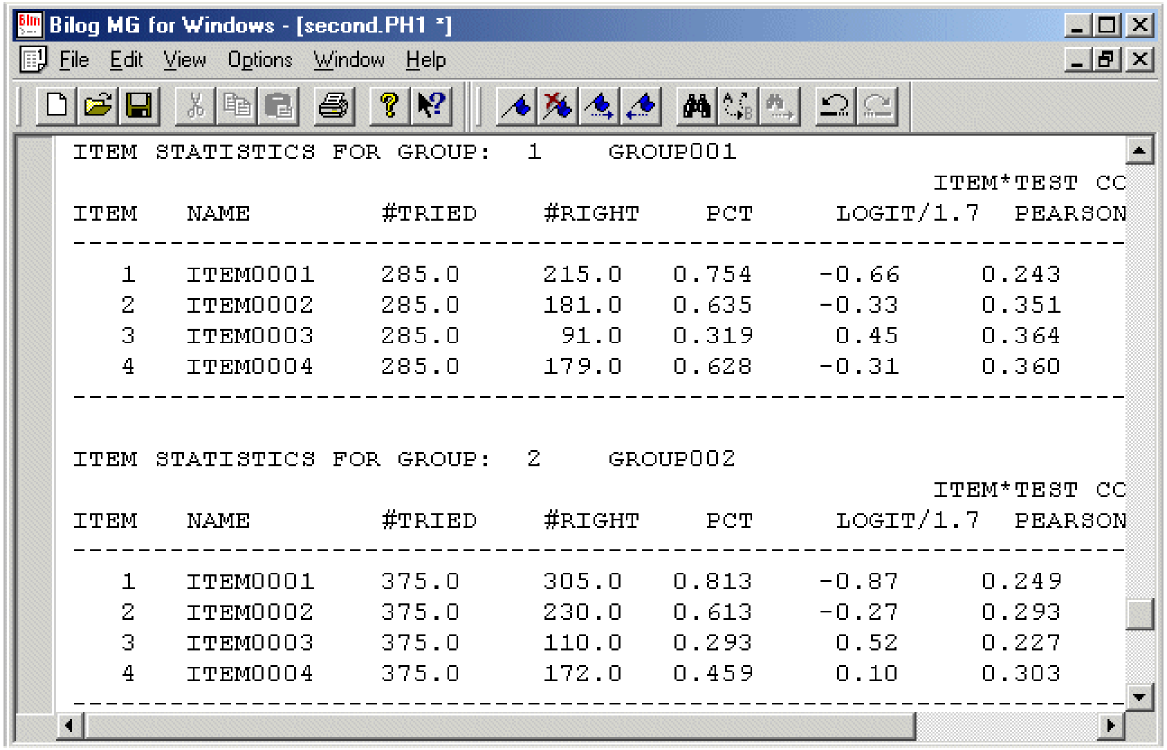


When the syntax displayed in the main BILOG-MG window is compared to the first example, we note the addition of two GROUP commands and the NGROUP and DIF keywords on the INPUT command. The revised format statement is also included. The NPARM keyword on the GLOBAL command (not shown here) indicates that a 1-PL model is requested.

The three phases of the analysis can be run separately using the **Classical Statistics Only**, **Calibration Only**, and **Scoring Only** options on the **Run** menu. To run the phases sequentially, select the **Stats, Calibration, and Scoring** option from this menu.

Output for the analysis is accessed as before using the **Output** menu from the main menu bar.

In the partial output from the **second.ph1** file for this DIF analysis classical item statistics are provided by group. Similar statistics are also given for the combined group (not shown below).



Bilog MG for Windows - [second.PH1 *]

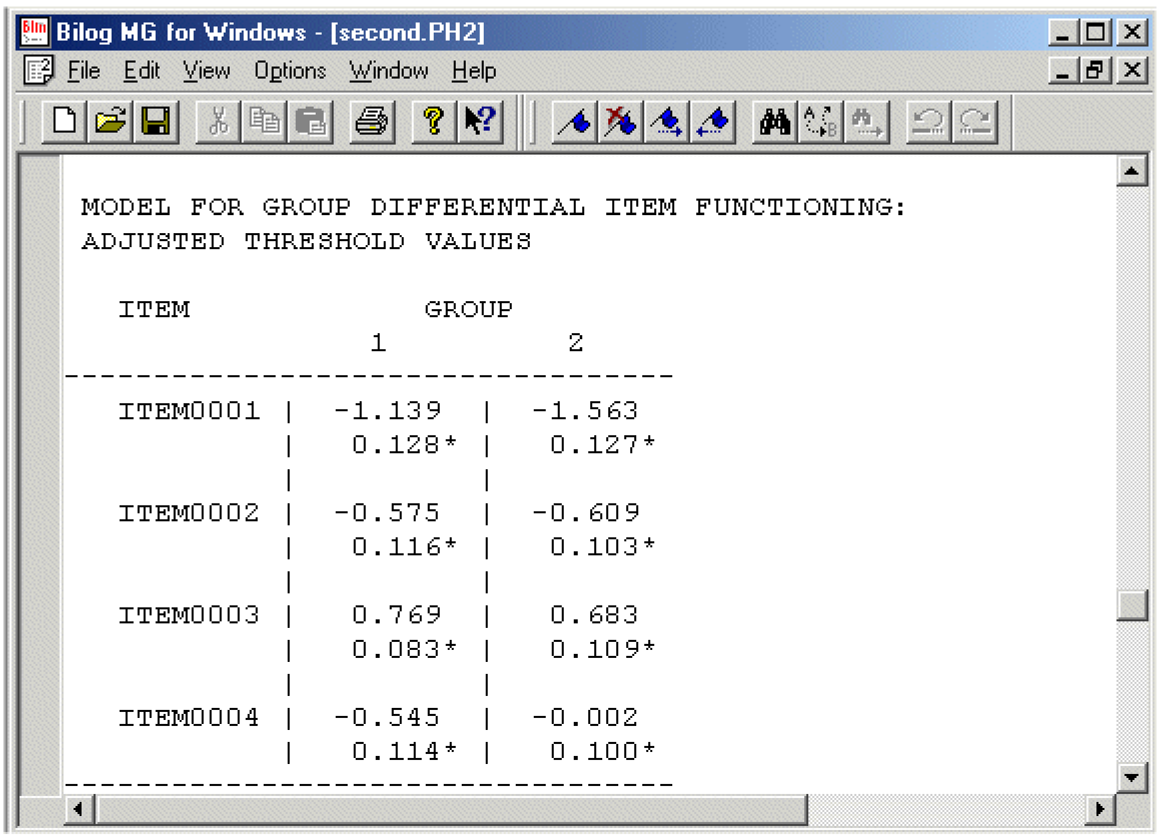
File Edit View Options Window Help

ITEM STATISTICS FOR GROUP: 1 GROUP001

ITEM	NAME	#TRIED	#RIGHT	PCT	LOGIT/1.7	ITEM*TEST CC PEARSON
1	ITEM0001	285.0	215.0	0.754	-0.66	0.243
2	ITEM0002	285.0	181.0	0.635	-0.33	0.351
3	ITEM0003	285.0	91.0	0.319	0.45	0.364
4	ITEM0004	285.0	179.0	0.628	-0.31	0.360

ITEM STATISTICS FOR GROUP: 2 GROUP002

ITEM	NAME	#TRIED	#RIGHT	PCT	LOGIT/1.7	ITEM*TEST CC PEARSON
1	ITEM0001	375.0	305.0	0.813	-0.87	0.249
2	ITEM0002	375.0	230.0	0.613	-0.27	0.293
3	ITEM0003	375.0	110.0	0.293	0.52	0.227
4	ITEM0004	375.0	172.0	0.459	0.10	0.303



Bilog MG for Windows - [second.PH2]

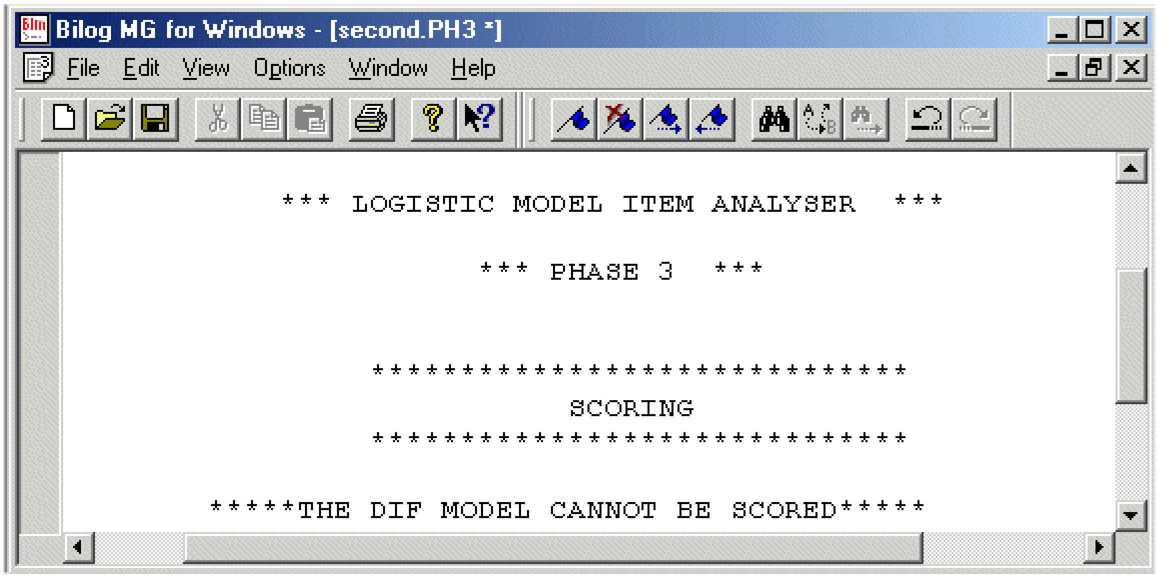
File Edit View Options Window Help

MODEL FOR GROUP DIFFERENTIAL ITEM FUNCTIONING:
ADJUSTED THRESHOLD VALUES

ITEM	GROUP 1	GROUP 2
ITEM0001	-1.139	-1.563
	0.128*	0.127*
ITEM0002	-0.575	-0.609
	0.116*	0.103*
ITEM0003	0.769	0.683
	0.083*	0.109*
ITEM0004	-0.545	-0.002
	0.114*	0.100*

The Phase 2 output in the **second.ph2** file provides item parameter estimates, and DIF specific output as shown above.

Although the **second.ph3** file is created as shown below, no scoring is performed in the case of a DIF analysis.



3 Syntax

3.1 Data structures: ITEMS, TEST, GROUP and FORM commands

In addition to conventional IRT analysis of one test administered to one group of examinees, BILOG-MG is capable of analyzing data from test development and scoring applications in which multiple alternative test forms, each consisting of multiple subtests or scales, are administered to persons in one or more groups. BILOG-MG relies on a system of four commands, ITEMS, TEST, FORM, and GROUP, describing the assignment of items to subtests, forms, and groups. The syntax of these commands is discussed in detail in the syntax section. Here a description is given of how the commands work together to accommodate a wide range of applications.

Related topics

- The FORM command
- The GROUP command
- The ITEMS command
- The TEST commands
- How the FORM and GROUP commands work
- **Setup** menu: **General** dialog box
- **Setup** menu: **Item Analysis** dialog box

The ITEMS command

The ITEMS command attaches names and numbers to items of the test instrument. In the TEST, FORM, and GROUP commands, the user can select items either by name or number. As a convenience to the user, the program can automatically create sequences of eight-character item names consisting of a user-supplied alphabetical section and a sequential numerical section. An ITEMS command is required in all applications of the program. It lists the entire set of items appearing in the test instrument. In the TEST, FORM, and GROUP commands, the full set, or subsets of it, may appear. The examples illustrate the use of the ITEMS command in a variety of applications of the program.

Related topics

- The FORM command
- The GROUP command
- The TEST commands
- How the FORM and GROUP commands work
- **Setup** menu: **General** dialog box

The TEST command

The TEST commands describe the subsets (or scales) that will be scored in the test. There is a separate TEST command for each subtest. A subtest may consist of a combination of items in the instrument including items that appear on different test forms making up the instrument. The TEST commands identify the items belonging to each subtest.

In addition, when the LENGTH command indicates *variant* items are present in a particular subtest (items that are included in the test to obtain item statistics for a subsequent form of the test but are not used in computing test scores), the user identifies these items *with* the corresponding subtest by means of an additional TEST command that immediately follows the TEST command of the subtest.

If the entire instrument is analyzed in a single subtest without variant items, the problem setup requires a single TEST command that lists all the items in the test instrument. Chapter 6 illustrate this type of application.

If multiple subtests of items are selected for analysis, a separate TEST command is required for each subscale. The example in Section 6.6 illustrates the problem setup for an analysis *with* multiple subtests within a single test form. The example discussed in Section 6.8 shows the setup for analysis with multiple subtests for an instrument *and* with multiple test forms. Section 6.7 illustrates the special TEST command setup for an instrument *with variant items.*

Related topics

- LENGTH command
- **Setup** menu: **General** dialog box
- **Setup** menu: **Item Analysis** dialog box
- **Technical** menu: **Assign Fixed Items** dialog box
- **Technical** menu: **Item Parameter Starting Values** dialog box

The FORM command

The FORM command controls the INPUTof the response record. It lists the items in the order in which they appear in the data records. Most applications of BILOG-MG require at least one FORM command.

There are two arrangements in which multiple forms data can be supplied to the program. We refer to them as the *expanded* format and the *compressed* format (see also the file structure specifications):

Expanded format

The response record of each examinee spans the entire set of items appearing in the test instrument. Each item of the test instrument has a unique location (column) in the INPUTrecords. A *not-presented* code appears in the locations of the items belonging to forms not presented to a given examinee. Expanded format is convenient for users who store data in two-dimensional (row by column) arrays typical of many database systems. This format requires only a single FORM command, even though the data arise from multiple forms. Note that the order of the items in the INPUTrecords, and thus the order of their listing on the FORM command, does not have to be the same as that in the list of names and numbers in the ITEMS command (although ordinarily it would be). Note also that a code to identify the form administered to a particular examinee is not read by the program from an expanded format record.

Compressed format

The data record for each examinee contains responses only to the items presented to that person, and the responses appear in the same column field of each record (the number of columns is equal to the number of items in the longest test form). Data entry in the compressed format is easier than in the expanded format and results in smaller data files.

With compressed-format data, the locations of the items in the INPUTrecords are not unique. An item in one record may occupy the same column as a different item in another record. A separate FORM command is therefore required for each test form in the instrument. In addition, each response record must contain a number identifying the FORM command that applies to that record. The number (1, 2, 3, etc.) refers to the order of the FORM command in the command file. The item list of the corresponding FORM command gives the names or numbers of the items in the order that they appear in the response field of the data records (see Section 3.2.16 for details). Internally, the program works by expanding the compressed records and inserting not-presented codes in locations corresponding to the forms *not* administered to the examinee.

Related topics

- **Setup** menu: **General** dialog box (see Section 2.3)
- **Setup** menu: **Item Analysis** dialog box

The GROUP command

GROUP commands are required whenever a multiple-group analysis is specified. The number of commands is equal to the number of groups in the analysis. GROUP commands serve two purposes. First, they identify groups of respondents for multiple-group analysis. Second, they identify the set of items administered to each group. Note that whenever a multiple-group analysis is requested, each response record must contain a number identifying the GROUP command that applies to that record.

Related topics

- **Setup** menu: **General** dialog box (see Section 2.3)
- **Setup** menu: **Item Analysis** dialog box

How the FORM and GROUP commands work

The FORM and GROUP commands control the INPUTof the individual response records. How they work together depends on the following:

- the format of the response records (expanded or compressed)

- the number of forms in the test instrument
- the assignment of test forms to the respondents
- the number of groups in the IRT analysis

The sections below describe how these factors determine the structure of the FORM and GROUP commands.

Instruments with a single test form

When an instrument consists of a single test form, a single FORM command is assumed, and it applies to all records. The program reads the entire response records according to the specifications on the command. If a FORM command is not included in the problem setup, the program reads the response records according to the order of items in the ITEMS command list. As in all applications with a single FORM command, the response records do not contain a form indicator.

Single-group analysis

The examples in Sections 6.1 and 6.3 illustrate the simple case of a single-group analysis of a single test form. The program reads all response records according to the specifications on the FORM or ITEMS commands. GROUP commands are not required for the analysis.

Multiple-group analysis

In multiple-group analysis of a single test form, the groups may represent naturally occurring subgroups within a population of respondents, or groups of respondents drawn from different populations. In either case, the structure of the FORM and GROUP commands is the same. A single FORM command applies to all response records and a separate GROUP command is required for each group of respondents in the analysis. Because all respondents receive the same test form, and thus respond to the same set of items, the lists of items in the GROUP commands are the same for all groups in the analysis. The lists include all of the items specified in the FORM or ITEMS commands.

The primary function of GROUP commands in applications of this type is to identify the groups of respondents for multiple-group analysis. The example in Section 6.2 shows how this command structure applies to examinations of differential item functioning in subgroups of a population. Group differences in the latent distributions of ability may also be examined in this way.

Instruments with multiple test forms

When an instrument consists of multiple test forms, the structure of the FORM and GROUP commands depends in part on whether the forms are administered to equivalent or nonequivalent groups of respondents. If the forms of the instrument are randomly assigned to respondents drawn from a single population, the groups are equivalent, and the data may be analyzed with a single-group IRT model. GROUP commands are not required in this case, but may be added to examine subgroup differences in item functioning. When test forms are administered to nonequivalent groups of respondents, the forms must contain common “linking” items, and a multiple-group analysis is necessary to place the items from the forms on the same scale. GROUP commands are required in this case.

The number of GROUP commands corresponds to the number of groups in the analysis. In multiple-form applications the response records may follow either of the two formats. The sections below show how the structure of the FORM and GROUP commands depends on these formats.

Single-group analysis

When there are multiple forms in the test instrument but only one group of examinees, multiple FORM commands are required if the compressed data format is used, but a GROUP command or a group indicator on the response records is not required. Section 6.4 illustrates an application of this type.

Multiple-group analysis

In the case of multiple forms and multiple groups, all such applications can be handled by *expanded* format. Only one FORM command is then required and the data records will not contain a forms indicator. Similarly, if the assignment of items to groups is performed in expanded format, including the codes for items presented to a given examinee in a given group, the GROUP commands require only the group names, not the item identifications. Specification of the items assigned to each group will, however, shorten the run time. The example in Section 6.5 illustrates this type of data structure. The expanded style of data entry is mandatory in applications where the test forms contain more than one subtest and the examinee is assigned to *different* groups for different subtests. This can occur in complex two-stage testing designs.

In more typical applications, however, whole forms rather than subtests are assigned to groups. In this case, the *compressed* style of data entry is suitable and may be more convenient. The GROUP commands must then contain, in addition to the group name, a list of all items on all forms assigned to the corresponding groups. The data records must include both a forms identifier and a group identifier. The advantage of this method is that response records need not contain codes for not-presented items. Examples illustrating this type of data INPUTare discussed in Sections 6.4 and 6.8 respectively.

Related topics

- **Setup** menu: **General** dialog box (see Section 2.3)
- **Setup** menu: **Item Analysis** dialog box

3.2 Using the command language

3.2.1 Overview of syntax

BILOG-MG uses command lines employing the general syntax:

```
>NAME  KEYWORD1=N, KEYWORD2=(list), …, OPTIONn;
```

The following rules apply:

- A greater-than sign (>) must be entered in column 1 of the first line of a command and followed without a space by the command name.
- All command names, keywords, options, and keyword values may be entered in upper and/or lower case.
- Command names, keywords, options, and keyword values may be entered in full or abbreviated to the first three characters.
- At least one space must separate the command name from any keywords or options.
- Commas must separate all keywords and options.
- The equals sign is used to set a keyword equal to a value, which may be integer, real or character. A real value must contain a decimal point. A character value must be enclosed in single quotes if it:
    - Contains more than eight characters
    - Begins with a numeral
    - Contains embedded blanks, commas, slashes, or semi-colons

For example: DFNAME='EXAMPL01.DAT', TNAME='20-ITEMS';

- A keyword may be vector valued, *i.e.*, set equal to a list of integer, real or character constants, separated with commas or spaces, and enclosed in left and right parentheses (as KEYWORD2 above).
- If the list is an arithmetic progression of integer or decimal numbers, the short form, *first (increment) last*, may be used. Thus, a selection of items 1, 3, 7, 8, 9, 10, 15 may be entered as 1,3,7(1)10,15. Real values may be entered in a similar way.
- If the values in the list are equal, the form, *value (0) number of values*, may be used. Thus 1.0, 1.0, 1.0, 1.0, 1.0 may be entered as 1.0(0)5.
- The *italic* elements in the format description are variables that the user needs to replace.
- Command lines may not exceed 80 columns. Continuation on one or more lines is permitted.
- Each command terminates with a semi-colon (;). The semi-colon signals the end of the command and the beginning of a new command.

3.2.2 Order of commands

The table below lists all available BILOG-MG commands in their necessary order. Commands marked below as "Required" must appear in each problem *in the order shown*. All other commands are optional. Note that, in the rest of this chapter, description for the commands follow an alphabetical order. The data layout must be described in a variable format statement. This statement is entered within parentheses.

Table 7.3: keywords and options in BILOG-MG

Command	Keywords and options	Required
TITLE1		*
TITLE2		*
COMMENT		
GLOBAL	CFNAME, DFNAME, IFNAME, LOGISTIC, MFNAME, NPARM, NTEST, NVTEST, NWGHT, OMITS, SAVE, PRN;	*
SAVE	CALIB, COVARIANCE, DIF, DRIFT, EXPECTED, ISTAT, MASTER, PARM, POST, SCORE, TSTAT, PDISTRIB;	
LENGTH	NITEMS, NVARIANT;	*
INPUT	DIAGNOSE, DIF, DRIFT, KFNAME, NALT, NFMT, NFNAME, NFORM, NGROUP, NIDCHAR, NTOTAL, OFNAME, PERSONAL, SAMPLE, TAKE, TYPE, EXTERNAL;	*
ITEMS	INAMES, INUMBERS;	*
TESTi	DISPERSN, GUESS, INAME, INTERCPT, INUMBER, SLOPE, THRESHLD, TNAME, FIX;	*
FORMj	INAMES, INUMBERS, LENGTH;	
GROUPk	GNAME, INAMES, INUMBERS, LENGTH;	
DRIFT	MAXPOWER, MIDPOINT;	
(variable format statement)		*
CALIB	ACCEL, COMMON, CRIT, CYCLES, DIAGNOSIS, EMPIRICAL, FIXED, FLOAT, GPRIOR, IDIST, NEWTON, NOFLOAT, NOGPRIOR, NORMAL, NOSPRIOR, NOTPRIOR, NQPT, NSD, PLOT, PRINT, READPRIOR, REFERENCE, RIDGE, SELECT, SPRIOR, TPRIOR, CHI, GROUP-PLOTS, NOADJUST, RASCH, NFULL;	*
QUADk (for group k)	POINTS, WEIGHTS;	
PRIORSi (for sub-test I)	ALPHA, BETA, SMU, SSIGMA, TMU, TSIGMA;	
SCORE	BIWEIGHT, FIT, IDIST, INFO, LOCATION, METHOD, NOPRINT, NQPT, PMN, POP, PSD, RSCTYPE, SCALE, YCOMMON, MOMENTS, DOMAIN, FILE, REFERENCE, READF, NFORMS;	
QUADSk (for group k)	POINTS, WEIGHTS;	

Note that if there are not variant items in the subtest, there is one TEST command for each subtest. If a subtest contains variant test items, there must be exactly two TEST commands for that subtest. The first identifies the main test items while the second identifies the variant test items.

Related topics

- Location of keywords in interface (see Section 2.13)

3.2.2 CALIB command

Purpose

To control the item parameter estimation procedure and the specification of prior distributions on the item parameters.

Format

```
>CALIB   NQPT=n, CYCLES=n, NEWTON=n, PRINT=n, CRIT=n, IDIST=n,
         PLOT=n, DIAGNOSIS=n, REFERENCE=n, SELECT=(list), RIDGE=(a,b,c),
         ACCEL=n, NSD=n, COMMON, EMPIRICAL, NORMAL, FIXED, TPRIOR, SPRIOR,
         GPRIOR, NOTPRIOR, NOSPRIOR, NOGPRIOR, READPRIOR, NOFLOAT, FLOAT,
         NOADJUST, GROUP-PLOT, RASCH, NFULL, CHI=(a,b);
```

Examples

This example uses simulated responses to illustrate nonequivalent groups equating of two forms of a 25-item multiple-choice examination administered to different populations. Separate latent distributions are estimated for each population (EMPIRICAL option). The indeterminacy in location and scale of the distributions is resolved by setting the mean and standard deviation of Group 1 to 0 and 1, respectively, with REF=1 on the CALIB command.

```
>CALIB NQPT=10, EMPIRICAL, CYCLES=25, NEWTON=5, CRIT=0.01, PLOT=0.05,
       REFERENCE=1, TPRIOR;
```

In the following example of vertical equating of test forms over three grade levels, students at each of three grade levels were given grade-appropriate versions of a arithmetic examination. The distributions of ability are assumed to be normal at each grade level (NORMAL option). The second group serves as the reference group in the calibration of the items. A prior is placed on the item thresholds by the addition of the TPRIOR option.

```
>CALIB   NQPT=20, NORMAL, CYCLE=30, TPRIOR, NEWTON=2, CRIT=0.01,
         REFERENCE=2;
```

In the following example of a 3-PL model, the PLOT keyword has been set to 0.99 so that all item response functions will be plotted. The FLOAT option is added to request the MML (under normal distribution assumptions) of the means of the prior distributions on the item parameters along with the parameters. This option should not be invoked when the data set is small and the items few. The acceleration constant (ACCEL keyword) is set to 0.5 instead of the default value of 1.0 for a single group analysis.

```
>CALIB NQPT=6, FLOAT, PLOT=0.99, CYCLES=15, NEWTON=3, ACCEL=0.5;
```

The next example, again of a 3-PL model, illustrates the command's usage in the presence of aggregate-level, multiple-matrix sampling data. In this case, the data come from eight forms of a rather difficult, multiple-choice instrument. Since aggregate-level data are always more informative than individual-level item responses, it is worthwhile to increase the number of quadrature points (NQPT), to set a stricter convergence criterion (CRIT), and to increase the CYCLES limit. A prior on the thresholds (TPRIOR) and a ridge constant of 0.8 (RIDGE) are required for convergence with the exceptionally difficult second subtest.

Aggregate-level data typically have smaller slopes in the 0,1 metric than do person-level data. Thus, the mean of the prior for log slopes is set to 0.5 with the READPRIOR option and the succeeding PRIOR commands as shown.

```
>CALIB   NQPT=3, CYCLES=24, NEWTON=4, CRIT=0.0050,
         RIDGE=(2, 0.8000, 2.0000), ACCEL=1.0000, TPRIOR, READPRI,
         NOFLOAT;
>PRIORS1 SMU=(0.5000(0)8);
>PRIORS2 SMU=(0.5000(0)8);
```

Related topics

- PRIORS command (see Section 3.2.10)
- QUAD command (see Section 3.2.11)
- **Setup** menu: **General** dialog box (see Section 2.3)
- **Setup** menu: **Item Analysis** dialog box
- **Technical** menu: **Calibration Options** dialog box (see Section 2.5)
- **Technical** menu: **Data Options** dialog box
- **Technical** menu: **Item Parameter Prior Constraints** dialog box

ACCEL keyword (optional)

Purpose

To set the acceleration constant for the E-steps.

Format

```
ACCEL=n
```

Default

1.0 for NGROUP=1 and 0.5 for NGROUP>1.

Related topics

- CALIB command: CYCLES keyword (see Section 3.2.2)
- INPUT command: NGROUP keyword (see Section 3.2.7)
- **Technical** menu: **Data Options** dialog box (see Section 2.5)

CHI keyword (optional)

Purpose

To specify the number of items required and the number of intervals used for χ^2 computations.

Format

```
CHI=(a,b)
```

where a is the number of items for computation of the χ^2 fit statistics, and b is the number of intervals into which the score continuum will be divided for purposes of computation of χ^2 item fit statistics.

Default

```
CHI=(20,9).
```

Example

In the CALIB command shown below, the CHI keyword is used to request the calculation of the χ^2 item fit statistics on 18 items and 7 intervals.

```
>CALIB   CYCLE=30, TPRIOR, NEWTON=2, CRIT=0.01, CHI=(18,7);
```

Related topics

- **Setup** menu: **Item Analysis** dialog box (see Section 2.3)

COMMON option (optional)

Purpose

To estimate a common value for the lower asymptote of all items in the 3PL model.

Format

```
COMMON
```

Default

Separate values for the lower asymptotes.

Example

If the CALIB command

```
>CALIB NQPT=10, CYCLES=30, NEWTON=5, COMMON;
```

is used for a 3PL model, output as shown below is obtained. Note that the asymptote parameter is estimated at a common value of 0.031 for all items.

```
SUBTEST SIM     ;  ITEM PARAMETERS AFTER CYCLE  15

ITEM  INTERCEPT    SLOPE    THRESHOLD   LOADING   ASYMPTOTE    CHISQ   DF
            S.E.        S.E.        S.E.        S.E.        S.E.      (PROB)
--------------------------------------------------------------------------
T01|   1.320  |   0.968  |  -1.363  |   0.695  |   0.031  |     3.5   6.0
   |   0.185* |   0.192* |   0.198* |   0.138* |   0.006* | (0.7490)
   |          |          |          |          |          |
T02|   1.516  |   0.984  |  -1.541  |   0.701  |   0.031  |     3.2   6.0
   |   0.212* |   0.197* |   0.212* |   0.140* |   0.006* | (0.7783)
   |          |          |          |          |          |
T03|   1.020  |   1.131  |  -0.902  |   0.749  |   0.031  |     2.8   6.0
   |   0.170* |   0.227* |   0.142* |   0.150* |   0.006* | (0.8294)
   |          |          |          |          |          |
T04|   0.603  |   0.787  |  -0.766  |   0.619  |   0.031  |     8.2   8.0
   |   0.118* |   0.134* |   0.155* |   0.105* |   0.006* | (0.4113)
   |          |          |          |          |          |
T05|   0.780  |   0.695  |  -1.123  |   0.571  |   0.031  |     6.3   7.0
   |   0.119* |   0.124* |   0.208* |   0.102* |   0.006* | (0.5066)
```

Related topics

- GLOBAL command: NPARM keyword (see Section 3.2.5)

CRIT keyword (optional)

Purpose

To set the convergence criterion for EM and Newton iterations.

Format

```
CRIT=n
```

Default

0.01.

Example

Here, the convergence criterion has been set to the more restrictive value of 0.0050 in order to deal with a more informative aggregate-level data set.

```
>CALIB NQPT=30, CYCLES=24, NEWTON=4, CRIT=0.0050,
       RIDGE=(2, 0.8000, 2.0000), ACCEL=1.0000, TPRIOR, READPRI,
       NOFLOAT;
```

Related topics

- CALIB command: CYCLES and NEWTON keywords (see Section 3.2.2)
- **Setup** menu: **Item Analysis** dialog box

CYCLES keyword (optional)

Purpose

To set the maximum number of EM cycles. If CYCLES=0 and NEWTON=0, item parameter estimates will be calculated from the classical item statistics from Phase 1 or from the starting values of the TEST command. The former will be corrected for guessing if the 3-parameter model is selected.

Format

```
CYCLES=n
```

Default

10 (for all subtests).

Examples

In this example of vertical equating of test forms over three grade levels, a maximum of 30 EM cycles and 2 Newton-Gauss iterations are requested.

```
>CALIB   NQPT=20,NORMAL,CYCLE=30,TPRIOR,NEWTON=2,CRIT=0.01,REFERENCE=2;
```

Here, the CYCLES limit is increased in order to deal with a more informative aggregate-level data set.

```
>CALIB NQPT=30, CYCLES=24, NEWTON=4, CRIT=0.0050, NOFLOAT,
       RIDGE=(2, 0.8000, 2.0000), ACCEL=1.0000, TPRIOR, READPRI;
```

Related topics

- CALIB command: NEWTON keyword (see Section 3.2.2)

- TEST command (see Section 3.2.15)
- **Setup** menu: **Item Analysis** dialog box (see Section 2.3)

DIAGNOSIS keyword (optional)

Purpose

To set the level of diagnostic printout.

Format

```
DIAGNOSIS=n
```

Default

0.

Example

When DIAGNOSIS is set to 1, for example, item parameter estimates are printed to the Phase 2 output phase at each iteration.

EMPIRICAL option (optional)

Purpose

To estimate the score distribution in the respondent population in the form of a discrete distribution on NQPT points. This empirical distribution is used in place of the prior in the MML estimation of the item parameters.

If NGROUP >1, separate score distributions are estimated for each group.

Format

```
EMPIRICAL
```

Default

Not-empirical (FIXED) if NGROUP=1; empirical if NGROUP>1.

Example

For this example, which comes from a simulation of non-equivalent groups equating, the EMPIRICAL option is used to estimate separate latent distributions for each population.

```
>CALIB NQPT=10, EMPIRICAL, CYCLES=25, NEWTON=5, CRIT=0.01, PLOT=0.05,
       REFERENCE=1, TPRIOR;
```

Related topics

- CALIB command: FIXED option (see Section 3.2.2)
- CALIB command: NQPT keyword
- INPUT command: NGROUP keyword (see Section 3.2.7)
- **Setup** menu: **Item Analysis** dialog box (see Section 2.3)

FIXED option (optional)

Purpose

To keep the prior distributions of ability in the population of respondents fixed at the values specified in the IDIST keyword and/or the QUAD commands.

Format

```
FIXED
```

Default

Same as for EMPIRICAL.

Related topics

- CALIB command: EMPIRICAL option (see Section 3.2.2)
- CALIB command: IDIST keyword
- QUAD command (see Section 3.2.11)
- **Technical** menu: **Calibration Options** dialog box (see Section 2.5)

FLOAT option (optional)

Purpose

To estimate the means of the prior distributions on the item parameters by marginal maximum likelihood (under normal distribution assumptions), along with the parameters. To keep the means of the prior distributions on the item parameters fixed at their specified values during estimation, the NOFLOAT option should be used.

Standard deviations of the priors are fixed in either case. The FLOAT option should *not* be invoked when the data set is small and the items few. The means of the item parameters may drift indefinitely during the estimation cycles under these conditions.

Format

```
FLOAT
```

Default

```
NOFLOAT if NGROUP=1; FLOAT if NGROUP>1
```

Example

In this example of a 3-PL model, the FLOAT option is added to request the MML (under normal distribution assumptions) of the means of the prior distributions on the item parameters along with the parameters. This option should not be invoked when the data set is small and the items few. The acceleration constant (ACCEL keyword) is set to 0.5 instead of the default value of 1.0 for a single group analysis.

```
>CALIB NQPT=6, FLOAT, PLOT=0.99, CYCLES=15, NEWTON=3, ACCEL=0.5;
```

Related topics

- INPUT command: NGROUP keyword (see Section 3.2.7)
- **Setup** menu: **Item Analysis** dialog box (see Section 2.3)

GPRIOR/NOGPRIOR option (optional)

Purpose

To select or suppress prior distributions on the lower asymptote (guessing) parameter. This may be needed in order to give plausible values for easy items (which carry little or no information about guessing).

Format

```
GPRIOR/NOGPRIOR
```

Default

- 1PL model, NOGPRIOR
- 2PL model, NOGPRIOR
- 3PL model, GPRIOR

Examples

For a 3PL model, priors on slopes and asymptote parameters are assumed. To remove these priors, the CALIB command

```
>CALIB NQPT=10, CYCLES=15, NOSPRIOR, NOGPRIOR;
```

may be used.

To remove the default prior distribution on the asymptote parameters and use a prior distribution on the thresholds instead, use

```
>CALIB NQPT=10, CYCLES=15, SPRIOR, NOGPRIOR, TPRIOR;
```

Related topics

- CALIB command: SPRIOR/NOSPRIOR and TPRIOR/NOTPRIOR options
- **Setup** menu: **Item Analysis** dialog box (see Section 2.3)

GROUP-PLOTS option (optional)

Purpose

To provide plots showing the proportions of correct responses for each separate group in a multiple-group analysis. These plots may provide more information than the combined plot provided by the PLOT keyword.

Format

```
GROUP-PLOTS
```

Default

Combined plots, if PLOT keyword is used.

Example

In the CALIB command from a two-group analysis below, the PLOT keyword has been set to 0.99 so that all item response functions will be plotted. In order to obtain plots by group, the GROUP-PLOT keyword has been added.

```
>CALIB NQPT=6, REFERENCE=2, PLOT=0.99, CYCLES=15, NEWTON=3, GROUP-PLOT;
```

Related topics

- CALIB command: PLOT keyword
- Technical menu: **Calibration Options** dialog box (see Section 2.5)

IDIST keyword (optional)

Purpose

To designate the type of prior distribution in the population of respondents.

Format

```
IDIST=n
```

n = 0	standard normal approximation
n = 1	separate, arbitrary discrete priors for each group for each subtest read from QUAD command
n = 2	separate, arbitrary discrete priors for each group read from QUAD command

Default

0.

Example

This example illustrates how user-supplied priors for the latent distributions are specified with IDIST=1 on the CALIB command. The points and weights for these distributions are supplied in the QUAD commands. Note that with IDIST=1, there are separate QUAD commands for each group for each subtest. Within each subtest, the points are the same for each group. This is a requirement of the program. But as the example shows, the points for the groups may differ by subtest.

```
>CALIB   IDIST=1,READPR,EMPIRICAL,NQPT=16,CYCLE=25,TPRIOR,NEWTON=5,
         CRIT=0.01,REFERENCE=1,NOFLOAT;
>QUAD1   POINTS=(-0.4598E+01 -0.3560E+01 -0.2522E+01 -0.1484E+01
         -0.4453E+00 0.5930E+00  0.1631E+01  0.2670E+01  0.3708E+01
         0.4746E+01),
         WEIGHTS=(0.2464E-05  0.4435E-03  0.1724E-01  0.1682E+00
         0.3229E+00 0.3679E+00  0.1059E+00  0.1685E-01  0.6475E-03
         0.8673E-05);
>QUAD2   POINTS=(-0.4598E+01 -0.3560E+01 -0.2522E+01 -0.1484E+01
         -0.4453E+00 0.5930E+00  0.1631E+01  0.2670E+01  0.3708E+01
         0.4746E+01),
         WEIGHTS=(0.2996E-04  0.1300E-02  0.1474E-01  0.1127E+00
         0.3251E+00 0.3417E+00  0.1816E+00  0.2149E-01  0.1307E-02
         0.3154E-04);
>PRIOR   TSIGMA=(1.5(0)35);
>QUAD1   POINTS=(-0.4000E+01 -0.3111E+01 -0.2222E+01 -0.1333E+01
         -0.4444E+00 0.4444E+00  0.1333E+01  0.2222E+01  0.3111E+01
         0.4000E+01),
         WEIGHTS=(0.1190E-03  0.2805E-02  0.3002E-01  0.1458E+00
         0.3213E+00 0.3213E+00  0.1458E+00  0.3002E-01  0.2805E-02
         0.1190E-03);
>QUAD2   POINTS=(-0.4000E+01 -0.3111E+01 -0.2222E+01 -0.1333E+01
         -0.4444E+00 0.4444E+00  0.1333E+01  0.2222E+01  0.3111E+01
         0.4000E+01),
         WEIGHTS=(0.1190E-03  0.2805E-02  0.3002E-01  0.1458E+00
```

```
         0.3213E+00 0.3213E+00  0.1458E+00  0.3002E-01  0.2805E-02
         0.1190E-03);
>PRIOR   TSIGMA=(1.5(0)35);
```

Related topics

- QUAD command (see Section 3.2.11)

NEWTON keyword (optional)

Purpose

To specify the number of Gauss-Newton (Fisher-scoring) iterations following EM cycles.

If CYCLES=0 and NEWTON=0, item parameter estimates will be calculated from the classical item statistics from Phase 1 or from the starting values of the TEST command. The former will be corrected for guessing if the 3-parameter model is selected.

Format

```
NEWTON=n
```

Default

2.

Example

In this example, the value of NEWTON is increased to 4 in order to deal with a more informative aggregate-level data set.

```
>CALIB NQPT=30, CYCLES=24, NEWTON=4, CRIT=0.0050,
       RIDGE=(2, 0.8000, 2.0000), ACCEL=1.0000, TPRIOR, READPRI,
       NOFLOAT;
```

Related topics

- CALIB command: CYCLES keyword
- TEST command (see Section 3.2.15)
- **Setup** menu: **Item Analysis** dialog box (see Section 2.3)

NFULL keyword (optional)

Purpose

To specify that the Fisher-scoring steps for estimating item parameters use the full information matrix (if the number of items n is less than p) or the block-diagonal approximation to the information matrix (if n is greater than or equal to p).

Format

```
NFULL=p
```

Default

```
p=20.
```

Example

The NFULL keyword is used on the CALIB command to request the use of the full information matrix in the Newton steps for this data set where only 4 items were presented to subjects. In the absence of the NFULL keyword, the block diagonal approximation to the information matrix would have been used in this case, as NITEMS=4 is less than the default threshold of 20 items.

```
>CALIB   TPRIOR,SPRIOR,NFULL=4;
```

Related topics

- LENGTH command: NITEMS keyword (see Section 3.2.9)
- **Technical** menu: **Data Options** dialog box (see Section 2.5)

NOADJUST option (optional)

Purpose

In multiple-group applications, each group has its own latent distribution. To resolve the indeterminacy of origin and scale of measurement in the IRT analysis, the user can choose to set the mean and standard deviation to 0.0 and 1.0 in a reference group specified by the REF keyword of the CALIB command; alternatively, the user can choose to assign these values to the combined distributions weighted by their sample sizes.

BILOG-MG routinely rescales the origin and scale of the latent distribution (*i.e.*, linearly transforms the quadrature points) exactly to these values even in the case of one group. The item slopes and thresholds are then linearly transformed to match the adjusted scale.

This results in small differences between the values estimated in BILOG and BILOG-MG because the posterior latent distribution has mean and standard deviation equal to only approximately zero and one. To obtain the BILOG values (when all other conditions of estimation are identical), the user may include the option NOADJUST in the CALIB command, as in the example below.

Format

```
NOADJUST
```

Default

Rescaling the origin and scale of the latent distribution.

Example

In the syntax below, a single subtest is analyzed in a single group analysis. The NOADJUST option is used on the CALIB command to suppress the adjustment of the rescaling of the latent distribution.

```
EXAMPLE 16: TRADITIONAL IRT ANALYSIS OF A FIFTEEN-ITEM TEST
    PARAMETERS OF ITEMS 6 THROUGH 10 ARE FIXED
…
>CALIB   CYCLES=15,NEWTON=3,NQPT=11,DIAGNOS=1,NOADJUST;
```

Related topics

- CALIB command: REFERENCE keyword
- **Technical** menu: **Calibration Options** dialog box (see Section 2.5)

NORMAL option (optional)

Purpose

To specify the estimation of the means and standard deviations of the prior distributions of ability in the population of respondents by marginal maximum likelihood (under normal distribution assumptions) along with the item parameters. If NGROUP>1, separate means and standard deviations are estimated for each group.

Format

```
NORMAL
```

Default

Same as for EMPIRICAL.

Example

In this example of vertical equating of test forms over three grade levels, the distributions of ability are assumed to be normal at each grade level (NORMAL on the CALIB command).

```
>CALIB   NQPT=20,NORMAL,CYCLE=30,TPRIOR,NEWTON=2,CRIT=0.01,REFERENCE=2;
```

Related topics

- CALIB command: EMPIRICAL option
- INPUT command: NGROUP keyword (see Section 3.2.7)

NQPT keyword (optional)

Purpose

To specify the number of quadrature points in MML estimation for each group.

Format

```
NQPT=n
```

Default

20 for each group when NGROUP>1; 10 (otherwise).

Examples

In this example of a nonequivalent groups equating analysis, the number of quadrature points is set to 10 instead of the default of 20 for multiple-group analyses.

```
>CALIB NQPT=10, EMPIRICAL, CYCLES=25, NEWTON=5, CRIT=0.01, PLOT=0.05,
       REF=1, TPRIOR;
```

Here, the value of NQPT is increased to 30 in order to deal with a more informative aggregate-level data set.

```
>CALIB NQPT=30, CYCLES=24, NEWTON=4, CRIT=0.0050,
       RIDGE=(2, 0.8000, 2.0000), ACCEL=1.0000, TPRIOR, READPRI,
       NOFLOAT;
```

Related topics

- INPUT command: NGROUP keyword (see Section 3.2.7)
- **Setup** menu: **Item Analysis** dialog box (see Section 2.3)

NSD keyword (optional)

Purpose

To specify the range of the prior distribution(s) for the population(s) in standard deviation units.

Format

```
NSD=n
```

Default

8 standard deviation units (from –4.0 to 4.0).

PLOT keyword (optional)

Purpose

To specify the significance level for the goodness-of-fit of the item-response functions to be plotted. All items for which the significance level is below the real-number value (decimal fraction) provided will be plotted.

Format

```
PLOT=n
```

n = 0.0	no plots produced
n = 1.0	plots all items
n = 0.01	(for example) plots only those of poor-fitting items for which the significance level is less than 0.01

Default

```
0.0.
```

Examples

Plots of the item-response functions of all items for which the goodness-of-fit statistic is less than 0.05 are requested.

```
>CALIB NQPT=10, EMPIRICAL, CYCLES=25, NEWTON=5, CRIT=0.01, PLOT=0.05,
       REFERENCE=1, TPRIOR;
```

In this example of a 3-PL model, the PLOT keyword has been set to 1.0 so that all item response functions will be plotted

```
>CALIB NQPT=6, FLOAT, PLOT=1.0, CYCLES=15, NEWTON=3, ACCEL=0.5;
```

PRINT keyword (optional)

Purpose

To print provisional item parameter estimates at each iteration during the calibration phase. If PRINT=1, provisional item parameter estimates are printed; if PRINT=0 printing is suppressed.

Format

```
PRINT=n
```

Default

0.

Example

If the following CALIB command is used for a 2-group DIF analysis, only the information shown below is printed concerning the iterative process:

```
>CALIB   NQPT=10,CYCLES=15,CRIT=0.005,NEWTON=2,REFERENCE=1, PRINT=0;

[E-M CYCLES]
-2 LOG LIKELIHOOD =       3152.375
CYCLE      1;   LARGEST CHANGE=   0.17572
 -2 LOG LIKELIHOOD =       3128.806
CYCLE      2;   LARGEST CHANGE=   0.15440
 -2 LOG LIKELIHOOD =       3117.237
…
```

When the PRINT keyword is set to 1

```
>CALIB   NQPT=10,CYCLES=15,CRIT=0.005,NEWTON=2,REFERENCE=1, PRINT=1;
```

the output provided in the Phase 2 output file is expanded and parameter estimates are given for each group after each cycle. The output obtained for both groups after the third EM cycle is given below as an example.

```
QUADRATURE POINTS, POSTERIOR WEIGHTS, MEAN AND S.D.:
 EM CYCLE:      3
```

```
 GROUP   1    MALES   ;  ITEM PARAMETERS AFTER CYCLE   3
 ITEM      INTERCEPT    SLOPE    THRESHOLD   LOADING   ASYMPTOTE
 ----------------------------------------------------------------
 SP1       |   1.378  |   1.128  |  -1.222  |   0.748  |   0.000
           |   0.151* |   0.053* |   0.000* |   0.000* |   0.000*
           |          |          |          |          |
 SP2       |   0.686  |   1.128  |  -0.608  |   0.748  |   0.000
           |   0.137* |   0.053* |   0.000* |   0.000* |   0.000*
           |          |          |          |          |
 SP3       |  -0.938  |   1.128  |   0.831  |   0.748  |   0.000
           |   0.140* |   0.053* |   0.000* |   0.000* |   0.000*
           |          |          |          |          |
 SP4       |   0.649  |   1.128  |  -0.575  |   0.748  |   0.000
           |   0.136* |   0.053* |   0.000* |   0.000* |   0.000*
 ----------------------------------------------------------------
                                                * STANDARD ERROR
LARGEST CHANGE =    0.101366

GROUP   2    FEMALES ;  ITEM PARAMETERS AFTER CYCLE   3
 ITEM      INTERCEPT    SLOPE    THRESHOLD   LOADING   ASYMPTOTE
 ----------------------------------------------------------------
 SP1       |   1.795  |   1.128  |  -1.591  |   0.748  |   0.000
           |   0.144* |   0.053* |   0.000* |   0.000* |   0.000*
           |          |          |          |          |
 SP2       |   0.582  |   1.128  |  -0.516  |   0.748  |   0.000
           |   0.117* |   0.053* |   0.000* |   0.000* |   0.000*
           |          |          |          |          |
 SP3       |  -1.069  |   1.128  |   0.947  |   0.748  |   0.000
           |   0.124* |   0.053* |   0.000* |   0.000* |   0.000*
           |          |          |          |          |
 SP4       |  -0.200  |   1.128  |   0.177  |   0.748  |   0.000
           |   0.115* |   0.053* |   0.000* |   0.000* |   0.000*
 ----------------------------------------------------------------
                                                * STANDARD ERROR
     LARGEST CHANGE =    0.101366

 PARAMETER       MEAN  STN DEV
 -----------------------------------
 GROUP:   1  NUMBER OF ITEMS:     4
 THRESHOLD     -0.393    0.869
 GROUP:   2  NUMBER OF ITEMS:     4
 THRESHOLD     -0.246    1.078
 -----------------------------------

-2 LOG LIKELIHOOD =       3112.870
```

RASCH option (optional)

Purpose

To rescale the parameter estimates according to Rasch-model conventions. That is, all the slopes will be rescaled so that their geometric mean equals 1.0, and the thresholds will be rescaled so that their arithmetic mean equals 0.0. If the 1-parameter model has been specified, all slope parameters will therefore equal 1.0.

Because the threshold parameters are constrained in other ways in DIF and DRIFT analysis, the RASCH option cannot be used with these models. The posterior latent distribution displayed in Phase 2 is not rescaled in the Rasch convention.

Format

```
RASCH
```

Default

No Rasch rescaling.

Example

In the syntax for a single-group analysis shown below, a 1-parameter model is fitted to the data (NPARM=1 on GLOBAL command). Rasch rescaling is requested on the CALIB command through inclusion of the RASCH keyword, and all slope parameters will therefore equal 1.0.

```
>GLOBAL  DFNAME='EXAMPL04.DAT',NIDCH=5,NPARM=1;
…
>CALIB   CYCLE=10,TPRIOR,NEWTON=2,CRIT=0.01,RASCH;
```

Related topics

- GLOBAL command: NPARM keyword (see Section 3.2.5)
- **Technical** menu: **Calibration Options** dialog box (see Section 2.5)

READPRIOR option (optional)

Purpose

To specify that the prior distributions for selected parameters will be read from the ensuing PRIORS command(s). Otherwise, default priors will be used for these parameters.

Format

```
READPRIOR
```

Default

- thresholds: normal, mean = 0, SD = 2.0
- log slope: normal, mean = 0, SD = 0.5
- asymptote: beta, with parameters set so that the mean is 1/NALT with a weight of 20 observations of respondents who are marking randomly.

Example

In this example, the mean of the prior for the log slopes has been set to 0.5 by use of the READPRI option of the CALIB command and the following PRIORS commands.

```
>CALIB   NQPT=30, CYCLES=24, NEWTON=4, CRIT=0.0050,
         RIDGE=(2, 0.8000, 2.0000), ACCEL=1.0000, TPRIOR, READPRIOR,
         NOFLOAT;
>PRIORS1 SMU=(0.5000(0)8);
>PRIORS2 SMU=(0.5000(0)8);
```

Related topics

- INPUT command: NALT keyword (see Section 3.2.7)
- PRIORS command (see Section 3.2.10)
- **Technical** menu: **Item Parameter Prior Constraints** dialog box (see Section 2.5)

REFERENCE keyword (optional)

Purpose

To resolve the indeterminacy of the location and scale of the latent variable when NGROUP>1.

When the groups originally came from one population as, for example, in two-stage testing, REFERENCE should be set to 0. When the groups represent separate populations, REFERENCE should be set to the value of one of the group indicators. It specifies the reference group for the DIF model and the reference cohort for the DRIFT model.

Format

```
REFERENCE=n
```

n = 0	The mean and standard deviation of the combined estimated distributions of the groups weighted by their sample sizes are set to 0 and 1, respectively.
n > 0	The mean and standard deviation of group n are set to 0 and 1, respectively.

Default

1.

Examples

In this example of a nonequivalent groups equating analysis, the indeterminacy in location and scale of the distributions is resolved by using REF=1 to specify Group 1 as the reference group. This sets the mean and standard deviation of Group 1 to 0 and 1, respectively.

```
>CALIB NQPT=10, EMPIRICAL, CYCLES=25, NEWTON=5, CRIT=0.01, PLOT=0.05,
       REFERENCE=1, TPRIOR;
```

Here, the second group serves as the reference group in the calibration of the items.

```
>CALIB NQPT=20,NORMAL,CYCLE=30,TPRIOR,NEWTON=2,CRIT=0.01,REFERENCE=2;
```

Related topics

- INPUT command: DIF option (see Section 3.2.7)
- INPUT command: DRIFT option
- INPUT command: NGROUP keyword
- **Setup** menu: **General** dialog box (see Section 2.3)

RIDGE keyword (optional)

Purpose

To add a ridge constant (if a = 2) to the diagonal elements of the information matrix to be inverted during the EM cycles and Newton iterations. The ridge constant starts at the value 0 and is increased by b if the ratio of a pivot and the corresponding diagonal elements of the matrix is less than c.

The old ridge option can be invoked with the RIDGE=1 specification. It is provided so users may duplicate old results from BILOG. The present default is an improvement of the old method.

Format

```
RIDGE=(a, b, c)
```

Default

(2, 0.1, 0.01).

Example

This example emanates from an analysis of aggregate-level data that includes some fairly difficult items. A ridge constant of 0.8 is required for convergence as one of the subtests is exceptionally difficult.

Aggregate-level data typically have smaller slopes in the 0,1 metric than do person-level data. For this reason, the mean of the prior for the log slopes has been set to -0.5 by use of the READPRI option of the CALIB command and the following PRIOR commands.

```
>CALIB   NQPT=30, CYCLES=24, NEWTON=4, CRIT=0.0050,
         RIDGE=(2, 0.8000, 2.0000), ACCEL=1.0000, TPRIOR, READPRI,
         NOFLOAT;
>PRIORS1 SMU=(0.5000(0)8);
>PRIORS2 SMU=(0.5000(0)8);
```

Related topics

- CALIB command: CYCLES keyword
- CALIB command: NEWTON keyword

SELECT keyword (optional)

Purpose

To select, with a vector of ones and zeros, subtests for which item-parameter calibration is desired.

Format

SELECT=($n_1, n_2, \ldots, n_{NTEST}$)

where

- $n = 0$ Do not calibrate subtest i
- $n = 1$ Calibrate subtest i

Default

Calibrate all subtests.

Example

In this example with three subtests, only the second subtest is to be calibrated.

```
>TEST1 INUMBERS=(1(1)10);
>TEST2 INUMBERS=(11(1)30);
>TEST3 INUMBERS=(31(1)45);
(5A1,45A1)
>CALIB NQPT=10, CYCLES=25, NEWTON=5, SELECT=(0,1,0);
```

Related topics

- GLOBAL command: NTEST keyword (see Section 3.2.5)
- **Setup** menu: **Item Analysis** dialog box (see Section 2.3)

SPRIOR/NOSPRIOR option (optional)

Purpose

The presence of this option selects or suppresses prior distributions on the slope parameter.

Priors on the slope parameters are sometimes required to prevent Heywood cases.

Format

```
SPRIOR/NOSPRIOR
```

Default

- 1PL model, NOSPRIOR
- 2PL model, SPRIOR
- 3PL model, SPRIOR.

Examples

In the case of a 1PL model, no priors are used by default and thus the two CALIB commands

```
>CALIB NQPT=10, CYCLES=15, NOSPRIOR;
```

and

```
>CALIB NQPT=10, CYCLES=15;
```

are equivalent.

In order to assume a prior distribution on the slopes in the 1PL case, the CALIB command

```
>CALIB NQPT=10, CYCLES=15, SPRIOR;
```

may be used.

In a 2PL model, a prior is placed on the slopes by default and thus the commands

```
>CALIB NQPT=10, CYCLES=15, SPRIOR;
```

and

```
>CALIB NQPT=10, CYCLES=15;
```

are equivalent.

Related topics

- CALIB command: GPRIOR/NOGRPRIOR and TPRIOR/NOTPRIOR options
- **Setup** menu: **Item Analysis** dialog box (see Section 2.3)

TPRIOR/NOTPRIOR option (optional)

Purpose

To select prior distributions on the threshold parameters. Although extreme threshold values do not affect the estimation of ability adversely, a diffuse prior distribution on the thresholds will keep their estimates within a reasonable range during the estimation cycle.

Format

```
TPRIOR/NOTPRIOR
```

Default

- 1PL model, NOTPRIOR
- 2PL model, NOTPRIOR
- 3PL model, NOTPRIOR.

Examples

In this example of vertical equating of test forms, a prior is placed on the item thresholds by the addition of the TPRIOR option to the CALIB command.

```
>CALIB   NQPT=20,NORMAL,CYCLE=30,TPRIOR,NEWTON=2,CRIT=0.01,REFERENCE=2;
```

This example emanates from an analysis of aggregate-level data that includes some fairly difficult items. A prior on the thresholds is required for convergence as one of the subtests is exceptionally difficult.

```
>CALIB   NQPT=30, CYCLES=24, NEWTON=4, CRIT=0.0050,
         RIDGE=(2, 0.8000, 2.0000), ACCEL=1.0000, TPRIOR, READPRI,
         NOFLOAT;
>PRIORS1 SMU=(0.5000(0)8);
>PRIORS2 SMU=(0.5000(0)8);
```

In the case of a 1PL model, no priors are used by default and thus the two CALIB commands

```
>CALIB NQPT=10, CYCLES=15, NOSPRIOR;
```

and

```
>CALIB NQPT=10, CYCLES=15;
```

are equivalent. In order to assume a prior distribution on the slopes in the 1PL case, the CALIB command

```
>CALIB NQPT=10, CYCLES=15, SPRIOR;
```

may be used.

In a 2PL model, a prior is placed on the slopes by default and thus the commands

```
>CALIB NQPT=10, CYCLES=15, SPRIOR;
```

and

```
>CALIB NQPT=10, CYCLES=15;
```

are equivalent. In a 2PL model, the command

```
>CALIB NQPT=10, CYCLES=15, SPRIOR, TPRIOR;
```

indicates that an additional prior distribution should be assumed for the threshold parameters.

For a 3PL model, priors on slopes and asymptote parameters are assumed. To remove these priors, the CALIB command

```
>CALIB NQPT=10, CYCLES=15, NOSPRIOR, NOGPRIOR;
```

may be used.

In a 3PL model, to remove the default prior distribution on the asymptote parameters and use a prior distribution on the thresholds instead, use

```
>CALIB NQPT=10, CYCLES=15, SPRIOR, NOGPRIOR, TPRIOR;
```

Related topics

- CALIB command: SPRIOR/NOSPRIOR and GPRIOR/NOGPRIOR options
- **Setup** menu: **Item Analysis** dialog box (see Section 2.3)

3.2.3 COMMENT command

(Optional)

Purpose

To enter one or more lines of explanatory remarks into the program output stream. This line and all subsequent lines preceding the GLOBAL command will be printed in the initial output stream. The maximum length of each line is 80 characters. A semicolon to signal the end of the command is not needed.

Format

```
>COMMENT
…text…
…text…
```

Example

```
EXAMPLE 4
SIMULATED RESPONSES TO TWO 20-ITEM PARALLEL TEST FORMS
>COMMENT
This example illustrates the equating of equivalent groups with the BILOG-
MG program. Two parallel test forms of 20 multiple-choice items were ad-
ministered to two equivalent samples of 200 examinees drawn from the same
population. There are not common items between the forms.
>GLOBAL DFNAME='EXAMPL04.DAT',NIDCH=5,NPARM=2;
```

Default

No comments.

Related topics

- GLOBAL command (see Section 3.2.5)

Setup menu: General dialog box (see Section 2.3)

3.2.3 DRIFT command

(Required only if DRIFT is specified in the INPUT command)

Purpose

To provide the maximum degree of the polynomial item parameter drift model and a vector of time points, $n_1, n_2, ..., n_n$.

Format

```
>DRIFT MAXPOWER=a, MIDPOINT=(n1,n2,...,nn);
```

Default

No DRIFT command.

Example

```
>DRIFT MAXPOWER=2, MIDPOINT=(-5,-2,0,2,4);
```

Related topics

- INPUT command: DRIFT option (see Section 3.2.7)

MAXPOWER keyword

(optional)

Purpose

To specify the maximum degree of the drift polynomial included in the model. The maximum degree must be less than the number of groups.

Format

```
MAXPOWER=n
```

Default

```
NGROUP-1.
```

Related topics

- INPUT command: NGROUP keyword (see Section 3.2.7)

MIDPOINT keyword

(optional)

Purpose

To specify a vector of time points (or midpoints of time intervals).

Format

`MIDPOINT=(` $n_1, n_2, \ldots, n_n$ `)`

Default

```
(1, 2, …, NGROUP-1)
```

Related topics

- INPUT command: NGROUP keyword

3.2.4 FORM command

(Required only if the NFORM keyword appears in the INPUT command)

Purpose

To supply the order of the item responses in the data records. Each FORM command gives the number of items in the form and lists the items in the order in which the item responses appear on the data records for that form. The items may be listed by name or number, but not by both.

When NFORMS > 1, the FORM command requires a form number in the data record. The form numbers must range in value from 1 to the number of forms. The form indicator field follows the case ID field and is INTEGER in the variable format statement. Because the same format statement is used to read the data records for all forms, the item responses, the case ID and weight, and the form and group indicators must occupy the same columns on all records. If the forms are of unequal length, the size of the item-response field on the format statement should equal the number of items in the longest form.

The order of the several FORM commands corresponds to the number of the respective form.

Format

```
>FORM LENGTH=n, INUMBERS=(list), INAMES=(list);
```

Default

None.

Example

Form 1 consists of items 1, 2, 3, 4, and 6, and form 2 consists of items 1, 6, 7, 8, 9, and 10. The data records are as follows:

```
SUBJECT001 1 21321
SUBJECT002 2 513122
…
SUBJECT999 1 21422
```

Responses to item 1 appear in column 14 of the data records for form 1 and at the end of the data records for form 2. The FORM commands and format statement are as follows:

```
>FORM1 LENGTH=5, INUMBERS=(1(1)5);
>FORM2 LENGTH=6, INUMBERS=(6(1)10,1);
(10A1, 1X, I1, 1X, 6A1)
```

Related topics

- INPUT command: NFORM keyword (see Section 3.2.7)
- Variable format statement (see Section 3.2.16)
- **Setup** menu: **General** dialog box (see Section 2.3)
- **Setup** menu: **Item Analysis** dialog box

INAMES keyword (optional)

Purpose

To specify the list of item names, as specified in the ITEMS command, in the order in which the item response appear on the data records for FORM*j*.

Format

INAMES=($n_1, n_2, ..., n_{LENGTH}$)

Default

When NFORM = 1, the sequence of items specified on the ITEMS command. When NFORM > 1, no sequence is specified.

Example

Assume, in the previous example, that the command

```
>ITEMS INAMES=(I1(1)I10);
```

appears earlier in the command file to give the name Ix to item x. Then the FORM1 command could be replaced with

```
>FORM1 LENGTH=5, INAMES=(I1(1)I4,I6);
```

Note that if the item names are in a sequence, they can be specified using the variable list format "first (increment) last", as "I1(1)I4" is used here to specify items 1 through 4.

Related topics

- ITEMS command
- INPUT command: NFORM keyword
- **Setup** menu: **General** dialog box

INUMBERS keyword (optional)

Purpose

To provide the list of item numbers, as specified in the ITEMS command, in the order in which the item response appear on the data records for FORM*j*.

Format

INUMBERS=($n_1, n_2, ..., n_{LENGTH}$)

Default

When NFORM = 1, the sequence of items specified on the ITEMS command. When NFORM > 1, none.

Related topics

- ITEMS command (see Section 3.2.8)
- INPUT command: NFORM keyword (see Section 3.2.7)
- **Setup** menu: **Item Analysis** dialog box (see Section 2.3)

LENGTH keyword (required)

Purpose

The number of items in FORM.

Format

```
LENGTH=n
```

Default

NTOTAL when NFORM = 1, none when NFORM > 1.

Related topics

- INPUT command: NFORM and NTOTAL keywords (see Section 3.2.7)
- ITEMS command (see Section 3.2.8)
- **Setup** menu: **Item Analysis** dialog box (see Section 2.3)

3.2.5 GLOBAL command

(Required)

Purpose

To supply input filenames and other information used in the three phases of the program. The GLOBAL keywords DFNAME, MFNAME, CFNAME, and IFNAME enable the user to assign specific names to the program's input files. A filename must be not more than 128 characters long and may include a drive prefix, a path name, and an extension. The filename must be enclosed in single quotes. Note that each line of the command file has a maximum length of 80 characters. If the filename does not fit on one line of 80 characters, the remaining characters should be placed on the next line, starting at column 1.

Format

```
>GLOBAL DFNAME=n, MFNAME=n, CFNAME=n, IFNAME=n, NPARM=n, NWGHT=n, NTEST=n,
        NVTEST=n, PRNAME=n, LOGISTIC, OMITS, SAVE;
```

Example

```
>GLOBAL DFNAME='EXAMPL04.DAT', NPARM=2;
```

Related topics

- **Data** menu: **Examinee Data** dialog box (see Section 2.4)
- **Data** menu: **Group-Level Data** dialog box
- **Save** menu (see Section 2.6)
- **Setup** menu: **General** dialog box (see Section 2.3)
- **Setup** menu: **Test Scoring** dialog box

CFNAME keyword (optional)

Purpose

To supply the name of the previously created calibration file (if any) to be read in. If data are read from a previously generated calibration file, DFNAME must not appear, and TYPE=0 must appear in the INPUT command.

The PARM keyword of the SAVE command must be specified to save updated parameter estimates to an external file.

Format

```
CFNAME=<'filename'>
```

Example

In a previous run, a calibration file was created as shown below. The calibration file was saved to **exampl03.cal** using the CALIB keyword on the SAVE command. Note that a calibration file will be created only if the SAMPLE keyword is also specified on the INPUT command, with a number less that the total number of examinees.

```
EXAMPLE:
CREATING A CALIBRATION FILE
>COMMENT
>GLOBAL  DFNAME='EXAMPL03.DAT',NPARM=2, SAVE;
>SAVE    CALIB='EXAMPL03.CAL';
>LENGTH  NITEMS=(45);
>INPUT   NTOTAL=45,SAMPLE=2000,NGROUP=2,KFNAME='EXAMPL03.DAT',NIDCHAR=5,
         NALT=5,NFORM=2,TYPE=1;
```

The previously created calibration file is now used as data source through the use of the CFNAME keyword on the GLOBAL command. Note that the TYPE keyword on the INPUT command is now set to 0, compared to 1 previously. The updated item parameter estimates are saved to the file **latest.prm** using the PARM keyword on the SAVE command.

```
EXAMPLE:
USING A MASTER FILE AS INPUT
>COMMENT
>GLOBAL  CFNAME='EXAMPL03.CAL',NPARM=2, SAVE;
>SAVE    PARM='LATEST.PRM';
>LENGTH  NITEMS=(45);
>INPUT   NTOTAL=45,SAMPLE=2000,NGROUP=2,NIDCHAR=5,
         NALT=5,NFORM=2,TYPE=0;
```

Related topics

- GLOBAL command: DFNAME keyword (see Section 3.2.5)
- INPUT command: TYPE keyword (see Section 3.2.7)
- SAVE command: CALIB keyword (see Section 3.2.13)

DFNAME keyword (optional)

Purpose

To supply the name of the raw data file that contains the original data. The format for this file is described in the section on input and output files.

Format

```
DFNAME=<'filename'>
```

Notes

The path to and filename of this file may be longer than 80 characters. However, as the maximum length of any line in the command file is 80 characters, multiple lines may be used. It is important to continue up to and including the 80th column when specifying a long path and filename.

For example, suppose the data file **exampl06.dat** is in a folder named:

```
C:\PROGRAM FILES\ITEM RESPONSE THEORY\IRT_2002\MARCH20\BILOG-MG-
VERSION1.2\EXAMPLES
```

The correct way to enter this information in the command file is to enclose the name and path in single quotes, and continue until column 80 is reached. Then proceed in column 1 of the next line as shown below:

```
>GLOBAL DFNAME='C:\PROGRAM FILES\ITEM RESPONSE THEORY\IRT_2002\MARCH20\BILOG-MG

-VERSION1.2\EXAMPLES\EXAMPL06.DAT', NTEST=1, NVTEST=1, NPARM=2, SAVE;
```

If the data are stored in the same folder as the command file, it is sufficient to type

```
DFNAME='EXAMPL06.DAT'
```

Examples

This example shows the use of the external data file **exampl03.dat**.

```
>GLOBAL  DFNAME='EXAMPL03.DAT';
>LENGTH  NITEMS=(45);
>INPUT  NTOTAL=45,KFNAME='EXAMPL03.DAT',NIDCHAR=5,
```

Note that this file is referenced on both the GLOBAL command (DFNAME keyword) and on the INPUT command (KFNAME keyword). This indicates that the answer key for correct responses is given at the top of the data file, as shown below:

```
ANSWER KEY  111111111111111111111111
person1     111111111222121221111121
person2     221111121222222222255222
```

Related topics

- **Data** menu: **Examinee Data** dialog box (see Section 2.4)
- **Data** menu: **Group-Level Data** dialog box

- Input files (see Section 3.2.18)
- Output files
- Variable format statement (see Section 3.2.16)

IFNAME keyword (optional)

Purpose

To supply the name of the previously created item parameter file (if any) to be used as input.

The PARM keyword of the SAVE command must be specified to save updated parameter estimates to an external file.

Format

```
IFNAME=<'filename'>
```

Example

The previously created parameter file **exampl03.par** is used as data source through the use of the IFNAME keyword on the GLOBAL command. The updated item parameter estimates are saved to the file **latest.par** using the PARM keyword on the SAVE command.

```
EXAMPLE:
USING A ITEM PARAMETER FILE AS INPUT
>COMMENT
>GLOBAL  IFNAME='EXAMPL03.PAR',NPARM=2, SAVE;
>SAVE    CALIB='LATEST.PAR';
>LENGTH  NITEMS=(45);
>INPUT   NTOTAL=45,SAMPLE=2000,NGROUP=2,NIDCHAR=5,
         NALT=5,NFORM=2;
```

Related topics

- GLOBAL command: IFNAME keyword
- INPUT command: TYPE keyword (see Section 3.2.7)
- SAVE command: PARM keyword (see Section 3.2.13)
- **Setup** menu: **Test Scoring** dialog box (see Section 2.3)

LOGISTIC option (optional)

Purpose

To assume the natural metric of the logistic response function in all calculations. Otherwise, the logit is multiplied by D = 1.7 to obtain the metric of the normal ogive model.

Format

```
LOGISTIC
```

Default

Normal ogive model.

Examples

For the 2-parameter model requested in this first GLOBAL command, the natural metric of the logistic response function is assumed:

```
>GLOBAL NPARM=2, LOGISTIC, DFNAME='EXAMPLE.DAT';
```

while a similar normal ogive model can be obtained by using the command:

```
>GLOBAL NPARM=2, DFNAME='EXAMPLE.DAT';
```

Related topics

- **Setup** menu: **General** dialog box (see Section 2.3)

MFNAME keyword (optional)

Purpose

To supply the name of a previously created master file to be read in. If data are read from a previously prepared master file, DFNAME must not appear, and TYPE=0 must appear in the INPUT command. The PARM keyword of the SAVE command may be specified to save updated parameter estimates to an external file.

Format

```
MFNAME='filename'
```

Example

The previously created master file **exampl03.mas** is used as data source through the use of the MFNAME keyword on the GLOBAL command. Note that the TYPE keyword on the INPUT command is now set to 0.

```
EXAMPLE:
USING A MASTER FILE AS INPUT
>GLOBAL  MFNAME='EXAMPL03.MAS',NPARM=2;
>LENGTH  NITEMS=(45);
>INPUT   NTOTAL=45,SAMPLE=2000,NGROUP=2,NIDCHAR=5,
         NALT=5,NFORM=2,TYPE=0;
```

Related topics

- INPUT command: TYPE keyword (see Section 3.2.7)
- SAVE command: MASTER keyword (see Section 3.2.13)

NPARM keyword (optional)

Purpose

To indicate the number of item parameters in the model:

- 1: 1-parameter logistic model
- 2: 2-parameter logistic model
- 3: 3-parameter logistic model

Format

```
NPARM=n
```

Default

```
NPARM=2.
```

Examples

The following GLOBAL commands are used to request a 1PL, 2PL and 3PL model respectively.

```
>GLOBAL NPARM=1, NWGHT=3, LOGISTIC;

>GLOBAL DFNAME='EXAMPL03.DAT', NPARM=2;

>GLOBAL NPARM=3, DFNAME=EXAMPL07.DAT;
```

Related topics

- **Setup** menu: **General** dialog box (see Section 2.3)

NTEST keyword (optional)

Purpose

To indicate the number of subtests.

Format

```
NTEST=n
```

Default

```
NTEST=1.
```

Examples

In the GLOBAL command below, the NTEST keyword is used to indicate that two subtests are used. Note the two TEST commands in the syntax. The LENGTH command is used to indicate the length of the subtests.

```
>GLOBAL NPARM=3, NTEST=2, DFNAME='EXAMPL08.DAT';
>LENGTH NITEM=(8,8);
>INPUT NTOTAL=16;
>ITEMS INUMBER=(1(1)16), INAMES=(N1(1)N8,A1(1)A8);
>TEST1 TNAME=NUMCON, INUMBER=(1(1)8);
>TEST2 TNAME=ALGCON, INUMBER=(9(1)16);
```

Related topics

- GLOBAL command: NVTEST keyword
- LENGTH command (see Section 3.2.9)
- **Setup** menu: **General** dialog box (see Section 2.3)
- TEST command

NVTEST keyword (optional)

Purpose

To indicate the number of subtests with variant items.

Format

```
NVTEST=n
```

Default

```
NVTEST=0.
```

Example

In the example below, both a main and variant test are used. In this case, NTEST is set to 1 to indicate the main test, and the NVTEST keyword is used to indicate the presence of a variant test. The first test command is that for the main test, while items for the variant test are selected by name in the next TEST command (here named TESTV purely for convenience). There are 20 main test items and 4 variant test items, selected from a total of 50 items in the data file. The LENGTH command is used to indicate the length of the subtests.

```
>GLOBAL NPARM=3, NTEST=1, NVTEST=1, DFNAME='EXAMPL06.DAT';
>LENGTH NITEM=24, NVARIANT=4;
>INPUT  NTOTAL=50, NIDCHAR=11;
>ITEMS  INUMBERS=(1(1)50), INAMES=(I26(1)I75);
>TESTM  TNAME=MAIN, INAMES=(I26, I27, I28, I29, I31, I33, I34, I35, I36,
        I38, I39, I47, I48, I49, I50, I54, I60, I64, I68, I72);
>TESTV  TNAME=VARIANT, INAMES=(I53, I59, I69,I73);
(11A1,T39,25A1/T13,25A1)
```

Related topics

- GLOBAL command: NTEST keyword
- LENGTH command (see Section 3.2.9)
- TEST command (see Section 3.2.15)

NWGHT keyword (optional)

Purpose

To specify the weighting of response records. A value larger than 0 is required when the data are INPUTin the form of response patterns and frequencies, or when the sampling procedure requires the use of case weights. The data file (TYPE) in the INPUT command must also be set appropriately. See the information on format statements for the data format with weights in Section 3.2.16.

Format

```
NWGHT=n
```

The type of weighting associated with valid values of n is:

- 0: none
- 1: for classical item statistics only
- 2: for IRT item calibration only
- 3: for both statistics and calibrations.

Default

```
NWGHT=0.
```

Example

In this example, the data are accumulated into answer patterns. TYPE=2 and NWGHT=3 are included to indicate this form of data.

```
>GLOBAL NPARM=1, NWGHT=3, LOGISTIC;
>LENGTH NITEMS=4;
>INPUT  NTOTAL=4, NGROUPS=2, DIF, NIDCHAR=2, TYPE=2;
```

Related topics

- INPUT command: TYPE keyword (see Section 3.2.7)
- Variable format statement (see Section 3.2.16)

OMITS option (optional)

Purpose

To specify that omits are treated as fractionally correct when the 3-parameter model is employed. The fraction is the reciprocal of the number of the alternatives in the multiple-choice items (see the NALT keyword on the INPUT command, Section 3.2.7). Also see Section 3.2.18 for more information on the specification of an omit key using the OFNAME keyword on the INPUT command.

Format

```
OMITS
```

Default

Omitted responses are treated as incorrect.

Examples

For the following 3-parameter model, an omitted response will be scored fractionally correct with the fraction equal to 1/5 (NALT=5). The omit response key can be found in the data file.

```
>GLOBAL  NPARM=3, LOGISTIC, DFNAME='EXAMPLE.DAT', OMITS;
>LENGTH  NITEMS=40;
>INPUT   NTOTAL=40, OFNAME='EXAMPLE.DAT', NALT=5;
```

In this example the omitted response will be scored fractionally correct with fraction 1/4. The key for omitted responses can be found in a separate, external file.

```
>GLOBAL  NPARM=3, LOGISTIC, DFNAME='EXAMPLE.DAT', OMITS;
>LENGTH  NITEMS=40;
>INPUT   NTOTAL=40, OFNAME='OMITKEY.DAT', NALT=4;
```

Related topics

- INPUT command: NALT and OFNAME keywords(see Section 3.2.7)
- Input files (see Section 3.2.18)
- **Setup** menu: **General** dialog box (see Section 2.3)

PRNAME keyword (optional)

Purpose

To specify the name of the file from which the provisional (i.e. starting) values of parameters of selected items will be obtained. The values are read in space-delimited, free-format form.

Format

```
PRNAME=<'filename'>
```

Contents:

The contents of the file are as follows:

Line 1:

The number of selected items in each subtest.

Remaining lines:

The serial position of each item selected from the corresponding subtest, followed by the slope, threshold, and chance success (guessing) probability of the item. If a two-parameter model is assumed, the latter should be entered as 0.

Default

None.

Example

```
5  5
 5   1.0  0.0   0.333
10   1.0  0.0   0.333
15   1.0  0.0   0.333
25   1.0  0.0   0.333
30   1.0  0.0   0.333
 5   1.1  0.5   0.233
10   1.1  0.5   0.233
15   1.1  0.5   0.233
25   1.1  0.5   0.233
30   1.1  0.5   0.233
```

Provisional values will be assigned to five items in each of two subtests. In each subtest, the 5-th, 10-th, 15-th, 25-th, and 30-th item will be assigned the values in the corresponding line.

The following is an example of a command file that will INPUTthese values. Note that PRINT has been set to 1 on the CALIB command to print the item parameters at cycle zero and show the assigned values.

```
EXAMPLE 15:
ASSIGNED STARTING VALUES FOR TWO SUBTESTS
>GLOBAL  DFNAME='EXAMPL03.DAT',PRNAME='EXAMPL15.PRM',NPARM=2,
         NTEST=2,SAVE;
>SAVE    PDISTRIB='EXAMPL15.PST',SCORE='EXAMPL15.SCO';
>LENGTH  NITEMS=(35,35);
>INPUT   NTOTAL=45,SAMPLE=2000,NGROUP=2,KFNAME='EXAMPL03.DAT',
         NALT=5,NFORMS=2,NIDCHAR=5;
>ITEMS   INUMBERS=(1(1)45),
         INAME=(C01(1)C45);
>TEST1   TNAME=SUBTEST1, INAME=(C01(1)C15,C21(1)C40);
>TEST2   TNAME=SUBTEST2, INAME=(C06(1)C25,C31(1)C45);
>FORM1   LENGTH=25,INUMBERS=(1(1)25);
>FORM2   LENGTH=25,INUMBERS=(21(1)45);
>GROUP1  GNAME=POP1,LENGTH=25,INUMBERS=(1(1)25);
>GROUP2  GNAME=POP2,LENGTH=25,INUMBERS=(21(1)45);
(T28,5A1,T25,I1,T25,I1/45A1)
>CALIB   IDIST=1,EMPIRICAL,NQPT=11,CYCLE=10,TPRIOR,NEWTON=1,
         CRIT=0.01,REF=1,NOFLOAT,PRINT=1;
>SCORE   IDIST=3,RSCTYPE=3,INFO=1,YCOMMON,POP,NOPRINT,MOMENTS;
```

Related topics

- CALIB command: PRINT keyword (see Section 3.2.2)
- **Setup** menu: **Test Scoring** dialog box (see Section 2.3)

SAVE option (optional)

Purpose

To indicate that a SAVE command will follow the GLOBAL command.

Format

```
SAVE
```

Default

No SAVE command to follow.

Example

In the syntax below, the item parameters and scale scores are saved to file through the use of the SCORE and PARM keywords on the SAVE command. Note that, in order to use the SAVE command, the SAVE option is added to the GLOBAL command.

```
>GLOBAL  DFNAME='EXAMPLE.DAT', NPARM=2, SAVE;
>SAVE    SCORE='EXAMPLE.SCO', PARM='EXAMPLE.PAR';
>LENGTH  NITEMS=(40);
```

Related topics

- SAVE command (see Section 3.2.13)
- **Save** menu (see Section 2.6)

3.2.6 GROUP command

(Required when NGROUP > 1 on the INPUT command)

Purpose

To specify information about the items in each particular group. When the NGROUP keyword on the INPUT command is greater than one, that same number of GROUP commands must follow the FORM commands. Each GROUP command specifies the group's name, the length of the group's form, and the items included in that form. Items may be identified by name or number, but not by both.

The GROUP command requires a group number in the data record. The group numbers must range in value from 1 to the number of groups. If NFORM > 1, the group indicator field follows the form indicator field. If NFORM = 1, the group indicator field follows the case ID field. The group indicator field is INTEGER in the variable format statement. If the subtest is personalized (the option PERSONAL is present in the INPUT command) there are NTEST group indicators for each subject.

The order of the several GROUP commands corresponds to the number of the respective group. If the same items are administered to all groups, the INUMBERS and INAMES lists are the same as those in the ITEMS command.

Format

```
>GROUP GNAME=n, LENGTH=n, INUMBERS=(list), INAMES=(list);
```

Default

No groups assumed.

Example

If the form(s) for group 1 consists of items 1, 2, 4, and 5, and the form(s) for group 2 consists of items 3 through 8, then the corresponding group commands are as follows:

```
>GROUP1 GNAME=GROUP1, LENGTH=4, INUMBERS=(1,2,4,5);
>GROUP2 GNAME=GROUP2, LENGTH=6, INUMBERS=(3(1)8);
```

Related topics

- GLOBAL command: NTEST and NVTEST keywords (see Section 3.2.5)
- INPUT command: NFORM keyword (see Section 3.2.7)
- INPUT command: PERSONAL option
- ITEMS command (see Section 3.2.8)
- LENGTH command (see Section 3.2.9)
- **Setup** menu: **General** dialog box (see Section 2.3)
- **Setup** menu: **Item Analysis** dialog box
- Variable format statement (see Section 3.2.16)

GNAME keyword (optional)

Purpose

To specify the name of GROUP*k* (up to eight characters).

Format

```
GNAME=character string
```

Default

Blanks.

Related topics

- **Setup** menu: **General** dialog box (see Section 2.3)

INAMES keyword (optional)

Purpose

To specify the list of item names, as specified in the ITEMS command, for all items in all forms administered to GROUP*k*.

Format

INAMES=($n_1, n_2, ..., n_{LENGTH}$)

Default

All names specified in the ITEMS command.

Example

Assume, in the previous example, that the command

```
>ITEMS INAMES=(I1(1)I8)
```

appears earlier in the command file to give the name Ix to item x. Then the two GROUP commands could be replaced with

```
>GROUP1 GNAME=GROUP1, LENGTH=4, INAMES=(I1,I2,I4,I5)
>GROUP2 GNAME=GROUP2, LENGTH=6, INAMES=(I3(1)I8)
```

Note the use of the list notation in the GROUP2 command to specify items I3 through I8.

Related topics

- GROUP command: LENGTH keyword
- INPUT command: NTOTAL keyword (see Section 3.2.7)
- ITEMS command (see Section 3.2.8)
- **Setup** menu: **General** dialog box (see Section 2.3)

INUMBERS keyword (optional)

Purpose

To provide a list of item numbers, as specified in the ITEMS command, for all items in all forms administered to GROUP*k*.

Format

INUMBERS=($n_1, n_2, ..., n_{LENGTH}$)

Default

All items specified in the ITEMS command.

Example

In the following example, the INUMBERS keywords specify the item list for each group. Note, again, the use of the "sequence" notation in the second statement to specify items 3 through 8.

```
>GROUP1 GNAME=GROUP1, LENGTH=4, INUMBERS=(1,2,4,5);
>GROUP2 GNAME=GROUP2, LENGTH=6, INUMBERS=(3(1)8);
```

Related topics

- GROUP command: LENGTH keyword
- INPUT command: NTOTAL keyword (see Section 3.2.7)
- ITEMS command (see Section 3.2.8)
- **Setup** menu: **Item Analysis** dialog box (see Section 2.3)

LENGTH keyword (optional)

Purpose

To specify the number of items in the test form(s) for GROUP*k* .

Format

```
LENGTH=n
```

Default

```
NTOTAL.
```

Example

In the following example, the LENGTH keyword in each GROUP command specifies the number of items for each group.

```
>GROUP1 GNAME=GROUP1, LENGTH=4, INUMBERS=(1,2,4,5);
>GROUP2 GNAME=GROUP2, LENGTH=6, INUMBERS=(3(1)8);
```

Related topics

- INPUT command: NTOTAL keyword (see Section 3.2.7)
- **Setup** menu: **General** dialog box (see Section 2.3)

3.2.7 INPUT command

(Required)

Purpose

To provide the information which describes the raw data file. One or more variable format statements describing the layout of the data must follow the FORM, GROUP, or DRIFT command, if present.

The keywords KFNAME, NFNAME, and OFNAME enable the user to assign specific names to the program's input files. A filename must be no more than 128 characters long and may include a drive prefix, a path name, and an extension. The filename must be enclosed in single quotes. Note that each line of the command file has a maximum length of 80 characters. If the filename does not fit on one line of 80 characters, the remaining characters should be placed on the next line, starting at column 1.

Format

```
>INPUT NTOTAL=n, NFMT=n, TYPE=n, SAMPLE=n, NALT=n, NIDCHAR=n, TAKE=n,
       NGROUP=n, NFORM=n, ISEED=n, DIAGNOSE=n, KFNAME=n, NFNAME=n,
       OFNAME=n, DRIFT, DIF, PERSONAL, EXTERNAL;
```

Examples

In the following example, responses from two groups are analyzed. There are two forms of a 25-item multiple-choice examination, with 5 items in common. In total, the responses of a sample of 2000 respondents to the 45 items are considered.

```
>INPUT NTOTAL=45, SAMPLE=2000, NGROUP=2, NFORM=2;
```

The INPUT command below is used to request a DIF analysis on 4 items administered to two groups.

```
>INPUT NTOTAL=4, DIF, NGROUP=2;
```

Related topics

- **Data** menu: **Examinee Data** dialog box (see Section 2.4)
- **Data** menu: **Group-Level Data** dialog box
- **Data** menu: **Item Keys** dialog box
- DRIFT command (see Section 3.2.3)
- FORM command (see Section 3.2.4)
- GROUP command (see Section 3.2.6)
- **Setup** menu: **General** dialog box (see Section 2.3)
- **Technical** menu: **Data Options** dialog box (see Section 2.5)
- Variable format statement (see Section 3.2.16)

DIAGNOSE keyword (optional)

Purpose

To specify a level of diagnostic printout for Phase 1. Larger values of n give increasing diagnostic output.

Format

```
DIAGNOSE=n
```

Default

No diagnostic printout.

Related topics

- Phase 1: INPUT(see Section 1.2)

DIF option (optional)

Purpose

To specify a differential equation modeling (DIF) analysis for multiple groups, which assumes common slopes and guessing parameter for all groups.

Format

```
DIF
```

Default

No DIF analysis.

Example

In the syntax below, a 1-parameter DIF model is fitted to data from two groups of examinees. DIF parameters are saved to the file **exampl01.dif** through use of the SAVE option on the GLOBAL command and the DIF option on the SAVE command.

```
>GLOBAL  NPARM=1,LOGISTIC,SAVE;
>SAVE    PARM='EXAMPL01.PAR',DIF='EXAMPL01.DIF';
>LENGTH  NITEMS=4;
>INPUT   NTOTAL=4,NGROUPS=2,DIF,NIDC=2;
```

Related topics

- DRIFT option
- GROUP command (see Section 3.2.6)
- INPUT command: NGROUP keyword
- SAVE command: DIF keyword (see Section 3.2.13)
- **Setup** menu: **General** dialog box (see Section 2.3)

DRIFT option (optional)

Purpose

To specify an item parameter drift model for multiple groups. A DRIFT command must also appear after the GROUP commands.

Format

```
DRIFT
```

Default

No DRIFT model.

Example

In the syntax below, a 2-parameter DRIFT model is fitted to data from two groups of examinees. DRIFT parameters are saved to the file **exampl01.drf** by using the SAVE option on the GLOBAL command and the DRIFT option on the SAVE command.

```
>GLOBAL  NPARM=1,LOGISTIC,SAVE;
>SAVE    PARM='EXAMPL01.PAR',DRIFT='EXAMPL01.DRF';
>LENGTH  NITEMS=4;
>INPUT   NTOTAL=4,NGROUPS=2,DRIFT,NIDC=2;
```

Related topics

- DRIFT command (see Section 3.2.3)
- GROUP command (see Section 3.2.6)
- INPUT command: NGROUP keyword
- SAVE command: DRIFT keyword (see Section 3.2.13)
- **Setup** menu: **General** dialog box (see Section 2.3)

EXTERNAL keyword (optional)

Purpose

To specify the computation of the item parameters with respect to an external variable, the values of which are supplied in the data records, rather than to a latent variable inferred from the item responses. When item parameters are estimated in this way and used to score test data of any other groups of examinees, the resulting scores are the best predictors of the ability measured by the external variable.

In each record of the calibration data, each test in the analysis must be represented by a value of the external variable and its corresponding standard error. These two quantities for each test in the data record must precede the item responses in the same order as the tests appear in their successive command lines. The columns of the data records devoted to these pairs of scores and standard errors must be identified in the INPUTvariable format statement.

Format

```
EXTERNAL
```

Default

Calibration with respect to a latent variable inferred from the item responses.

Example

Suppose a group of students took an end-of-term reading test and math test routinely administered to all students in a metropolitan school district. Suppose these students were also part of the sample for a state assessment of reading and math achievement. If scores and standard errors on the assessment tests for these students were available to the district, the district test could be calibrated to best predict the state reading and math scores of students of all students in the district. For this purpose, the state test results would serve as the external variables for calibrating items of the local tests to predict the state assessment's scores.

For the sake of generality, suppose also that there are three random parallel forms of the district tests and that these forms are assigned at random to students in two successive school grades. Then there will be two groups of students in the analysis and the record layout of the data might be the following:

- Columns 1-4: Student ID
- Column 6: test form number
- Column 8: grade group number
- Columns 10-13: state reading test score
- Columns 15-18 state reading test standard error
- Columns 20-23 state math test score
- Columns 25-28 state math test standard error
- Columns of 30-59: local reading test item responses
- Columns 60-89: local math test item responses

The format statement for reading the data records would be

```
(4A1,1X,I1,1X,I1,2(1X,F4.1,1X,F4.1),1X,60A1)
```

and the item parameter file from the calibration could be saved for use in scoring other students.

Related topics

- **Data, Examinee Data /Data, Group-Level Dat**a dialog boxes (see Section 2.4)

- Variable format statement (see Section 3.2.16)

ISEED keyword (optional)

Purpose

To specify the seed for the random number generator used for sampling subjects.

By default, the same seed will always be used for sampling subjects when the SAMPLE keyword on the INPUT command is used. ISEED may be used to change the seed, thus producing a different random sample of subjects.

Format

```
ISEED=n
```

Default

```
ISEED=1.
```

Related topics

- INPUT command: SAMPLE and TAKE keywords
- **Technical** menu: **Data Options** dialog box (see Section 2.5)

KFNAME keyword (optional)

Purpose

To specify the name of the file which contains the answer key. This key consists of the correct response alternative for each item, in the same format as the corresponding response records. Any single ASCII character can be used as a response alternative. If the answer key is in the same file as the item response data, the key must precede the first response record. If KFNAME does not appear on the INPUT command, then the data are assumed to be scored 1 for correct and 0 for incorrect.

When NFORM > 1, separate answer, not-presented, and omit keys must be specified for each form in the order of the forms to which they apply. Again, if they are in the same file as the response data, all keys must precede the first response record.

Format

```
KFNAME=<'filename'>
```

Default

No answer key.

Notes

The path to and filename of this file may be longer than 80 characters. As the maximum length of any line in the command file is 80 characters, multiple lines should be used. It is important to continue up to and including the 80th column when specifying a long path and filename.

For example, suppose the data file **exampl06.dat** is in a folder named:

```
C:\PROGRAM FILES\ITEM RESPONSE THEORY\IRT_2002\MARCH20\BILOG-MG-
VERSION1.2\EXAMPLES
```

The correct way to enter this information in the command file is to enclose the name and path in apostrophes, and continue until column 80 is reached. Then proceed in column 1 of the next line as shown below:

```
>GLOBAL DFNAME='C:\PROGRAM FILES\ITEM RESPONSE THEORY\IRT_2002\MARCH20\BILOG-MG
-VERSION1.2\EXAMPLES\EXAMPL06.DAT', NTEST=1, NVTEST=1, NPARM=2, SAVE;
```

If the data are stored in the same folder as the command file, it is sufficient to type

```
DFNAME='EXAMPL06.DAT'
```

Example

In the analysis of single subject data from the file **exampl04.dat**, the answer key appears at the top of the file as indicated by the use of the KFNAME keyword.

```
>INPUT   NTOTAL=40,NFORM=2,KFNAME='EXAMPL04.DAT',NALT=5;
```

As two forms are used, answer keys are given by form before the actual data, and in the same format as the data records. The first few lines of **exampl04.dat** are as follows:

```
ANSWER KEY  FORM 1              1
11111111111111111111
ANSWER KEY  FORM 2              2
11111111111111111111
Samp1                     12    1
11111111122212122111
Samp1                     12    1
11222212221222222112
```

Related topics

- INPUT command: NFORM, NFNAME, and OFNAME keywords
- **Data** menu: **Item Keys** dialog box (see Section 2.4)

NALT keyword (optional)

Purpose

To specify the maximum number of response alternatives in the raw data. 1/NALT is used as the automatic starting value for estimating lower asymptotes (guessing parameters) of the 3-parameter model.

Format

```
NALT=n
```

Default

5 for the 3PL model; 1000 for the 1PL and 2PL models.

Examples

In the case of the following 2-parameter model, 5 responses to each item are given in the data file.

The correct response to each item is noted in the answer key, which appears at the top of the data file (indicated by the KFNAME keyword on the INPUT command).

```
>GLOBAL  DFNAME='EXAMPL03.DAT',NPARM=2;
>LENGTH  NITEMS=(45);
>INPUT   NTOTAL=45,SAMPLE=2000,NGROUP=2,KFNAME='EXAMPL03.DAT',NIDCHAR=5,
         NALT=5,NFORM=2,TYPE=1;
```

When a 3-parameter model is fitted to the same data, 1/5 will be used as starting value for the lower asymptote (guessing parameter) of each item.

```
>GLOBAL  DFNAME='EXAMPL03.DAT',NPARM=3;
>LENGTH  NITEMS=(45);
>INPUT   NTOTAL=45,SAMPLE=2000,NGROUP=2,KFNAME='EXAMPL03.DAT',NIDCHAR=5,
         NALT=5,NFORM=2,TYPE=1;
```

In the following example, a 2-parameter model is fitted to the data. No answer key is given, and it is assumed that the 2 response alternatives (NALT=2) are coded 1 for correct responses and 0 for incorrect responses. If more than 2 response alternatives are present and no code is given, all responses other than 1 will be assumed incorrect.

```
>GLOBAL  DFNAME='EXAMPL04.DAT',NPARM=2;
>LENGTH  NITEMS=(40);
>INPUT   NTOTAL=40,NALT=2;
```

Related topics

- GLOBAL command: NPARM keyword (see Section 3.2.5)
- **Setup** menu: **General** dialog box (see Section 2.3)

NFMT keyword (optional)

Purpose

To specify the number of format records for reading the respondent data records.

Format

```
NFMT=n
```

Default

1.

Examples

In the format statement below, item responses are read from two lines: the first 25 responses are read on the first line of data for each examinee and the second 25 on the second line of data. Although responses are read over two lines, the format statement fits comfortably on one line in the command file, and thus NFMT=1.

```
(11A1,T39,25A1/T13,25A1)
```

If, however, a large data file is used as input, and it becomes necessary to write the format statement over multiple lines in the command file, the value assigned to NFMT should be adjusted to reflect this. For example, NFMT=2 for the following format statement in which 15 items are selected and columns between items are passed over using the "X" operator:

```
(11A1,1X,A1,2X,A1,1X,A1,3X,A1,1X,A1,2X,A1,1X,A1,3X,A1,1X,A1,2X,A1,1X,A1,
  3X,A1,1X,A1,2X,A1,1X,A1)
```

Related topics

- **Data** menu: **Examinee Data** dialog box (see Section 2.4)
- **Data** menu: **Group-Level Data** dialog box
- Variable format statement (see Section 3.2.16)

NFNAME keyword (optional)

Purpose

To specify the name of the file which contains the not-presented key. This key must be given in the same format as the corresponding response records. Any single ASCII character can be used to represent a not-presented item. If the not-presented key is in the same file as the item response data, the key must precede the first response record. If this key appears in the same file as the answer key, it must appear in the file after the answer key. If NFNAME does not appear on the INPUT command, then all items are assumed presented.

When NFORM > 1, separate answer, not-presented, and omit keys must be provided for each form in the order of the forms to which they apply. Again, if they are in the same file as the response data, all keys must precede the first response record.

Format

```
NFNAME=<'filename'>
```

Default

No not-presented key.

Examples

In the analysis of single subject data from the file **exampl04.dat**, the not-presented key appears at the top of the file as indicated below, using the NFNAME keyword.

```
>INPUT   NTOTAL=40,NFORM=2,NFNAME='EXAMPL04.DAT',NALT=5;
>ITEMS   INUMBERS=(1(1)40),INAME=(T01(1)T40);
>TEST    TNAME=SIM;
>FORM1   LENGTH=20,INUMBERS=(1(1)20);
>FORM2   LENGTH=20,INUMBERS=(21(1)40);
(T28,5A1,T25,I1/40A1)
```

As two forms are used, the not-presented keys are given by form before the actual data, and in the same format as the data records. The first few lines of **exampl04.dat** are as follows:

```
Not-P  KEY  FORM 1            1
aaaaaaaaaaaaaaaaaaaa
Not-P  KEY  FORM 2            2
aaaaaaaaaaaaaaaaaaaa
Samp1                  12     1
11a11111122212122111
Samp1                  12     1
112222122a1222222112
```

Alternatively, the lines

```
Not-P  KEY  FORM 1             1
aaaaaaaaaaaaaaaaaaaa
Not-P  KEY  FORM 2             2
aaaaaaaaaaaaaaaaaaaa
```

can be saved to a not-presented key file **exampl04.nfn**, and referenced as such in a revised INPUT command:

```
>INPUT   NTOTAL=40,NFORM=2,NFNAME='EXAMPL04.NFN',NALT=5;
```

If both a not-presented key and an omit key are used for the two forms, the following lines should appear at the top of the data file when the data file is referenced by the NFNAME and OFNAME keywords in the INPUT command:

```
>INPUT   NTOTAL=40,NFORM=2,NFNAME='EXAMPL04.DAT',
         OFNAME='EXAMPL04.DAT',NALT=5;

Not-P  KEY  FORM 1             1
aaaaaaaaaaaaaaaaaaaa
Omit   KEY  FORM 1             1
bbbbbbbbbbbbbbbbbbbb
Not-P  KEY  FORM 2             2
aaaaaaaaaaaaaaaaaaaa
Omit   KEY  FORM 2             2
bbbbbbbbbbbbbbbbbbbb
```

Related topics

- **Data** menu: **Item Keys** dialog box (see Section 2.4)
- GLOBAL command: DFNAME keyword (see Section 3.2.5)
- INPUT command: NFORM keyword
- INPUT command: KFNAME keyword
- INPUT command: OFNAME keyword

NFORM keyword (optional)

Purpose

To specify the number of test forms. If NFORM > 1, the response records must contain an indicator specifying the form to which the examinee responded. This keyword is used in combination with the FORM command and the variable format statement.

The NFORM keyword is required when multiple-form data is supplied to the program in compressed form (see INPUTfile format discussed in Section 3.2.18 for more details). If the instrument consists of a single test form, or multiple-form data is supplied to the program in expanded format, the NFORM keyword, with NFORM=1, is required by the program if the order of items on the response records does not correspond to the order of items in the ITEMS command list.

Format

```
NFORM=n
```

Default

No FORM commands will be read and the order of items in the response records is assumed to be the same as that in the ITEMS command.

Example

In the following example, two forms were administered to two groups of examinees. As both the NFORM and NGROUP keywords are used on the INPUT command, both FORM and GROUP commands are given.

```
>INPUT   NTOTAL=45,NGROUP=2,NIDCHAR=5,NALT=5,NFORM=2;
>ITEMS   INUMBERS=(1(1)45), INAME=(C01(1)C45);
>TEST    TNAME=CHEMISTRY;
>FORM1   LENGTH=25,INUMBERS=(1(1)25);
>FORM2   LENGTH=25,INUMBERS=(21(1)45);
>GROUP1  GNAME=POP1,LENGTH=25,INUMBERS=(1(1)25);
>GROUP2  GNAME=POP2,LENGTH=25,INUMBERS=(21(1)45);
```

Note that the format statement contains both a form and a group indicator.

```
(5A1,T25,I1,T25,I1,25A1)
```

Related topics

- FORM command (see Section 3.2.4)
- Input files (see Section 3.2.18)
- ITEMS command (see Section 3.2.8)
- **Setup** menu: **General** dialog box (see Section 2.3)
- Variable format statement (see Section 3.2.16)

NGROUP keyword (optional)

Purpose

To specify the number of groups or cohorts of respondents. If NGROUP > 1, the response records must contain an indicator specifying the group or cohort to which the respondent belongs. This keyword is used in combination with the GROUP command and the variable format statement, where a group indicator is added.

Format

```
NGROUP=n
```

Default

1.

Related topics

- FORM command (see Section 3.2.4)
- GROUP command (see Section 3.2.6)
- **Setup** menu: **General** dialog box (see Section 2.3)
- Variable format statement (see Section 3.2.16)

NIDCHAR keyword (required)

Purpose

To specify the number of characters in the respondent's identification field. Valid values are 1 to 30.

Format

```
NIDCHAR=n
```

Default

30.

Example

Data from two groups, found on two forms are analyzed in this example. The NIDCHAR keyword is set to 5, indicating that the subject ID field is 5 columns in length. This corresponds with the format statement, where the first entry, for the subject ID, is 5A1.

```
>INPUT   NTOTAL=45,NGROUP=2,NIDCHAR=5,NALT=5,NFORM=2,TYPE=1;
(5A1,T25,I1,T25,I1/25A1)
```

Related topics

- **Data** menu: **Examinee Data** dialog box (see Section 2.4)
- **Data** menu: **Group-Level Data** dialog box
- Variable format statement (see Section 3.2.16)

NTOTAL keyword (optional)

Purpose

To specify the total number of unique items in the respondent data records. The number includes all main and variant test items on all forms.

Format

```
NTOTAL=n
```

Default

0.

Examples

In this example, responses from two groups are analyzed. There are two forms of a 25-item multiple-choice examination, with 5 items in common. In total, the responses of a sample of 2000 respondents to the 45 items are considered.

```
>INPUT NTOTAL=45, SAMPLE=2000, NGROUP=2, NFORM=2;
```

The INPUT command below is used to request a DIF analysis on 4 items administered to two groups.

```
>INPUT NTOTAL=4, DIF, NGROUP=2;
```

In the following example, responses to 50 items are read from the data file. From the 50, 20 are selected as Main Test items and 4 as Variant Test items. Items for the main test are selected by name in the TESTM command; items for the variant test are selected by name in the TESTV command.

```
>GLOBAL  DFNAME='EXAMPL06.DAT', NTEST=1,NVTEST=1,NPARM=2;
>LENGTH  NITEM=24,NVARIANT=4;
>INPUT   NTOTAL=50,KFNAME='EXAMPL06.DAT',SAMPLE=200,NIDCH=11;
>ITEMS   INUMBERS=(1(1)50),INAME=(I26(1)I75);
>TESTM   TNAME=MAINTEST,
         INAMES=(I26,I27,I28,I29,I31,I33,I34,
         I35,I36,I38,I39,I47,I48,I49,I50,I54,I60,I64,I68,I72);
>TESTV   TNAME=VARIANT,INAMES=(I53,I59,I69,I73);
```

Related topics

- **Setup** menu: **General** dialog box (see Section 2.3)

OFNAME keyword (optional)

Purpose

To specify the name of the file which contains the omit key. This key must be specified in the same format as the response records. Any single ASCII character can be used to represent a not-presented item. If the not-presented key is in the same file as the item response data, the key must precede the first response record. If this key appears in the same file as the answer and/or not-presented keys, it must appear in the file after the both keys.

If OFNAME does not appear on the INPUT command, omits will not be distinguished from incorrect responses. When NFORM > 1, separate answer, not-presented, and omit keys must be provided for each form in the order of the forms to which they apply. Again, if they are in the same file as the response data, all keys must precede the first response record.

Format

```
OFNAME=character string
```

Default

No omit key.

Examples

In the analysis of single subject data from the file **exampl04.dat**, the omit key appears at the top of the file as indicated by the use of the OFNAME keyword.

```
>INPUT   NTOTAL=40,NFORM=2,OFNAME='EXAMPL04.DAT',NALT=5;
>ITEMS   INUMBERS=(1(1)40),INAME=(T01(1)T40);
>TEST    TNAME=SIM;
>FORM1   LENGTH=20,INUMBERS=(1(1)20);
>FORM2   LENGTH=20,INUMBERS=(21(1)40);
(T28,5A1,T25,I1/40A1)
```

As two forms are used, omit keys are given by form before the actual data, and in the same format as the data records. The first few lines of **exampl04.dat** are as follows:

```
Omit   KEY  FORM 1             1
bbbbbbbbbbbbbbbbbbbb
Omit   KEY  FORM 2             2
bbbbbbbbbbbbbbbbbbbb
Samp1                   12     1
11a11111122212122111
Samp1                   12     1
112222122a1222222112
```

Alternatively, the lines

```
Omit   KEY  FORM 1            1
bbbbbbbbbbbbbbbbbbbb
Omit   KEY  FORM 2            2
bbbbbbbbbbbbbbbbbbbb
```

can be saved to a omit key file **exampl04.ofn**, and referenced as such in a revised INPUT command:

```
>INPUT   NTOTAL=40,NFORM=2,NFNAME='EXAMPL04.OFN',NALT=5;
```

If both a not-presented key and an omit key are used for the two forms, the following lines should appear at the top of the data file when the data file is referenced by the NFNAME and OFNAME keywords in the INPUT command:

```
>INPUT   NTOTAL=40,NFORM=2,NFNAME='EXAMPL04.DAT',
         OFNAME='EXAMPL04.DAT',NALT=5;

Not-P  KEY  FORM 1            1
aaaaaaaaaaaaaaaaaaaa
Omit   KEY  FORM 1            1
bbbbbbbbbbbbbbbbbbbb
Not-P  KEY  FORM 2            2
aaaaaaaaaaaaaaaaaaaa
Omit   KEY  FORM 2            2
bbbbbbbbbbbbbbbbbbbb
```

Related topics

- **Data** menu: **Item Keys** dialog box (see Section 2.4)
- GLOBAL command: DFNAME and NPARM keywords (see Section 3.2.5)
- INPUT command: KFNAME, NFNAME, and NFORM keywords

PERSONAL option (optional)

Purpose

To specify the assumption that the group or cohort assignment of an examinee is personalized by subtest. The response records must contain NTEST indicators, one for each subtest, specifying the groups or group cohorts to which the respondent belongs. The NTEST group indicators must be specified in the variable format statement in the same order as the subtests.

The PERSONAL option is especially useful for two-stage tests that measure ability in more than one area. Assignment to the second-stage booklets may differ among areas.

Format

```
PERSONAL
```

Default

None.

Related topics

- **Data** menu: **Examinee Data** dialog box (see Section 2.4)
- **Data** menu: **Group-Level Data** dialog box
- GLOBAL command: NTEST keyword (see Section 3.2.5)
- Variable format statement (see Section 3.2.16)

SAMPLE keyword (optional)

Purpose

To specify the number of respondents to be randomly sampled from the raw data file.

Format

```
SAMPLE=n
```

Default

1000.

Example

Here data are read from the file **exampl03.dat**, which also contains the answer key (DFNAME and KFNAME keywords). Although the data file contains only 400 records, a sample of 2000 is requested.

```
>GLOBAL  DFNAME='EXAMPL03.DAT',NPARM=2;
>LENGTH  NITEMS=(45);
>INPUT   NTOTAL=45,SAMPLE=2000,NGROUP=2,KFNAME='EXAMPL03.DAT',NIDCHAR=5,
         NALT=5,NFORM=2,TYPE=1;
```

If the first few records of the data file are to be used, the TAKE keyword should be used instead.

Related topics

- **Data** menu: **Examinee Data** dialog box (see Section 2.4)
- **Data** menu: **Group-Level Data** dialog box
- INPUT command: TAKE keyword

TAKE keyword (optional)

Purpose

To specify an analysis using only the first n respondents in the data file. This option is useful for testing the problem setup on a smaller number of respondents when the sample size is large. Note that the maximum value for this keyword is the actual number of respondents in the data file. To obtain a random sample of the respondents, the SAMPLE keyword should be used. TAKE and SAMPLE are mutually exclusive keywords.

Format

```
TAKE=n
```

Default

Take all data specified by SAMPLE.

Examples

In the following example, data are read from the file **exampl03.dat**, which also contains the answer key (DFNAME and KFNAME keywords). Although the data file contains only 400 records, a sample of 2000 is requested.

```
>GLOBAL  DFNAME='EXAMPL03.DAT',NPARM=2;
>LENGTH  NITEMS=(45);
>INPUT   NTOTAL=45,SAMPLE=2000,NIDCHAR=5,NALT=5,TYPE=1;
```

If, however, only the first 100 records are to be used in the analysis, the modified INPUT command

```
>INPUT   NTOTAL=45,TAKE=100,NIDCH=5,NALT=5,TYPE=1;
```

should be used.

Related topics

- **Data** menu: **Examinee Data** dialog box (see Section 2.4)
- **Data** menu: **Group-Level Data** dialog box
- INPUT command: SAMPLE keyword

TYPE keyword (optional)

Purpose

To specify the type of data file to be used in the analysis:

- 0: no raw data to read in
- 1: single-subject data to read in
- 2: single-subject data with case weights
- 3: number tried, number right data, no case weights
- 4: number tried, number right data, case weights

Format

```
TYPE=n
```

Default

1.

Examples

In a preliminary run, an item parameter file was created as shown below. The item parameter file was saved to **exampl03.par** using the PARM keyword on the SAVE command. As single-subject data were used in this run TYPE was set to 1 in the INPUT command.

```
EXAMPLE:
CREATING A ITEM PARAMETER FILE
>COMMENT
>GLOBAL  DFNAME='EXAMPL03.DAT',NPARM=2, SAVE;
>SAVE    PARM='EXAMPL03.PAR';
>LENGTH  NITEMS=(45);
>INPUT   NTOTAL=45,SAMPLE=2000,NGROUP=2,KFNAME='EXAMPL03.DAT',NIDCH=5,
         NALT=5,NFORM=2,TYPE=1;
```

The previously created calibration file is now used as INPUTthrough the use of the IFNAME keyword on the GLOBAL command. Note that the TYPE keyword on the INPUT command is now set to 0, compared to 1 previously. The updated item parameter estimates are saved to the file **latest.par** using the PARM keyword on the SAVE command.

```
EXAMPLE:
USING A ITEM PARAMETER FILE AS INPUT
>COMMENT
>GLOBAL  CFNAME='EXAMPL03.PAR',NPARM=2, SAVE;
>SAVE    CALIB='LATEST.PAR';
>LENGTH  NITEMS=(45);
>INPUT   NTOTAL=45,SAMPLE=2000,NGROUP=2,NIDCHAR=5,
         NALT=5,NFORM=2,TYPE=0;
```

Related topics

- **Data** menu: **Examinee Data** dialog box (see Section 2.4)
- **Data** menu: **Group-Level Data** dialog box
- GLOBAL command: IFNAME keyword (see Section 3.2.5)
- SAVE command: PARM keyword (see Section 3.2.13)

3.2.8 ITEMS command

(Required)

Purpose

To specify the names and corresponding numbers for all items in the data records. The items may be listed in any order, but the order in which the names appear must correspond with the order of the numbers. The names and numbers specified in the ITEMS command are used to refer to the items in the TEST, FORM, and GROUP commands.

Strings of consecutive numbers may be abbreviated as m(1)n, where m is the number of the first item and *n* is the number of the last item. Strings of up to 8 character names including consecutive numbers may be abbreviated as Xm(1)Xn, where X is a string of up to 4 letters of the alphabet, m is the up-to-4 character integer number of the first item and n is the up-to-4 character integer number of the last item.

Format

```
>ITEMS INUMBERS=(list), INAMES=(list);
```

Default

None.

Examples

In the first example, 15 items are assigned the names MATH01 through MATH15.

```
>ITEMS   INAME=(MATH01(1)MATH15);
```

In the syntax that follows, 16 items belonging to 2 subtests are identified. From the LENGTH command, we see that each subtest has 8 items. The ITEMS command is used to first number these items, and then to assign the names N1 through N8 to items belonging to the first subtest. Items belonging to the second subtest are named A1 through A8. On the TEST commands, items are referenced by number. Referencing by the names assigned in the ITEMS command is another option.

```
>LENGTH  NITEMS=(8,8);
>INPUT   NTOTAL=16,NALT=5,NIDCHAR=9,TYPE=3;
>ITEMS   INUMBERS=(1(1)16),INAMES=(N1(1)N8,A1(1)A8);
>TEST1   TNAME=NUMCON,INUMBERS=(1(1)8);
>TEST2   TNAME=ALGCON,INUMBERS=(9(1)16);
```

Related topics

- FORM command (see Section 3.2.4)

- GROUP command (see Section 3.2.6)
- TEST command (see Section 3.2.15)
- **Setup** menu: **General** dialog box (see Section 2.3)

INAMES keyword (optional)

Purpose

To specify a list of NTOTAL unique names (up to eight characters each). Item names that do not begin with letters must be enclosed in single quotes.

Strings of up to 8 character names including consecutive numbers may be abbreviated as Xm(1)Xn, where X is a string of up to 4 letters of the alphabet, m is the up-to-4 character integer number of the first item and n is the up-to-4 character integer number of the last item.

Format

INAMES=($n_1, n_2, \ldots, n_{NTOTAL}$)

Default

```
1, 2, …, NTOTAL.
```

Related topics

- INPUT command: NTOTAL keyword (see Section 3.2.7)
- **Setup** menu: **General** dialog box (see Section 2.3)

INUMBERS keyword (optional)

Purpose

To specify the list of NTOTAL unique numbers. Strings of consecutive numbers may be abbreviated as m(1)n, where m is the number of the first item and n is the number of the last item.

Format

INUMBERS=($n_1, n_2, \ldots, n_{NTOTAL}$)

Default

```
1, 2, …, NTOTAL.
```

Example

In the syntax that follows, 16 items belonging to 2 subtests are identified. From the LENGTH command we see that each subtest has 8 items. The ITEMS command is used to first number these items, and then to assign the names N1 through N8 to items belonging to the first subtest. Items belonging to the second subtest are named A1 through A8. On the TEST commands, items are referenced by number. Referencing by the names assigned in the ITEMS command is another option.

```
>LENGTH  NITEMS=(8,8);
>INPUT   NTOTAL=16,NALT=5,NIDCHAR=9,TYPE=3;
>ITEMS   INUMBERS=(1(1)16),INAMES=(N1(1)N8,A1(1)A8);
>TEST1   TNAME=NUMCON,INUMBERS=(1(1)8);
>TEST2   TNAME=ALGCON,INUMBERS=(9(1)16);
```

Related topics

- INPUT command: NTOTAL keyword (see Section 3.2.7)

3.2.9 LENGTH command

(Required)

Purpose

To supply the number of items in subtests and the number of variant items in the subtests.

Format

```
>LENGTH NITEMS=(list), NVARIANT=(list);
```

Example

Consider two subtests. Subtest 1 has a total of 20 items; subtest 2 has a total of 15 items. Five of the items in subtest 1 are variant items. None of the items in subtest 2 are variant items.

Note that the number of variant tests has to be specified using the NVTEST keyword on the GLOBAL command. The corresponding number of TEST commands must also be included in the syntax.

```
>GLOBAL DFNAME='EXAMPL04.DAT',NTEST=2,NVTEST=1;
…
>LENGTH NITEMS=(20,15), NVARIANT=(5,0);
```

Related topics

- GLOBAL command: NVTEST and NVTEST keywords (see Section 3.2.5)

- **Setup** menu: **Item Analysis** dialog box (see Section 2.3)
- TEST command (see Section 3.2.15)

NITEMS keyword (required)

Purpose

To provide a list of the number of items in the successive subtests to be analyzed. If a subtest contains variant items, they are included in this count of items.

Format

NITEMS=($n_1, n_2, ..., n_{NTESTS}$)

Default

None.

Example

In the example below, 20 of the 24 items are selected as main test items and 4 as variant test items. The number of variant tests is specified using the NVTEST keyword on the GLOBAL command. The TEST command for the main test is followed by a TEST command in which the variant items are specified by item number.

```
>GLOBAL DFNAME='example.dat', NTEST=1, NVTEST=1;
>LENGTH NITEM=24, NVARIANT=4;
>INPUT  NTOTAL=24;
>ITEMS  INUMBER=(1(1)24);
>TESTM  TNAME=MAINTEST,
        INUMBER=(1(1)20);
>TESTV  TNAME=VARIANT,
        INUMBER=(21(1)24);
```

Related topics

- GLOBAL command: NTEST and NVTEST keywords (see Section 3.2.5)
- **Setup** menu: **Item Analysis** dialog box (see Section 2.3)
- TEST command (see Section 3.2.15)

NVARIANT keyword (optional)

Purpose

To specify the number of variant items, if any, in the successive subtests to be analyzed. Although parameter estimates for these items will be obtained, these items are not used in scoring of tests/forms.

Format

```
NVARIANT=(nv1,nv2,...,nvNVTESTS)
```

Default

0.

Related topics

- GLOBAL command: NTEST and NVTEST keywords (see Section 3.2.5)
- **Setup** menu: **Item Analysis** dialog box (see Section 2.3)
- TEST command (see Section 3.2.15)

3.2.10 PRIORS command

(Optional)

Purpose

To specify prior distributions for constrained estimation of the item parameters of the main test and for the variant items, if any. This command is required when the READPRIOR keyword appears in the CALIB command.

There is one prior command for each subtest. Values are read in order of the items in the subtest beginning with the main test items and ending with the variant test items. If NGROUP>1, more than one set of prior means and standard deviations for the item thresholds may be required when the DIF or DRIFT models are specified. See the TMU and TSIGMA keywords below.

Format

```
>PRIORS TMU=(list),TSIGMA=(list),SMU=(list) SSIGMA=(list),
       ALPHA=(list), BETA=(list);
```

Notes

If the same value applies to all items of the subtest, you may use the “repeat” form: “value (0) number of values” (see Section 3.2.2).

For a mean of p with a weight of n observations for the beta prior distribution, set

```
ALPHA=np+1
BETA=n(1-p)+1
```

To set an item parameter to a fixed value, set the mean of the prior to the parameter value and set the corresponding standard deviation to a very small value. Suitable values for TSIGMA are 0.005, for SSIGMA, 0.001 and for ALPHA and BETA, n = 1000. The priors for free parameters should be set to the default values above. The PRIORS command for each test should appear immediately after the QUAD commands for that test.

Examples

The following example emanates from an analysis of aggregate-level, multiple-matrix sampling data. Aggregate-level data typically have smaller slopes in the 0,1 metric than do person-level data. For this reason, the mean of the prior for the log slopes has been set to 0.5 by the use of the READPRIOR option on the CALIB command and the successive PRIOR commands. The NOFLOAT option is used to keep the means of the prior distributions on the item parameters fixed at their specified values during estimation.

```
>CALIB   NQPT=30, CYCLES=24, NEWTON=4, CRIT=0.0050,
         RIDGE=(2, 0.8000, 2.0000), ACCEL=1.0000, TPRIOR, READPRI,
         NOFLOAT;
>PRIORS1 SMU=(0.5000(0)8);
>PRIORS2 SMU=(0.5000(0)8);
```

The next example illustrates how user-supplied priors for the latent distributions are specified with IDIST=1 on the CALIB command. The points and weights for these distributions are supplied in the corresponding QUAD commands. Note that with IDIST=1, there are separate QUAD commands for each group for each subtest. Within each subtest the points are the same for each group. This is a requirement of the program. But as the example shows, the points for the groups may differ by subtest. The PRIOR command for each subtest is placed after the QUAD commands for that subtest. In this example, only the prior for the standard deviations of the thresholds is supplied on the PRIOR command. Default values are used for the other prior distributions. The means of the distributions are kept fixed at their specified values by using the NOFLOAT option on the CALIB command.

```
>GLOBAL  DFNAME='EXAMPL03.DAT',NPARM=2,NTEST=2;
>LENGTH  NITEMS=(35,35);
>INPUT   NTOT=45,SAMPLE=2000,NGROUP=2,KFNAME='EXAMPL03.DAT',NALT=5,
         NFORMS=2,NIDCHAR=5;
>ITEMS   INUMBERS=(1(1)45), INAME=(C01(1)C45);
>TEST1   TNAME=SUBTEST1,INAME=(C01(1)C15,C21(1)C40);
>TEST2   TNAME=SUBTEST2,INAME=(C06(1)C25,C31(1)C45);
>FORM1   LENGTH=25,INUMBERS=(1(1)25);
>FORM2   LENGTH=25,INUMBERS=(21(1)45);
>GROUP1  GNAME=POP1,LENGTH=25,INUMBERS=(1(1)25);
>GROUP2  GNAME=POP2,LENGTH=25,INUMBERS=(21(1)45);
(T28,5A1,T25,I1,T25,I1/45A1)
>CALIB   IDIST=1,READPR,EMPIRICAL,NQPT=16,CYCLE=25,TPRIOR,NEWTON=5,
         CRIT=0.01,REFERENCE=1,NOFLOAT;
>QUAD1   POINTS=(-0.4598E+01,-0.3560E+01,-0.2522E+01,-0.1484E+01,
         -0.4453E+00,0.5930E+00,0.1631E+01,0.2670E+01,0.3708E+01,
          0.4746E+01),
         WEIGHTS=(0.2464E-05,0.4435E-03,0.1724E-01,0.1682E+00,
```

```
          0.3229E+00,0.3679E+00,0.1059E+00,0.1685E-01,0.6475E-03,
          0.8673E-05);
>QUAD2    POINTS=(-0.4598E+01,-0.3560E+01,-0.2522E+01,-0.1484E+01,
          -0.4453E+00,0.5930E+00,0.1631E+01,0.2670E+01,0.3708E+01,
           0.4746E+01),
          WEIGHTS=(0.2996E-04,0.1300E-02,0.1474E-01,0.1127E+00,
          0.3251E+00,0.3417E+00,0.1816E+00,0.2149E-01,0.1307E-02,
          0.3154E-04);
>PRIOR    TSIGMA=(1.5(0)35);
>QUAD1    POINTS=(-0.4000E+01,-0.3111E+01,-0.2222E+01,-0.1333E+01,
          -0.4444E+00,0.4444E+00,0.1333E+01,0.2222E+01,0.3111E+01,
           0.4000E+01),
          WEIGHTS=(0.1190E-03,0.2805E-02,0.3002E-01,0.1458E+00,
          0.3213E+00,0.3213E+00,0.1458E+00,0.3002E-01,0.2805E-02,
          0.1190E-03);
>QUAD2    POINTS=(-0.4000E+01,-0.3111E+01,-0.2222E+01,-0.1333E+01,
          -0.4444E+00,0.4444E+00,0.1333E+01,0.2222E+01,0.3111E+01,
           0.4000E+01),
          WEIGHTS=(0.1190E-03,0.2805E-02,0.3002E-01,0.1458E+00,
          0.3213E+00,0.3213E+00,0.1458E+00,0.3002E-01,0.2805E-02,
          0.1190E-03);
>PRIOR    TSIGMA=(1.5(0)35);
```

Suppose IDIST=1, NGROUP=2, and NTEST=2. The setup for the QUAD and PRIOR commands is as follows:

```
>QUAD1  (specifications for Group 1, subtest 1)
>QUAD2  (specifications for Group 2, subtest 1)
>PRIOR1 (specifications for Groups 1 and 2, subtest1)
>QUAD1  (specifications for Group 1, subtest2)
>QUAD2  (specifications for Group 2, subtest 2)
>PRIOR2 (specifications for Groups 1 and 2, subtest 2)
```

Related topics

- CALIB command: READPRIOR option (see Section 3.2.2)
- GLOBAL command: NTEST keyword (see Section 3.2.5)
- INPUT command: NGROUP keyword (see Section 3.2.7)
- QUAD command (see Section 3.2.11)
- **Technical** menu: **Item Parameter Prior Constraints** dialog box

ALPHA keyword (optional)

Purpose

To specify the real-valued “alpha” parameters for the beta prior distribution of lower asymptote (guessing) parameters.

Format

ALPHA=($n_1, n_2, ..., n_N$)

Default

`20p+1.`

Related topics

- CALIB command: READPRIOR option (see Section 3.2.2)
- GLOBAL command: NPARM keyword (see Section 3.2.5)
- PRIORS command: BETA keyword
- **Technical** menu: **Item Parameter Prior Constraints** dialog box (see Section 2.5)

BETA keyword (optional)

Purpose

To specify the real-valued "beta" parameters for the beta prior distribution of lower asymptote (guessing) parameters.

Format

BETA=($n_1, n_2, ..., n_N$)

Default

`20(1-p)+1.`

Related topics

- CALIB command: READPRIOR option (see Section3.2.2)
- GLOBAL command: NPARM keyword (see Section 3.2.5)
- PRIORS command: ALPHA keyword
- **Technical** menu: **Item Parameter Prior Constraints** dialog box (see Section 2.5)

SMU keyword (optional)

Purpose

To provide real-valued prior means for the item slopes.

Format

SMU=($n_1, n_2, ..., n_N$)

Default

1.0.

Example

In the following example, SMU is used to specify prior means for the item slopes.

```
>CALIB   NQPT=30, CYCLES=24, NEWTON=4, CRIT=0.0050, NOFLOAT,
         RIDGE=(2, 0.8000, 2.0000), ACCEL=1.0000, SPRIOR, READPRI;
>PRIORS1 SMU=(0.5000(0)8);
>PRIORS2 SMU=(0.5000(0)8);
```

Related topics

- CALIB command: READPRIOR option (see Section 3.2.2)
- **Technical** menu: **Item Parameter Prior Constraints** dialog box (see Section 2.5)

SSIGMA keyword (optional)

Purpose

To specify real-valued prior standard deviations of the item slopes.

Format

SSIGMA=($n_1, n_2, ..., n_N$)

Default

1.64872127.

Example

In the calibration of a single subtest with 35 items, the following PRIOR command is used to provide a real-valued prior standard deviation of 1.75 for the item slopes.

```
>CALIB   READPRI, NQPT=16,CYCLE=25, NEWTON=5;
>PRIOR   SSIGMA=(1.75(0)35);
```

Related topics

- CALIB command: READPRIOR option (see Section 3.2.2)
- **Technical** menu: **Item Parameter Prior Constraints** dialog box (see Section 2.5)

TMU keyword (optional)

Purpose

To specify real-valued prior means for the item thresholds (DIF) or polynomial coefficients (DRIFT) including intercept.

Format

TMU=($n_1, n_2, ..., n_{MITM_1}, n_1, n_2, ..., n_{MITM_2}, ..., n_1, n_2, ..., n_{MITM_L}$)

Default

0.0.

Example

In the example, below, PRIOR commands are used to specify prior distributions for the constrained estimation of the thresholds in the calibration of two subtests with 8 items each.

```
>CALIB   NQPT=30, CYCLES=24, NEWTON=4, CRIT=0.0050, READPRI;
>PRIORS1 TMU=(2.0500(0)8);
>PRIORS2 TMU=(2.0500(0)8);
```

Related topics

- CALIB command: READPRIOR option (see Section 3.2.2)
- INPUT command: DIF option (see Section 3.2.7)
- INPUT command: DRIFT option
- **Technical** menu: **Item Parameter Prior Constraints** dialog box (see Section 2.5)

TSIGMA keyword (optional)

Purpose

To specify real-valued prior standard deviations of the threshold parameters.

If neither the DIF nor the DRIFT model is selected, L = 1. If the DIF model is selected, L = NGROUP. If the DRIFT model is selected, L = MAXPOWER.

Format

TSIGMA=($n_1, n_2, ..., n_{MITM_1}, n_1, n_2, ..., n_{MITM_2}, ..., n_1, n_2, ..., n_{MITM_L}$)

Default

2.0.

Related topics

- CALIB command: READPRIOR option (see Section 3.2.2)
- DRIFT command: MAXPOWER keyword (see Section 3.2.3)
- INPUT command: DIF option (see Section 3.2.7)
- INPUT command: DRIFT option
- INPUT command: NGROUP keyword
- **Technical** menu: **Item Parameter Prior Constraints** dialog box (see Section 2.5)

3.2.11 QUAD command

(Required if IDIST = 1 or 2 on CALIB command)

Purpose

To read in user-supplied quadrature points and weights, or points and ordinates of the discrete finite representations of the prior distribution for the groups. This command follows directly after the CALIB command.

If:

- IDIST = 0: This command is not used.
- IDIST = 1: There must be a separate QUAD command for each group for each subtest. For any subtest, the points for each group must have the same values
- IDIST = 2: There must be a separate QUAD command for each group. The same set of QUAD commands applies to all subtests. The points for each group must have the same values.

Format

```
>QUAD POINTS=(list), WEIGHTS=(list);
```

Example

This example illustrates user-supplied priors for the latent distributions are specified with IDIST=1 on the CALIB command. The points and weights for these distributions are supplied in the QUAD commands. Note that with IDIST=1, there are separate QUAD commands for each group for each subtest.

Within each subtest the points are the same for each group. This is a requirement of the program. But as the example shows, the points for the groups may differ by subtest. The PRIOR command for each subtest is placed after the QUAD commands for that subtest.

```
>GLOBAL  DFNAME='EXAMPL03.DAT',NPARM=2,NTEST=2;
>LENGTH  NITEMS=(35,35);
>INPUT   NTOT=45,SAMPLE=2000,NGROUP=2,KFNAME='EXAMPL03.DAT',NALT=5,
         NFORMS=2,NIDCHAR=5;
>ITEMS   INUMBERS=(1(1)45), INAME=(C01(1)C45);
>TEST1   TNAME=SUBTEST1,INAME=(C01(1)C15,C21(1)C40);
>TEST2   TNAME=SUBTEST2,INAME=(C06(1)C25,C31(1)C45);
>FORM1   LENGTH=25,INUMBERS=(1(1)25);
>FORM2   LENGTH=25,INUMBERS=(21(1)45);
>GROUP1  GNAME=POP1,LENGTH=25,INUMBERS=(1(1)25);
>GROUP2  GNAME=POP2,LENGTH=25,INUMBERS=(21(1)45);
(T28,5A1,T25,I1,T25,I1/45A1)
>CALIB   IDIST=1,READPR,EMPIRICAL,NQPT=16,CYCLE=25,TPRIOR,NEWTON=5,
         CRIT=0.01,REFERENCE=1,NOFLOAT;
>QUAD1   POINTS=(-0.4598E+01,-0.3560E+01,-0.2522E+01,-0.1484E+01,
         -0.4453E+00,0.5930E+00,0.1631E+01,0.2670E+01,0.3708E+01,
          0.4746E+01),
         WEIGHTS=(0.2464E-05,0.4435E-03,0.1724E-01,0.1682E+00,
         0.3229E+00,0.3679E+00,0.1059E+00,0.1685E-01,0.6475E-03,
         0.8673E-05);
>QUAD2   POINTS=(-0.4598E+01,-0.3560E+01,-0.2522E+01,-0.1484E+01,
         -0.4453E+00,0.5930E+00,0.1631E+01,0.2670E+01,0.3708E+01,
          0.4746E+01),
         WEIGHTS=(0.2996E-04,0.1300E-02,0.1474E-01,0.1127E+00,
         0.3251E+00,0.3417E+00,0.1816E+00,0.2149E-01,0.1307E-02,
         0.3154E-04);
>PRIOR   TSIGMA=(1.5(0)35);
>QUAD1   POINTS=(-0.4000E+01,-0.3111E+01,-0.2222E+01,-0.1333E+01,
         -0.4444E+00,0.4444E+00,0.1333E+01,0.2222E+01,0.3111E+01,
          0.4000E+01),
         WEIGHTS=(0.1190E-03,0.2805E-02,0.3002E-01,0.1458E+00,
         0.3213E+00,0.3213E+00,0.1458E+00,0.3002E-01,0.2805E-02,
         0.1190E-03);
>QUAD2   POINTS=(-0.4000E+01,-0.3111E+01,-0.2222E+01,-0.1333E+01,
         -0.4444E+00,0.4444E+00,0.1333E+01,0.2222E+01,0.3111E+01,
          0.4000E+01),
         WEIGHTS=(0.1190E-03,0.2805E-02,0.3002E-01,0.1458E+00,
         0.3213E+00,0.3213E+00,0.1458E+00,0.3002E-01,0.2805E-02,
         0.1190E-03);
>PRIOR   TSIGMA=(1.5(0)35);
```

Related topics

- CALIB command: IDIST keyword (see Section 3.2.2)
- GLOBAL command: NTEST keyword (see Section 3.2.5)
- INPUT command: NGROUP keyword (see Section 3.2.7)
- **Technical** menu: **Calibration Prior Latent Distribution** dialog box (see Section 2.5)

POINTS keyword (optional)

Purpose

To specify the location of quadrature points.

If:

- IDIST = 1: a set of NQPT real-numbered values (with decimal points) of the quadrature points must be supplied for each group for each subtest.
- IDIST = 2: one set of points is required for each group.

Format

POINTS=($n_1, n_2, ..., n_{NQPT}$)

Default

Supplied by program.

Example

See the example given above.

Related topics

- CALIB command: IDIST and NQPT keywords (see Section 3.2.2)
- GLOBAL command: NTEST keyword (see Section 3.2.5)
- INPUT command: NGROUP keyword (see Section 3.2.7)
- **Technical** menu: **Calibration Prior Latent Distribution** dialog box (see Section 2.5)

WEIGHTS keyword (optional)

Purpose

To supply the weights for the quadrature points.

If:

- IDIST = 1 on the CALIB command: A set of NQPT positive fractions (with decimal points and summing to 1.0) for weights for quadrature points must be supplied for each group for each subtest. This set of points applies to all subtests.
- IDIST = 2: One set of weights is required for each group. This set of weights applies to all subtests.

Format

WEIGHTS=($n_1, n_2, ..., n_{NQPT}$)

Default

Supplied by program.

Related topics

- CALIB command: IDIST and NQPT keywords (see Section 3.2.2)
- GLOBAL command: NTEST keyword (see Section 3.2.5)
- INPUT command: NGROUP keyword (see Section 3.2.7)
- **Technical** menu: **Calibration Prior Latent Distribution** dialog box (see Section 2.5)

3.2.12 QUADS command

(Required command if IDIST =1 or IDIST =2 on SCORE command)

Purpose

To supply arbitrary prior distributions of scale scores for the respondents when EAP estimation is selected. This command follows directly after the SCORE command.

If:

- IDIST = 0: This command is not required
- IDIST = 1: There must be as separate QUADSj command for each group for each subtest.
- IDIST = 2: There must be a separate QUADSj command for each group. The same set of QUADS commands applies to all subtests.
- IDIST = 3: This command is not required.
- IDIST = 4: This command is not required.

If there are multiple groups (NGROUPS > 1) and IDIST = 1 or 2, the POINTS must have the same values for all groups. The WEIGHTS may differ by group, and the POINTS may differ by subtest.

Format

```
>QUADS POINTS=(list), WEIGHTS=(list);
```

Example

In the 2-group example below, an illustration is given of the use of user-supplied priors for the scale scores (IDIST=2) for the respondents when EAP estimation is selected (METHOD=2). The points and weights for these distributions are supplied in the QUADS commands. Note that with IDIST=2, there are separate QUADS commands for each group.

```
>SCORE     NQPT = 10, METHOD = 2, IDIST=2, INFO=1, YCOMMON, POP;
>QUADS1    POINTS=(-0.4598E+01,-0.3560E+01,-0.2522E+01,-0.1484E+01,
           -0.4453E+00,0.5930E+00,0.1631E+01,0.2670E+01,0.3708E+01,
```

```
           0.4746E+01),
          WEIGHTS=(0.2464E-05,0.4435E-03,0.1724E-01,0.1682E+00,
          0.3229E+00,0.3679E+00,0.1059E+00,0.1685E-01,0.6475E-03,
          0.8673E-05);
>QUADS2   POINTS=(-0.4598E+01,-0.3560E+01,-0.2522E+01,-0.1484E+01,
          -0.4453E+00,0.5930E+00,0.1631E+01,0.2670E+01,0.3708E+01,
           0.4746E+01),
          WEIGHTS=(0.2996E-04,0.1300E-02,0.1474E-01,0.1127E+00,
          0.3251E+00,0.3417E+00,0.1816E+00,0.2149E-01,0.1307E-02,
          0.3154E-04);
```

Related topics

- GLOBAL command: NTEST keyword (see Section 3.2.5)
- INPUT command: NGROUP keyword (see Section 3.2.7)
- SCORE command: IDIST keyword (see Section 3.2.14)
- **Technical** menu: **Scoring Prior Latent Distribution** dialog box (see Section 2.5)

POINTS keyword (optional)

Purpose

To specify real-numbered, non-negative values (with decimal points) for the NQPT points of the arbitrary discrete prior distribution.

Format

POINTS=($n_1, n_2, \ldots, n_{NQPT}$)

Default

Supplied by program.

Example

See example above.

Related topics

- GLOBAL command: NTEST keyword (see Section 3.2.5)
- INPUT command: NGROUP keyword (see Section 3.2.7)
- SCORE command: IDIST or NQPT keywords (see Section 3.2.14)
- **Technical** menu: **Scoring Prior Latent Distribution** dialog box (see Section 2.5)

WEIGHTS keyword (optional)

Purpose

To specify real-numbered, non-negative values (with decimal points) for the NQPT weights of the arbitrary discrete prior distribution. The sum of the weights must equal unity.

Format

WEIGHTS=($n_1, n_2, \ldots, n_{NQPT}$)

Default

Supplied by program.

Example

See the example above.

Related topics

- GLOBAL command: NTEST keyword (see Section 3.2.5)
- INPUT command: NGROUP keyword (see Section 3.2.7)
- SCORE command: IDIST or NQPT keywords (see Section 3.2.14)
- **Technical** menu: **Scoring Prior Latent Distribution** dialog box (see Section 2.5)

3.2.13 SAVE command

(Required when SAVE is specified on the GLOBAL command)

Purpose

This command is used to supply output filenames. The filenames must be less than 128 characters long and may contain a drive prefix, a path name, and an extension. The filename must be enclosed in single quotes. Note that each line of the command file has a maximum length of 80 characters. If the filename does not fit on one line of 80 characters, the remaining characters should be placed on the next line, starting at column 1. All output files other than the MASTER and CALIB files are saved in a formatted form. See Section 3.2.18 on output files for more information. Note that, in order to use the SAVE command, the SAVE option must be included in the GLOBAL command.

Format

```
>SAVE MASTER=n, CALIB=n, PARM=n, SCORE=n, COVARIANCE=n, TSTAT=n, POST=n,
      EXPECTED=n, ISTAT=n, DIF=n, DRIFT=n, PDISTRIB=n;
```

Example

In the syntax below, the item parameters and scale scores are saved to file through use of the SCORE and PARM keywords on the SAVE command. Note that, in order to use the SAVE command, the SAVE keyword is added to the GLOBAL command.

```
>GLOBAL DFNAME='EXAMPLE.DAT', NPARM=2, SAVE;
>SAVE   SCORE='EXAMPLE.SCO', PARM='EXAMPLE.PAR';
>LENGTH NITEMS=(40);
```

Related topics

- GLOBAL command: SAVE option (see Section 3.2.5)
- Output files (see Section 3.2.18)
- **Save** menu (see Section 2.6)

CALIB keyword (optional)

Purpose

To specify a filename for the calibration data file that is to be saved. The original response data are sampled and calibrated, then saved as a temporary binary file. If no sampling occurs, this temporary file cannot be created. Upon normal termination of the program this temporary file is deleted automatically. By assigning a specific name to the calibration data file, the user can save and reuse it as a master data file in subsequent analyses.

Format

```
CALIB=<'filename'>
```

Default

Do not save.

Example

The calibration file is saved to **exampl03.cal** using the CALIB keyword on the SAVE command.

```
>GLOBAL  DFNAME='EXAMPL03.DAT',NPARM=2, SAVE;
>SAVE    CALIB='EXAMPL03.CAL';
```

Related topics

- GLOBAL command: SAVE option (see Section 3.2.5)
- Output files (see Section 3.2.18)
- **Save** menu (see Section 2.6)

COVARIANCE keyword (optional)

Purpose

To specify a filename for the external file to which the covariances of item parameter estimates for each item are written. This file is written automatically in the calibration phase (Phase 2) as a temporary file, which passes necessary information to the scoring phase (Phase 3). Normally, it is deleted at the termination of the program, but by assigning a specific name to this file the user can save it as a permanent file.

Format

```
COVARIANCE=<'filename'>
```

Default

Do not save.

Example

A covariance file from a previous calibration can be used to compute test information by specifying the name of the file with the COVARIANCE keyword on the SAVE command. During the scoring phase, the item information indices will be added to this file if requested. This feature is intended for use when scoring is based on a previously created item parameter file. It must be used in conjunction with an IFNAME specification on the GLOBAL command, as shown below:

```
>GLOBAL IFNAME='EXAMPLE.PAR', SAVE;
>SAVE   COV='EXAMPLE.COV';
```

Related topics

- GLOBAL command: SAVE option (see Section 3.2.5)
- GLOBAL command: IFNAME keyword
- Output files (see Section 3.2.18)
- **Save** menu (see Section 2.6)

DIF keyword (optional)

Purpose

To specify a filename for saving the DIF parameters if requested and computed during the calibration phase (Phase 2) to an external file.

Format

```
DIF=<'filename'>
```

Default

Do not save.

Example

The DIF parameters are saved to the file **exampl03.dif** using the DIF keyword on the SAVE command.

```
>GLOBAL  DFNAME='EXAMPL03.DAT',NPARM=2, SAVE;
>SAVE    DIF='EXAMPL03.DIF';
>INPUT   NGROUPS=2, DIF, …;
```

Related topics

- GLOBAL command: SAVE option (see Section 3.2.5)
- INPUT command: DIF option (see Section 3.2.7)
- Output files (see Section 3.2.18)
- **Save** menu (see Section 2.6)

DRIFT keyword (optional)

Purpose

To specify a filename for saving the DRIFT parameters computed during the calibration phase (Phase 2) to an external file.

Format

```
DRIFT=<'filename'>
```

Default

Do not save.

Example

In the following example, the DRIFT parameters are saved to the file **exampl03.dri** using the DRIFT keyword on the SAVE command.

```
>GLOBAL  DFNAME='EXAMPL03.DAT',NPARM=2, SAVE;
>SAVE    DRIFT='EXAMPL03.DRI';
>INPUT   NGROUPS=2, DRIFT, …;
```

Related topics

- GLOBAL command: SAVE option (see Section 3.2.5)

- INPUT command: DRIFT option (see Section 3.2.7)
- Output files (see Section 3.2.18)
- **Save** menu (see Section 2.6)

EXPECTED keyword (optional)

Purpose

To specify the filename to which the expected frequencies of correct responses, attempts, and proportions of correct responses, attempts, and proportions of correct responses for each item at each quadrature point by group will be saved. This file will also contain standardized posterior residuals and model proportions of correct responses.

Format

```
EXPECTED=<'filename'>
```

Default

Do not save.

Example

In the following example, the expected frequencies are saved to **exampl03.frq** using the EXPECTED keyword on the SAVE command.

```
>GLOBAL  DFNAME='EXAMPL03.DAT',NPARM=2, SAVE;
>SAVE    EXPECTED='EXAMPL03.FRQ';
```

Related topics

- GLOBAL command: SAVE option (see Section 3.2.5)
- Output files (see Section 3.2.18)
- **Save** menu (see Section 2.6)

ISTAT keyword (optional)

Purpose

To specify a filename for saving the classical item statistics computed in Phase 1 of the program to an external file.

Format

```
ISTAT=<'filename'>
```

Default

Do not save.

Example

The classical item statistics are saved to the file **exampl03.sta** using the ISTAT keyword on the SAVE command.

```
>GLOBAL  DFNAME='EXAMPL03.DAT',NPARM=2, SAVE;
>SAVE    ISTAT='EXAMPL03.STA';
```

Related topics

- GLOBAL command: SAVE option (see Section 3.2.5)
- Output files (see Section 3.2.18)
- **Save** menu (see Section 2.6)

MASTER keyword (optional)

Purpose

To specify a filename for the master data file. The original response data are scored and stored as a temporary binary file. Upon normal termination of the program this temporary file is deleted automatically. By assigning a specific name to this master data file, the user can save and reuse it as an INPUTfile in subsequent analyses.

Format

```
MASTER=<'filename'>
```

Default

Do not save.

Example

The master file is saved to **exampl03.mas** using the MAS keyword on the SAVE command.

```
>GLOBAL  DFNAME='EXAMPL03.DAT',NPARM=2, SAVE;
>SAVE    MAS='EXAMPL03.MAS';
```

Related topics

- GLOBAL command: SAVE option (see Section 3.2.5)

- Output files (see Section 3.2.18)
- SAVE command: CALIB keyword
- **Save** menu (see Section 2.6)

PARM keyword (optional)

Purpose

To specify a filename for the item parameter file.

Item parameter estimates are saved in a formatted form as an external output file. This file can be used as initial estimates of item parameters for further iterations or as final estimates of the item parameters for scoring new data.

In either case, the user must specify the name of the previously created item parameter file using the IFNAME keyword of the GLOBAL command.

Format

```
PARM=<'filename'>
```

Default

Do not save.

Example

In the syntax below, the item parameters are saved to file through use of the SCORE and PARM keywords on the SAVE command. Note that, in order to use the SAVE command, the SAVE option is added to the GLOBAL command.

```
>GLOBAL DFNAME='EXAMPLE.DAT', NPARM=2, SAVE;
>SAVE   PARM='EXAMPLE.PAR';
>LENGTH NITEMS=(40);
```

The use of this file as initial estimates for further iterations is illustrated in the syntax below:

```
>GLOBAL DFNAME='EXAMPLE.DAT', IFNAME='EXAMPLE.PAR', NPARM=2;
>LENGTH NITEMS=(40);
```

Related topics

- GLOBAL command: IFNAME keyword (see Section 3.2.5)
- GLOBAL command: SAVE option
- Output files (see Section 3.2.18)
- **Save** menu (see Section 2.6)

PDISTRIB keyword (optional)

Purpose

To save the points and weights of the posterior latent distribution at the end of Phase 2 to an external file. These quantities can be included as prior values following the SCORE command for later EAP estimation of ability from previously estimated item parameters.

Format

```
PDISTRIB=<'filename'>
```

Default

Do not save.

Related topics

- SCORE command (see Section 3.2.14)
- **Save** menu (see Section 2.6)

POST keyword (optional)

Purpose

To save the case weight and marginal probability for each observation to an external output file.

Format

```
POST=<'filename'>
```

Default

Do not save.

Example

The case weights and marginal probabilities are saved to the file **exampl03.pos** using the POST keyword on the SAVE command.

```
>GLOBAL  DFNAME='EXAMPL03.DAT',NPARM=2, SAVE;
>SAVE    POST='EXAMPL03.POS';
```

Related topics

- GLOBAL command: SAVE option (see Section 3.2.5)
- Output files (see Section 3.2.18)
- **Save** menu (see Section 2.6)

SCORE keyword (optional)

Purpose

To specify a filename when the score file is to be saved.

Format

```
SCORE=<'filename'>
```

Default

Do not save.

Example

In the following example, the score file is saved to **exampl03.sco** using the SCORE keyword on the SAVE command.

```
>GLOBAL  DFNAME='EXAMPL03.DAT',NPARM=2, SAVE;
>SAVE    SCORE='EXAMPL03.SCO';
```

Related topics

- GLOBAL command: SAVE option (see Section 3.2.5)
- Output files (see Section 3.2.18)
- **Save** menu (see Section 2.6)

TSTAT keyword (optional)

Purpose

To specify a filename when the tables of test information statistics are to be saved.

Format

```
TSTAT=<'filename'>
```

Default

Do not save.

Example

The test information statistics file is saved to **exampl03.tsa** using the TSTAT keyword on the SAVE command.

```
>GLOBAL  DFNAME='EXAMPL03.DAT',NPARM=2, SAVE;
>SAVE    TSTAT='EXAMPL03.TSA';
```

Related topics

- GLOBAL command: SAVE option (see Section 3.2.5)
- Output files (see Section 3.2.18)
- **Save** menu (see Section 2.6)

3.2.14 SCORE command

(Optional)

Purpose

To initiate the scoring of individual examinees or of response patterns; to compute item and test information and plot information curves; to rescale scores to a specified mean and standard deviation in either the sample or the latent distribution.

Format

```
>SCORE METHOD=n, NQPT=(list), IDIST=n, PMN=(list), PSD=(list), RSCTYPE=n,
       LOCATION=(list), SCALE=(list), INFO=n, BIWEIGHT, FIT, NOPRINT,
       YCOMMON, POP, MOMENTS, FILE, READF, REFERENCE=n, NFORMS=n;
```

Examples

The aggregate scores for the following analysis of school-level data are estimated by the EAP method using the empirical distributions from Phase 2. The number of quadrature points is set to 12 per subtest.

The scores are rescaled to a mean of 250 and a standard deviation of 50 in the latent distribution of schools (IDIST=3, LOCATION=250, SCALE=50). The fit of the data to the group-level model is tested for each school (FIT).

```
>SCORE NQPT=(12, 12), IDIST=3, RSCTYPE=4, LOCATION=(250.0000, 250.0000),
       SCALE=(50.0000, 50.0000), FIT;
```

The next SCORE command gives the specifications for a scoring phase that includes an information analysis (INFO=2) with expected information indices for a normal population (POP). Rescaling of the scores and item parameters to mean 0 and standard deviation 1 in the estimated latent distribution has been requested (RSC=3). Printing of the students' scores on the screen is suppressed (NOPRINT).

```
>SCORE NQPT=6, NOPRINT, RSCTYPE=3, INFO=2, POP;
```

In the following SCORE command, the EAP scale scores of Phase 3 are computed from the responses to items in the main test as specified by setting METHOD to 2. Printing of scores is suppressed (NOPRINT).

```
>SCORE   METHOD=2,NOPRINT;
```

In this score command, Maximum Likelihood estimates of ability (METHOD=1) are rescaled to a mean of 250 and standard deviation of 50 in Phase 3 (RSCTYPE=3, LOCATION=250, SCALE=50).

```
>SCORE METHOD=1, RSCTYPE=3, LOCATION=(250.0000), SCALE=(50.0000), INFO=1,
       NOPRINT;
```

Related topics

- **Phase 2: CALIBRATE** (see Section 1.2)
- **Setup** menu: **Test Scoring** dialog box (see Section 2.3)
- **Technical** menu: **Score Options** dialog box (see Section 2.5)
- **Technical** menu: **Scoring Prior Latent Distribution** dialog box

BIWEIGHT option (optional)

Purpose

To request the calculation of biweighted estimates robust to isolated deviant responses. (See also Mislevy & Bock, 1982.)

Format

```
BIWEIGHT
```

Related topics

- **Setup** menu: **Test Scoring** dialog box (see Section 2.3)

DOMAIN keyword (optional)

Purpose

To convert the Phase 3 estimates into domain scores if the user supplies a file containing the item parameters for a sample of previously calibrated items. The FILE keyword on the SCORE command is used to specify this parameter file. Weights can be applied to the items to improve the representation of the domain specifications. This conversion may be useful as an aid to the interpretation of test results (see Bock, Thissen, & Zimowski (1997).)

Note that the formula for the domain scores that appears in the paper cited here contains typographical errors. The computation of the domain scores in the program uses the corrected formula. The domain scores will appear in the score listing following the test percent correct score for each case in the Phase 3 output file.

Format

```
DOMAIN=n
```

where n represents the number of items in the domain.

Default

No domain scores.

Related topics

- SCORE command: FILE keyword
- **Technical** menu: **Score Options** dialog box (see Section 2.5)

FILE keyword (required if DOMAIN keyword is used)

Purpose

To specify the name of the file containing the item parameters to be used for the domain score conversions.

The first line of the file referenced by the FILE keyword must contain a variable format statement (in parentheses) describing the column layout of the weights and parameter in the following lines of the file. The values must be read in order—item weight, slope, threshold, and guessing parameter. The weights will be automatically scaled to sum to 1.0 by the program. The domain score will appear in the score listing following the test percent correct score for each case. Note that the parameter file produced by the SAVE command does not have the layout described above.

Format

```
FILE=<'filename'>
```

Default

No domain scores or supplied file.

Related topics

- SCORE command: DOMAIN keyword
- **Technical** menu: **Score Options** dialog box (see Section 2.5)

FIT option (optional)

Purpose

To request the computation of a likelihood ratio χ^2 goodness-of-fit statistic for each response pattern. This statistic is intended only for use with aggregate-level data.

Format

```
FIT
```

Default

No fit statistic.

Example

The aggregate scores for this analysis of school-level data are estimated by the EAP method using the empirical distributions from Phase 2. The fit of the data to the group-level model is tested for each school (FIT).

```
>SCORE NQPT=(12, 12), IDIST=3, RSCTYPE=4, LOCATION=(250.0000, 250.0000),
       SCALE=(50.0000, 50.0000), FIT;
```

Related topics

- **Setup** menu: **Test Scoring** dialog box (see Section 2.3)

IDIST keyword (optional)

Purpose

To designate the type of prior distribution of scale scores. IDIST = 0 applies to both MAP and EAP estimation. IDIST = 1, 2, 3, or 4 applies only to EAP estimation.

Format

```
IDIST=n
```

- n = 0 Standard normal approximation.
- n = 1 Separate, arbitrary discrete priors for each group for each subtest read

from QUADSj command.

- n = 2 Separate arbitrary discrete priors for each group read from QUADSj command.
- n = 3 Empirical prior estimated during Phase 2.
- n = 4 35-point rectangular prior on the interval ± 3.5. These scores may be transformed to the 150 – 850 range by setting LOCATION = 500.0 and SCALE = 100.0.

Default

0.

Examples

In the following aggregate-level example, IDIST=3 is used to estimate scores by the EAP method by using the empirical distributions from Phase 2.

```
>SCORE NQPT=(12, 12), IDIST=3, RSCTYPE=4, LOCATION=(250.0000, 250.0000),
       SCALE=(50.0000, 50.0000), FIT;
```

In the next example, EAP estimates of ability are calculated (METHOD=2) using the information in the posterior distributions from Phase 2 (IDIST=3). The ability estimates are rescaled to a mean of 0 and standard deviation of 1 by specifying RSCTYPE=3 on the SCORE command.

```
>SCORE   METHOD=2, IDIST=3, NOPRINT, RSCTYPE=3;
```

Related topics

- QUADS command (see Section 3.2.12)
- **Setup** menu: **Test Scoring** dialog box (see Section 2.3)
- **Technical** menu: **Scoring Prior Latent Distribution** dialog box (see Section 2.5)

INFO keyword (optional)

Purpose

To select information output.

Format

```
INFO=n
```

- n = 0 none
- n = 1 test information curves
- n = 2 test information curves and table of information statistics

Default

0.

Examples

The following SCORE command gives the specifications for a scoring phase that includes an information analysis (INFO=2) with expected information indices for a normal population (POP).

```
>SCORE NQPT=6, NOPRINT, RSCTYPE=4, INFO=2, POP;
```

In the following SCORE command, Maximum Likelihood estimates of ability (METHOD=1) are rescaled to a mean of 250 and standard deviation of 50 in Phase 3.

```
>SCORE METHOD=1, RSCTYPE=3, LOCATION=(250.0000), SCALE=(50.0000),
       INFO=1, NOPRINT;
```

Related topics

- SCORE command: POP and YCOMMON options

LOCATION keyword (optional)

Purpose

To specify real-valued location constants (with decimal points) for rescaling.

Format

LOCATION=($n_1, n_2, \ldots, n_{NTEST}$)

Default

0.0.

Examples

The scores are rescaled to a mean of 250 and a standard deviation of 50 in the latent distribution of schools (IDIST=3, LOCATION=250, SCALE=50). The fit of the data to the group-level model is tested for each school (FIT).

```
>SCORE NQPT=(12, 12), IDIST=3, RSCTYPE=4, LOCATION=(250.0000, 250.0000),
       SCALE=(50.0000, 50.0000), FIT;
```

In the next SCORE command, Maximum Likelihood estimates of ability (METHOD=1) are rescaled to a mean of 250 and standard deviation of 50 in Phase.

```
>SCORE METHOD=1, RSCTYPE=3, LOCATION=(250.0000), SCALE=(50.0000),
       INFO=1, NOPRINT;
```

Related topics

- SCORE command: RSCTYPE keyword
- SCORE command: SCALE keyword
- **Setup** menu: **Test Scoring** dialog box (see Section 2.3)

METHOD keyword (optional)

Purpose

To specify the method of estimating scale scores. If ML is selected, it is advisable to use the PMN keyword to set bounds on the estimated scores. If EAP or MAP is selected, the PMN and PSD keywords may be used to specify the means and standard deviations of the prior distributions.

Format

```
METHOD=n
```

- n = 1 Maximum likelihood (ML)
- n = 2 Expected a posteriori (EAP) (Bayes)
- n = 3 Maximum a posteriori (MAP) (Bayes modal)

Default

2.

Examples

In this score command, Maximum Likelihood estimates of ability (METHOD=1) are rescaled to a mean of 250 and standard deviation of 50 in Phase 3 (RSCTYPE=3, LOCATION=250, SCALE=50).

```
>SCORE METHOD=1, RSCTYPE=3, LOCATION=(250.0000), SCALE=(50.0000),
       INFO=1, NOPRINT;
```

Related topics

- SCORE command: PMN keyword
- SCORE command: PSD keyword
- **Setup** menu: **Test Scoring** dialog box (see Section 2.3)

MOMENTS option (optional)

Purpose

To request the computation and listing of the coefficients of skewness and kurtosis of the ability estimates and of the latent distribution.

Format

```
MOMENTS
```

Default

No computation or listing.

Examples

The MOMENTS keyword on the SCORE commands below is used to obtain the coefficients of skewness and kurtosis for the rescaled ability.

```
>SCORE   NQPT=11,RSCTYPE=3,LOCATION=250,SCALE=50,NOPRINT,INFO=1,
         POP,MOMENT;

>SCORE   IDIST=3,RSCTYPE=3,INFO=1,YCOMMON,POP,NOPRINT,MOMENTS;
```

Related topics

- **Technical** menu: **Score Options** dialog box (see Section 2.5)

NFORM keyword (optional)

Purpose

To indicate the number of additional FORM commands after the SCORE command. It is used when scoring is to be performed using these additional form specifications. The reference form for scoring is set using the REFERENCE keyword on the SCORE command.

Format

```
NFORMS=n
```

Default

No additional FORM commands are expected.

Example

In the example below, two additional form commands follow the SCORE command. The first is the references group (as set by the REFERENCE keyword) while the READF keyword instructs the program to read and process the additional FORM commands.

```
>SCORE IDIST=3,RSCTYPE=3,INFO=1,YCOMMON,POP,NOPRINT,REF=1,NFORMS=2,READF;
>FORM1 LENGTH=25,INUM=(1(1)25);
>FORM2 LENGTH=25,INUM=(21(1)45);
```

Related topics

- FORM command (see Section 3.2.4)
- SCORE command: READF option
- SCORE command: REFERENCE keyword

NOPRINT option (optional)

Purpose

To suppress the display of the scores on screen and in the printed output of Phase 3.

To shorten the run time for scoring a large subject response file, it is advisable to specify an external file using the SCORE keyword in the SAVE command, and the NOPRINT option. In this way, scores for all subjects are computed but are stored only in the external file.

Format

```
NOPRINT
```

Default

Scores will appear both on screen and in the Phase 3 output file.

Examples

The EAP scale scores of Phase 3 are computed from the responses to items in the main test as specified by setting METHOD to 2. Printing of scores is suppressed (NOPRINT).

```
>SCORE  METHOD=2, NOPRINT;
```

Related topics

- **Setup** menu: **Test Scoring** dialog box (see Section 2.3)

NQPT keyword (optional)

Purpose

To set the number of quadrature points for each subtest with the NQPT keyword when EAP estimation is selected by the METHOD keyword.

To reduce computing time when there are items not-presented, use 2 x square root of the maximum number of items per respondent as the number of quadrature points.

Format

NQPT=($n_1, n_2, \ldots, n_{NTEST}$)

Default

Computed by program as a function of number of items in complete data.

Example

The aggregate scores for this analysis of school-level data are estimated by the EAP method using the empirical distributions from Phase 2. The number of quadrature points is set to 12 per subtest.

```
>SCORE NQPT=(12, 12), IDIST=3, RSCTYPE=4, LOCATION=(250.0000, 250.0000),
       SCALE=(50.0000, 50.0000), FIT;
```

Related topics

- GLOBAL command: NTEST keyword (see Section 3.2.5)
- SCORE command: METHOD keyword
- **Technical** menu: **Scoring Prior Latent Distribution** dialog box (see Section 2.5)

PMN keyword (optional)

Purpose

To specify real-numbered means (with decimal points) of the normal prior distributions for each group for each subtest.

Format

PMN=($n_{1,1}, n_{1,2}, \ldots, n_{1,NGROUP}, n_{2,1}, n_{2,2}, \ldots, n_{2,NGROUP}, \ldots, n_{NTEST,1}, n_{NTEST,2}, \ldots, n_{NTEST,NGROUP}$)

Default

0.0.

Example

In the following two-group analysis for one subtest, the PMN and PSD keywords are used on the SCORE command to provide the means and standard deviations of the normal prior distributions for each subtest.

```
>SCORE PMN=(0.00051,-0.16191), PSD=(0.00001,0.89707);
```

Related topics

- GLOBAL command: NTEST keyword (see Section 3.2.5)
- INPUT command: NGROUP keyword (see Section 3.2.7)
- SCORE command: PSD keyword
- **Technical** menu: **Scoring Prior Latent Distribution** dialog box (see Section 2.5)

POP option (optional)

Purpose

To request the calculation of the expected information for the population when INFO > 0. This includes an estimate of the classical reliability coefficient for each subtest. The score metric after rescaling is used in these calculations.

Format

```
POP
```

Default

No expected information calculated for population.

Example

This SCORE command gives the specifications for a scoring phase that includes an information analysis (INFO=2) with expected information indices for a normal population (POP). Rescaling of the scores and item parameters to mean 0 and standard deviation 1 in the estimated latent distribution has been requested (RSC=3). Printing of the students' scores on the screen is suppressed (NOPRINT).

```
>SCORE NQPT=6, NOPRINT, RSCTYPE=4, INFO=2, POP;
```

Related topics

- SCORE command: INFO keyword
- Phase 3: SCORING (see Section 1.2)

PSD keyword (optional)

Purpose

To specify real-numbered standard deviations (with decimal points) of the normal prior distributions for each group for each subtest.

Format

PSD=($n_{1,1}, n_{1,2}, ..., n_{1,NGROUP}, n_{2,1}, n_{2,2}, ..., n_{2,NGROUP}, ..., n_{NTEST,1}, n_{NTEST,2}, ..., n_{NTEST,NGROUP}$)

Default

1.0.

Example

In the following two-group analysis for one subtest, the PMN and PSD keywords are used on the SCORE command to provide the means and standard deviations of the normal prior distributions for each subtest.

```
>SCORE PMN=(0.00051, -0.16191), PSD=(0.00001, 0.89707);
```

Related topics

- GLOBAL command: NTEST keyword (see Section 3.2.5)
- INPUT command: NGROUP keyword (see Section 3.2.7)
- SCORE command: PMN keyword
- **Technical** menu: **Scoring Prior Latent Distribution** dialog box (see Section 2.5)

READF option (optional)

Purpose

To indicate the presence of multiple FORM commands after the SCORE command. It is used to indicate that scoring is to be performed using this form specification. The reference form for scoring is set using the REFERENCE keyword on the SCORE command.

Format

```
READF
```

Default

No additional FORM commands are expected.

Example

In the example below, two additional form commands follow the SCORE command. The first is the references group (as set by the REFERENCE keyword) while the READF keyword instructs the program to read and process the additional FORM commands.

```
>SCORE IDIST=3,RSCTYPE=3,INFO=1,YCOMMON,POP,NOPRINT,REF=1,NFORMS=2,READF;
>FORM1 LENGTH=25,INUM=(1(1)25);
>FORM2 LENGTH=25,INUM=(21(1)45);
```

Related topics

- FORM command (see Section 3.2.4)
- SCORE command: REFERENCE and NFORM keywords

REFERENCE keyword (optional)

Purpose

To set the reference group for scoring when scoring is performed by forms, as specified with the READF and NFORM keywords on the same command. Note that, if this keyword is omitted while the READF and NFORM keywords are present, the reference form specified in the CALIB command will be used.

Format

```
REFERENCE=n
```

Default

Set by REFERENCE keyword on CALIB command.

Example

In the example below, two additional form commands follow the SCORE command. The first is the references group (as set by the REFERENCE keyword) while the READF keyword instructs the program to read and process the additional FORM commands.

```
>SCORE IDIST=3,RSCTYPE=3,INFO=1,YCOMMON,POP,NOPRINT,REF=1,NFORMS=2,READF;
>FORM1 LENGTH=25,INUM=(1(1)25);
>FORM2 LENGTH=25,INUM=(21(1)45);
```

Related topics

- CALIB command: REFERENCE keyword (see Section 3.2.2)
- SCORE command: READF option
- SCORE command: NFORM keyword

RSCTYPE keyword (optional)

Purpose

To specify the type of rescaling required.

Format

```
RSCTYPE=n
```

Uses the LOCATION and SCALE constants specified by the options below. Note that there is no option 2.

- 0: no rescaling
- 1: linear transformation of scores: new score = SCALE x old score + LOCATION
- 3: rescale to SCALE and LOCATION in the sample of scale score estimates
- 4: only if EAP estimation has been selected: Set the mean of the latent population distribution equal to LOCATION and set the standard deviation equal to SCALE.

Default

0.

Examples

The aggregate scores for this analysis of school-level data are estimated by the EAP method using the empirical distributions from Phase 2. The number of quadrature points is set to 12 per subtest. The scores are rescaled to a mean of 250 and a standard deviation of 50 in the latent distribution of schools (IDIST=3, LOCATION=250, SCALE=50). The fit of the data to the group-level model is tested for each school (FIT).

```
>SCORE NQPT=(12, 12), IDIST=3, RSCTYPE=4, LOCATION=(250.0000, 250.0000),
       SCALE=(50.0000, 50.0000), FIT;
```

Related topics

- SCORE command: LOCATION and SCALE keywords
- **Setup** menu: **Test Scoring** dialog box (see Section 2.3)

SCALE keyword (optional)

Purpose

To specify real-valued scale constants (with decimal points) for rescaling.

Format

SCALE=($n_1, n_2, ..., n_{NTEST}$)

Default

1.0.

Examples

In the following example, Maximum Likelihood estimates of ability (METHOD=1) are re-scaled to a mean of 250 and standard deviation of 50 in Phase 3 (RSCTYPE=3, LOCATION=250, SCALE=50).

```
>SCORE METHOD=1, RSCTYPE=3, LOCATION=(250.0000), SCALE=(50.0000),
       INFO=1, NOPRINT;
```

Related topics

- SCORE command: LOCATION and RSCTYPE keywords
- **Setup** menu: **Test Scoring** dialog box (see Section 2.3)

YCOMMON option (optional)

Purpose

To specify that the test information curves for subtests should be expressed in comparable units when INFO > 0. If INFO = 0, the curves for subsets will be adjusted separately to make each plot fill the available space.

Format

YCOMMON

Default

Plots adjusted separately.

Example

The following SCORE command specifies a scoring phase that includes an information analysis (INFO=2) with expected information indices for a normal population (POP).

Test information curves for subtests will be expressed in comparable units and printed to the Phase 3 output file.

```
>SCORE INFO=2, POP, YCOMMON;
```

Related topics

- SCORE command: INFO keyword
- SCORE command: POP option

3.2.15 TEST command

(Required)

Purpose

To identify the main test items and the variant test items (if any) in each of the NTEST subtests. If the subtest contains only main test items, there is only one TEST command for that subtest. If there are variant items in the subtest, two TEST commands are required for that subtest. The first describes the main test items, while the second describes the variant test items. There are as many TEST commands as there are main and variant subtests specified in the NTEST and NVTEST keywords of the GLOBAL command.

Items may be identified by name or number, but not by both. The names or numbers must correspond to those listed in the ITEMS command. If numbers are supplied, the program will refer to the names supplied in the ITEMS command only for printing of item information. Starting values for estimating the item parameters may also be supplied in the TEST command. Note that parameter estimation for variant items is non-iterative and does not require starting values.

Format

```
>TEST TNAME=n, INUMBER=(list), INAME=(list), INTERCPT=(list),
      SLOPE=(list), THRESHLD=(list), GUESS=(list), DISPERSN=(list),
      FIX=(list);
```

Default

All items are used.

Examples

In the example below, two subtests are used, each with 8 items. The NTEST keyword on the GLOBAL command indicates that two subtests are to be used, and two TEST commands follow the ITEMS command. The TEST commands are assigned names through the TNAME keyword and items are referenced by number.

```
>GLOBAL  NPARM=3,NTEST=2,DFNAME='EXAMPL08.DAT';
>LENGTH  NITEMS=(8,8);
>INPUT   NTOTAL=16,NALT=5,NIDCHAR=9,TYPE=3;
>ITEMS   INUMBERS=(1(1)16),INAMES=(N1(1)N8,A1(1)A8);
>TEST1   TNAME=NUMCON,INUMBERS=(1(1)8);
>TEST2   TNAME=ALGCON,INUMBERS=(9(1)16);
```

In the next example, the ITEMS command lists the four items in the order that they will be read from the data records. The INAMES and INUMBERS keywords assign each item a name and a corresponding number. Because there is only one form, the NFORM keyword is not required in the INPUT command and a FORM command is not required. Because examinees in both groups are presented all the items listed in the ITEMS command, the TEST command needs contain only the test name.

```
>GLOBAL  NPARM=1,NWGHT=3,LOGISTIC;
>LENGTH  NITEMS=4;
>INPUT   NTOTAL=4,NGROUPS=2,DIF,NIDCHAR=2,TYPE=2;
>ITEMS   INAMES=(SP1(1)SP4),INUMBERS=(1(1)4);
>TEST    TNAME=SPELL;
>GROUP1  GNAME=MALES;
>GROUP2  GNAME=FEMALES;
```

Related topics

- GLOBAL command NTEST and NVTEST keywords (see Section 3.2.5)
- ITEMS command (see Section 3.2.8)
- **Setup** menu: **General** dialog box (see Section 2.3)
- **Setup** menu: **Item Analysis** dialog box
- **Technical** menu: **Item Parameter Starting Values** dialog box (see Section 2.5)
- **Technical** menu: **Assign Fixed Items** dialog box

DISPERSN keyword (optional)

Purpose

To specify positive real-numbered starting values for dispersion (2- and 3-parameter models only).

Starting values may be specified for slopes or for dispersions, but not for both.

Format

DISPERSN=($n_1, n_2, ..., n_{n(i)}$)

Default

1/slope.

Example

In the syntax below, starting values for the dispersion and intercepts of the four items considered in this 3-parameter model are provided on the TEST command.

```
EXAMPLE:
USING STARTING VALUES
>GLOBAL  NPARM=3,LOGISTIC';
>LENGTH  NITEMS=4;
>INPUT   NTOTAL=4,NIDCHAR=2;
>ITEMS   INAME=(SP01,SP02,SP03,SP04),INUMBERS=(1(1)4);
>TEST    TNAME=SPELL,
         INTERCPT = (1.284,0.287,-1.912,-0.309),
         DISPERSN=(0.957,0.623,0.545,0.620);
```

Related topics

- **Technical** menu: **Item Parameter Starting Values** dialog box (see Section 2.5)
- TEST command: SLOPE keyword

FIX keyword (optional)

Purpose

To specify whether the parameters of specific items are free to be estimated or are to be held fixed at their starting values. This keyword appears in the *j*-th TEST command as FIX=($n_1, n_2, ..., n_{LENGTH(j)}$) where

- n_i = 0 if parameters of item *i* of test *j* are free to be estimated, or
- n_i =1 if these item parameters are to be held fixed at their starting values.

The starting values may be entered by the SLOPE, THRESHLD, and GUESSING keywords of the *j*-th TEST command, or read from an existing item parameter file (IFNAME) designated by IFNAME=<'filename'> on the GLOBAL command and saved in a previous job by the PARM=<'filename'> keyword on the SAVE command; or, alternatively, read from a file of provisional item parameters, designated by the PRNAME=<'filename'> keyword on the GLOBAL command. When only a few items are to be fixed, this method is the most convenient. If all items are designated as fixed, and the INFO keyword appears on the SCORE command, the required information and reliability analysis will be performed in Phase 3.

In order for this procedure to work, however, the program must have data to process in Phases 1 and 2 for at least a few cases. Some artificial response data can be used for this purpose. The only calculations that will be performed in Phase 2 are preparations for the information analysis in Phase 3. The number of EM cycles in the CALIB command can therefore be set to 2 and the number of NEWTON cycles to 1. The NOADJUST option must also be invoked.

Format

FIX=($u_1, u_2, ..., u_{LENGTH(j)}$)

Default

Do not fix.

Example

The following command file shows the fixing of five items by specifying values in a PRNAME file.

```
EXAMPLE 16: TRADITIONAL IRT ANALYSIS OF A FIFTEEN-ITEM TEST
            PARAMETERS OF ITEMS 6 THROUGH 10 ARE FIXED
>GLOBAL  NPARM=3, DFNAME='EXAMPL07.DAT',PRNAME='EXAMPL7f.PRM',SAVE;
>SAVE    SCORE='EXAMPL7.SCO',PARM='EXAMPL7.PAR';
>LENGTH  NITEMS=15;
>INPUT   NTOTAL=15,NALT=5,KFNAME='EXAMPL07.KEY',SAMPLE=600,NIDCHAR=4;
>ITEMS   INAME=(MATH01(1)MATH15);
>TEST    TNAME=PRETEST,FIX=(0(0)5,1(0)5,0(0)5);
         (2X,4A1,T25,15A1)
>CALIB   CYCLES=15,NEWTON=3,NQPT=11,NOADJUST;
>SCORE   NQPT=11,RSCTYPE=3,LOCATION=250,SCALE=50,NOPRINT,INFO=1,POP;
```

The **exampl7f.prm** file contains the following 6 lines:

```
  5
  6  1.27168     0.10504   0.14011
  7  1.79009     0.10221   0.07543
  8  0.81238     0.24523   0.22179
  9  1.33017    -0.22387   0.15453
 10  1.06557     0.58430   0.08921
```

DIAGNOS has been set equal to 1 to produce more detailed output which show that these values do not change during the Phase 2 estimation cycles. They will, of course, be rescaled along with those of the estimated items in Phase 3.

Related topics

- CALIB command: NOADJUST option (see Section 3.2.2)
- GLOBAL command: IFNAME and PRNAME keywords (see Section 3.2.5)
- SAVE command: PARM keyword (see Section 3.2.13)
- SCORE command: INFO keyword (see Section 3.2.14)
- TEST command: GUESS, SLOPE, and THRESHLD keywords
- **Technical** menu: **Assign Fixed Items** dialog box (see Section 2.5)

GUESS keyword (optional)

Purpose

To specify starting values for the lower asymptote (guessing) parameters (3-parameter model only). These values should be positive fractional numbers with decimal points.

Format

GUESS=($n_1, n_2, ..., n_{n(i)}$)

Default

0.0.

Example

In the syntax below, starting values for the slopes and guessing parameters of the four items considered in this 3-parameter model are provided on the TEST command.

```
EXAMPLE:
USING STARTING VALUES
>GLOBAL  NPARM=3,LOGISTIC';
>LENGTH  NITEMS=4;
>INPUT   NTOTAL=4,NIDCHAR=2;
>ITEMS   INAME=(SP01,SP02,SP03,SP04),INUMBERS=(1(1)4);
>TEST    TNAME=SPELL,
         SLOPE=(1.045,1.604,1.836,1.613),
         GUESS=(0.189,0.168,0.101,0.152);
```

Related topics

- GLOBAL command: NPARM keyword (see Section 3.2.5)
- **Technical** menu: **Item Parameter Starting Values** dialog box (see Section 2.5)

INAMES keyword (optional)

Purpose

To provide a list of names, as specified in the ITEMS command for items in TEST. Item names that do not begin with letters must be enclosed in single quotes.

Format

INAME=($n_1, n_2, ..., n_{n(i)}$)

Default

If NTEST =1, and NVTEST = 0, all NTOTAL items are as specified in the INPUT command. There is no default if NTEST > 1 or NVTEST $\neq 0$.

Example

In the following example, responses to 50 items are read from those of 100 items in the data file. From the 50, 20 are selected as Main Test items and 4 as Variant Test items. Items for the main test are selected by name in the TESTM command; items for the variant test are selected by name in the TESTV command. The item names correspond to the sequence numbers in the original set of 100 items.

```
>GLOBAL  DFNAME='EXAMPL06.DAT', NTEST=1,NVTEST=1,NPARM=2;
>LENGTH  NITEM=24,NVARIANT=4;
>INPUT   NTOTAL=50,KFNAME='EXAMPL06.DAT',SAMPLE=200,NIDCH=11;
>ITEMS   INUMBERS=(1(1)50),INAME=(I26(1)I75);
>TESTM   TNAME=MAINTEST,
         INAMES=(I26,I27,I28,I29,I31,I33,I34,
         I35,I36,I38,I39,I47,I48,I49,I50,I54,I60,I64,I68,I72);
>TESTV   TNAME=VARIANT,INAMES=(I53,I59,I69,I73);
```

Related topics

- GLOBAL command: NTEST and NVTEST keywords (see Section 3.2.5)
- INPUT command: NTOTAL keyword (see Section 3.2.7)
- ITEMS command (see Section 3.2.8)
- LENGTH command (see Section 3.2.9)
- **Setup** menu: **General** dialog box (see Section 2.3)

INTERCPT keyword (optional)

Purpose

To specify real-numbered starting values (with decimal points) for estimating the item intercept. Starting values may be specified for intercepts or for thresholds, but not for both.

Format

`INTERCPT=(` $n_1, n_2, ..., n_{n(i)}$ `)`

Default

Supplied by the program.

Example

In the syntax below, starting values for the intercepts of the four items considered in this 3-parameter model are provided on the TEST command.

```
EXAMPLE:
USING STARTING VALUES
>GLOBAL  NPARM=3,LOGISTIC';
>LENGTH  NITEMS=4;
>INPUT   NTOTAL=4,NIDCHAR=2;
>ITEMS   INAME=(SP01,SP02,SP03,SP04),INUMBERS=(1(1)4);
>TEST    TNAME=SPELL,INTERCPT = (1.284,0.287,-1.912,-0.309);
```

Related topics

- TEST command: THRESHLD keyword
- **Technical** menu: **Item Parameter Starting Values** dialog box (see Section 2.5)

INUMBERS keyword (optional)

Purpose

To provide a list of numbers, as specified in the ITEMS command for items in TEST. If TEST refers to main test items, $n(i)$ is the number of main test items. If TEST refers to variant test items, $n(i)$ is the number of variant test items.

The notation "first (increment) last" in these lists may be used when the item numbers form an arithmetic progression.

Format

`INUMBER=(` $n_1, n_2, ..., n_{n(i)}$ `)`

Default

If NTEST=1, and NVTEST=0, all NTOTAL items as specified in the INPUT command. There is no default if NTEST>1 of NVTEST $\neq 0$.

Examples

For the case where NTEST=1 and NVTEST=1 in the GLOBAL command, NITEMS=10 and NVARIANT=4 in the LENGTH command, and NTOT=10 in the INPUT command, the main test items of subtest *i* might be specified in the first TEST command with ITEMS=(1,2,3,6,8,10). The variant test items of subtest *i* might be specified in the second TEST command with ITEMS=(4,5,7,9).

In the example below, two subtests are used, each with 8 items. The NTEST keyword on the GLOBAL command indicates that two subtests are to be used, and two TEST commands follow the ITEMS command. The subtests are assigned names through the TNAME keyword and items are referenced by number.

```
>GLOBAL  NPARM=3,NTEST=2,DFNAME='EXAMPL08.DAT';
>LENGTH  NITEMS=(8,8);
>INPUT   NTOTAL=16,NALT=5,NIDCHAR=9,TYPE=3;
>ITEMS   INUMBERS=(1(1)16),INAMES=(N1(1)N8,A1(1)A8);
>TEST1   TNAME=NUMCON,INUMBERS=(1(1)8);
>TEST2   TNAME=ALGCON,INUMBERS=(9(1)16);
```

Related topics

- GLOBAL command: NTEST or NVTEST keywords (see Section 3.2.5)
- ITEMS command (see Section 3.2.8)
- LENGTH command (see Section 3.2.9)
- **Setup** menu: **Item Analysis** dialog box (see Section 2.3)

SLOPE keyword (optional)

Purpose

To provide starting values for slopes (2- and 3-parameter models only). These starting values should be positive, real numbers with decimal points. Starting values may be specified for slopes or for dispersions, but not for both.

Format

SLOPE=($n_1, n_2, ..., n_{n(i)}$)

Default

1.0.

Example

In the syntax below, starting values for the intercepts and slopes of the four items considered in this 3-parameter model are provided on the TEST command.

```
EXAMPLE:
USING STARTING VALUES
>GLOBAL  NPARM=3,LOGISTIC';
>LENGTH  NITEMS=4;
>INPUT   NTOTAL=4,NIDCHAR=2;
>ITEMS   INAME=(SP01,SP02,SP03,SP04),INUMBERS=(1(1)4);
>TEST    TNAME=SPELL,
         INTERCPT = (1.284,0.287,-1.912,-0.309),
         SLOPE=(1.045,1.604,1.836,1.613);
```

Related topics

- **Technical** menu: **Item Parameter Starting Values** dialog box (see Section 2.5)
- TEST command: DISPERSN keyword

THRESHLD keyword (optional)

Purpose

To specify real-numbered starting values (with decimal points) for estimating the item thresholds. Starting values may be specified for intercepts or for thresholds, but not for both.

Format

THRESHLD=($n_1, n_2, ..., n_{n(i)}$)

Default

0.0.

Example

In the syntax below, starting values for the slopes and thresholds of the four items considered in this 3-parameter model are provided on the TEST command.

```
EXAMPLE:
USING STARTING VALUES
>GLOBAL  NPARM=3,LOGISTIC';
>LENGTH  NITEMS=4;
>INPUT   NTOTAL=4,NIDCHAR=2;
>ITEMS   INAME=(SP01,SP02,SP03,SP04),INUMBERS=(1(1)4);
>TEST    TNAME=SPELL, SLOPE=(1.045,1.604,1.836,1.613),
         THRESHLD=(-1.229,-0.179,1.041,0.192);
```

Related topics

- **Technical** menu: **Item Parameter Starting Values** dialog box (see Section 2.5)
- TEST command: INTERCPT keyword

TNAME keyword (optional)

Purpose

To supply a name for subtest *i* (up to eight characters), if there are not variant test items in subtest *i*; or name of the main test items in subtest *i,* if there are variant test items in subtest *i*.

Format

```
TNAME=character string
```

Default

None.

Examples

In the example below, two subtests are used, each with 8 items. The NTEST keyword on the GLOBAL command indicates that two subtests are to be used, and two TEST commands follow the ITEMS command. The TEST commands are assigned names through the TNAME keyword and items are referenced by number.

```
>GLOBAL  NPARM=3,NTEST=2,DFNAME='EXAMPL08.DAT';
>LENGTH  NITEMS=(8,8);
>INPUT   NTOTAL=16,NALT=5,NIDCHAR=9,TYPE=3;
>ITEMS   INUMBERS=(1(1)16),INAMES=(N1(1)N8,A1(1)A8);
>TEST1   TNAME=NUMCON,INUMBERS=(1(1)8);
>TEST2   TNAME=ALGCON,INUMBERS=(9(1)16);
```

In the next example, the ITEMS command lists the four items in the order that they will be read from the data records. Because examinees in both groups are presented all the items listed in the ITEMS command, the TEST command needs contain only the test name.

```
>GLOBAL  NPARM=1,NWGHT=3,LOGISTIC;
>LENGTH  NITEMS=4;
>INPUT   NTOTAL=4,NGROUPS=2,DIF,NIDCHAR=2,TYPE=2;
>ITEMS   INAMES=(SP1(1)SP4),INUMBERS=(1(1)4);
>TEST    TNAME=SPELL;
>GROUP1  GNAME=MALES;
>GROUP2  GNAME=FEMALES;
```

Related topics

- GLOBAL command: NTEST and NVTEST keywords (see Section 3.2.5)
- ITEMS command (see Section 3.2.8)
- **Setup** menu: **General** dialog box (see Section 2.3)
- TEST command: INUMBERS keyword

3.2.16 TITLE command

(Required)

Purpose

To provide a label that will be used throughout the output to identify the problem run. The first two lines of the command file are always the title lines. If the title fits on one line, a second, blank line must be entered before the next command starts.

The maximum length of each line is 80 characters. The text will be printed verbatim at the top of each output section, as well as at the start of some output files. The two title lines are required at the start of the command file. No special delimiters (> or ;) are required.

Format

```
…text…
…text…
```

Example

```
EXAMPLE 4
SIMULATED RESPONSES TO TWO 20-ITEM PARALLEL TEST FORMS
```

Related topics

- **Setup** menu: **General** dialog box (see Section 2.3)

3.2.17 Variable format statement

(Required)

Purpose

To supply variable format statements describing the column assignments of fields in the data records.

Format

```
(aA1,nX,Ib,Ic,Fd.e,Tw,fA1)
```

where:

a is the number of columns in the ID field

b is the number of columns in the form indicator field, if any.

c is the number of columns in the group indicator field, if any. If PERSONAL is present on the INPUT command there will be multiple-group indicator fields.

d is the number of columns in the case-weight or pattern-frequency field, if any.

e is the number of columns to the right of the decimal place in the case-weight or pattern frequency field, if any.

f 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.

Notes

Columns skipped between fields are indicated by nX, where n is the number of columns to be passed over.

If the fields in the data records are not in the above order, the format tab designator (Tw) may be inserted in front of any of the fields (w is the position of the first column of the field, counting from column one). Check the INPUTdata carefully when left tabs are used.

A forward slash (/) means "skip to the next line". For example,

```
(5A1,5X,15A1/10X,15A1)
```

would read the case ID and 15 item responses from line 1; then, skip ten columns and read 15 item responses from line 2.

The variable format statement for aggregate-level data has the general form:

```
(aA1,Ib,Ic,Fd.e,f(Fg.h,Fi.j))
```

where:

- g is the number of columns in the "number tried" field
- h is the number of columns to the right of the decimal point in the "number tried" field
- i is the number of columns in the "number right" field.
- j is the number of columns to the right of the decimal point in the "number right" field.

Examples

The following example uses simulated responses to illustrate nonequivalent groups equating of two forms of a 25-item multiple-choice examination administered to different populations. The two forms have five items in common: C21, C22, C23, C24, and C25. The items for each group are specified in the GROUP1 and GROUP2 commands. Note that the item lists on the GROUP commands are the same as those on the FORMS command. This is because Group 1 took Form 1 of the examination and Group 2 took Form 2 of the examination.

As an answer key is provided in the raw data file (KFNAME=EXAMPL03.DAT on the INPUT command), the answer key appears first. Note that, when multiple forms are used, an answer key for each form should be provided. The answer key is in the same format as the data. For each examinee, two lines of data are provided. The first line contains identifying information and the second the item responses.

The first information read from the data file is the examinee's ID, which is in column 35 (5A1). For the first examinee the ID is 0001, and for the last 0200. Using the T operator to move to column 25, the form indicator is read next (I1). Because the values for form and group are the same for any given subject, a single form/group indicator appears on each data record. The indicator is read twice, first for forms and then for groups. The "/" operator is used to move to the first column of the second line. The 25 item responses are then read as (25A1).

```
>GLOBAL DFNAME='EXAMPL03.DAT', NPARM=2;
>LENGTH NITEMS=(45);
>INPUT  NTOTAL=45, SAMPLE=2000, NIDCHAR=5, NALT=5, NGROUP=2,
        NFORM=2, TYPE=1, KFNAME='EXAMPL03.DAT';
>ITEMS  INAMES=(C01(1)C45), INUMBERS=(1(1)45);
>TEST1  TNAME='CHEMISTR', INUMBERS=(1(1)45);
>FORM1  LENGTH=25, INUMBERS=(1(1)25);
>FORM2  LENGTH=25, INUMBERS=(21(1)45);
>GROUP1 GNAME='POP1', LENGTH=25, INUMBERS=(1(1)25);
>GROUP2 GNAME='POP2', LENGTH=25, INUMBERS=(21(1)45);
(T35,5A1,T25,I1,T25,I1/25A1)

ANSWER KEY  FORM 1             1
1111111111111111111111111
ANSWER KEY  FORM 2             2
1111111111111111111111111
Samp1     GROUP1        11     1  00001
1111111112221212211111121
Samp1     GROUP1        11     2  00002
2211111212222222222255222
Samp2     GROUP2        22     2  00199
2422221211222211222221121
…
Samp2     GROUP2        22   100  00200
1111111111111111212111111
```

The following 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.

As in the previous example, two lines of data are provided for each examinee. The first line contains identifying information and the second the item responses. The first information read from the data file is the examinee's ID, which is in column 35 (5A1). For the first examinee the ID is 0001, and for the last 0200. Using the “T” operator to move to column 25, the form indicator is read next (I1). The “/” operator is used to move to the first column of the second line. The 20 item responses per form are then read in.

```
>GLOBAL DFNAME='EXAMPL04.DAT', NIDCH=5, NPARM=2;
>LENGTH NITEMS=(40);
>INPUT NTOTAL=40, SAMPLE=2000, NALT=5, NIDCHAR=5, NFORM=2,
       KFNAME='EXAMPL04.DAT';
>ITEMS INAMES=(T01(1)T40), INUMBERS=(1(1)40);
>TEST1 TNAME='SIM', INMBERS=(1(1)40);
>FORM1 LENGTH=20, INUMBERS=(1(1)20);
>FORM2 LENGTH=20, INUMBERS=(21(1)40);
(T35,5A1,T25,I1/40A1)

ANSWER KEY  FORM 1            1
11111111111111111111
ANSWER KEY  FORM 2            2
11111111111111111111
Samp1    GROUP1        12     1  00001
11111111122212122111
Samp1    GROUP1        12     1  00002
11222212221222222112
…
```

Two hundred students at each of three grade levels, grades four, six, and eight, were given grade-appropriate versions of a 20-item arithmetic examination. Items 19 and 20 appear in the grade 4 and 6 forms; items 37 and 38 appear in the grade 6 and 8 forms. Because each item is assigned a unique column in the data records, a FORM command is not required. Both an answer and not-presented key are given at the top of the raw data file (KFNAME=EXAMPL05.DAT, NFNAME=EXAMPL05.DAT on the INPUT command). In the case of the answer key, a “1” represents a correct response. A not-presented item is indicated by a blank, ” “.

As no FORM command is required, only a group indicator has to be read in. The case ID, given in column 35, is read first (5A1), followed by the group indicator in column 25 (I1). The 56 item responses are read from the second line of data (56A1) after using the “/”operator to move to the start of this line.

```
>GLOBAL  DFNAME='EXAMPL05.DAT',NPARM=2;
>LENGTH  NITEMS=(56);
>INPUT   NTOTAL=56,SAMPLE=2000,NGROUPS=3,KFNAME='EXAMPL05.DAT',
         NFNAME='EXAMPL05.DAT',NIDCHAR=5;
>ITEMS   INUMBERS=(1(1)56),INAME=(M01(1)M56);
>TEST    TNAME=MATH;
>GROUP1  GNAME='GRADE 4',LENGTH=20,INUMBERS=(1(1)20);
>GROUP2  GNAME='GRADE 6',LENGTH=20,INUMBERS=(19(1)38);
>GROUP3  GNAME='GRADE 8',LENGTH=20,INUMBERS=(37(1)56);
(T35,5A1,T25,I5/56A1)

ANSWER KEY
11111111111111111111111111111111111111111111111111111111
NOT-PRESENTED KEY
```

```
Samp1     GROUP1        1      1  00001
11111112221211222212
Samp1     GROUP1        1      1  00002
21121211121111121212
Samp3     GROUP3        3      3  00199
                                    12212212212211112121
Samp3     GROUP3        3      3  00200
                                    11111111121111111111
```

The following example illustrates the use of the TYPE=3 specification on the INPUT command to analyze aggregate-level, multiple-matrix sampling data. The data in **exampl08.dat** are numbers tried and numbers correct for items from eight forms of a matrix sampled assessment instrument. The groups are selected 8-th grade students from 32 public schools. The first record for each school contains the data for the items of a Number Concepts scale, NUMCON, and the second record contains the data for items of an Algebra Concepts scale, ALGCON. An answer key is not relevant for aggregate-level data in number-tried, number-right summary form. Note the format statement for reading the two sets of eight number-tried, number-right observations from the two data lines. Again, the "/" operator is used to move to the start of the second line of data for each school.

```
>GLOBAL DFNAME='EXAMPL08.DAT', NPARM=3, NTEST=2;
>LENGTH NITEMS=(8, 8);
>INPUT NTOTAL=16, NALT=5, TYPE=3, NIDCHAR=9;
>ITEMS INAMES=(N1(1)N8, A1(1)A8), INUMBER=(1(1)16);
>TEST1 TNAME='NUMCON', INUMBERS=(1(1)8);
>TEST2 TNAME='ALGCON', INUMBERS=(9(1)16);
(1X,9A1,5X,8(2F3.0)/15X,8(2F3.0))
```

```
 SCHOOL  1 NUM   1  0  3  2  2  1  4  4  3  2  2  1  4  3  4  1
 SCHOOL  1 ALG   1  0  3  1  2  0  3  2  3  2  2  1  4  1  4  0
 SCHOOL  2 NUM   5  3  4  4  3  2  3  3  2  2  4  3  4  3  5  3
 SCHOOL  2 ALG   5  2  4  2  3  2  3  2  2  2  4  2  4  2  5  3
```

The next example illustrates the use of BILOG-MG with multiple groups and multiple subtests. Based on previous test performance, examinees are assigned to two groups for adaptive testing. Out of a set of 45 items, group 1 is assigned items 1 through 25, and group 2 is assigned items 21 through 45; thus, there are 5 items linking the test forms administered to the groups.

Twenty of the 25 items presented to group 1 belong to subtest 1 (items 1-15 and 21-25). Twenty items also belong to subtest 2 (items 6-25). Of the 25 items presented to group 2, 20 belong to subtest 1 (items 21-40) and 20 to subtest 2 (items 21-25 and 31-45).

In all, there are 35 items from the set of 45 assigned to each subtest. (This extent of item overlap between subtests is not realistic, but it illustrates that more than one subtest can be scored adaptively provided they each contain link items between the test forms.)

Note that, in this case, the item responses on the second line of data for each examinee represent responses to different items. When we previously considered these data, the response in the first column of the second line represented the response to item 1, regardless of group membership. Here, that response would be the response to item 1 for a member of group 1, but the response to item 21 for an examinee from group 2.

```
>GLOBAL  DFNAME='EXAMPL03.DAT',NPARM=2,NTEST=2,SAVE;
>SAVE    SCORE='EXAMPL09.SCO';
>LENGTH  NITEMS=(35,35);
>INPUT   NTOTAL=45, SAMPLE=2000, NGROUP=2, KFNAME='EXAMPL03.DAT', NALT=5,
         NFORMS=2,NIDCHAR=5;
>ITEMS   INUMBERS=(1(1)45), INAME=(C01(1)C45);
>TEST1   TNAME=SUBTEST1, INAME=(C01(1)C15,C21(1)C40);
>TEST2   TNAME=SUBTEST2, INAME=(C06(1)C25,C31(1)C45);
>FORM1   LENGTH=25,INUMBERS=(1(1)25);
>FORM2   LENGTH=25,INUMBERS=(21(1)45);
>GROUP1  GNAME=POP1,LENGTH=25,INUMBERS=(1(1)25);
>GROUP2  GNAME=POP2,LENGTH=25,INUMBERS=(21(1)45);
(T35,5A1,T25,I1,T25,I1/45A1)

ANSWER KEY  FORM 1             1
1111111111111111111111111
ANSWER KEY  FORM 2             2
1111111111111111111111111
Samp1     GROUP1        11     1  00001  1.0      .48900
1111111112221212211111121
Samp1     GROUP2        22     2  00002  1.0     -.92734
2211111212222222222255222
```

Default

None.

Related topics

- Data Structures (see Chapter 1.4)
- **Data** menu: **Examinee Data** dialog box (see Section 2.4)
- **Data** menu: **Group-Level Data** dialog box
- FORM command (see Section 3.2.4)
- GLOBAL command: DFNAME and NWGHT keywords (see Section 3.2.5)
- INPUT command: NFORM, TYPE and PERSONAL keywords (see Section 3.2.7)

3.2.18 Input and output files

Input files

The following data files contain problem information that must be supplied by the user as needed. Any text editor that writes an ASCII file may be used to prepare these files.

File	Keyword
Answer key	KFNAME on INPUT command
Not-presented key	NFNAME on INPUT command
Omit key	OFNAME on INPUT command
Original data file	DFNAME on GLOBAL command
Provisional starting values	PRNAME on GLOBAL command

Note:

The assignment of specific names to these files in the INPUT command causes the program to read external files.

These files may be combined into one file, using the order above. The arrangement is to construct an arbitrarily named file consisting of the answer key, if any, the not-presented key, if any, the omit key, if any and the item-response data. Any of the above files would then have the name of that file. Section 6.5 illustrates the combination of an answer key and not-presented key within the data file.

Format of the input records

- The keys and the data records must have the same fixed-column formats.
- The fields of the data records are read in the following order:
    - The respondent identification field (up to 30 columns of characters as specified by the NIDCHAR keyword on the INPUT command).
    - The form number (only if NFORMS>1).
    - The group number or numbers (integer) (only if specified by a value larger than 1 for the NGROUP keyword of the INPUT command).
    - A real-valued case weight for the respondent or frequency for a response pattern (only if specified by the NWGHT keyword of the GLOBAL command).
    - The individual item-response records or patterns.
- The type of entries in the item-response field is determined by the TYPE keyword of the INPUT command and by the presence or absence of the KFNAME keyword of the INPUT command:
- if KFNAME is not present, the item responses are scored 1 = correct and 0 = not correct.
- if KFNAME is present, the item responses are arbitrary single ASCII characters, the correct alternatives of which appear in the same columns of the answer key.
- In either of the above types of data, not-presented items may be coded by an arbitrary character defined in the corresponding column of the not-presented key. (See the NFNAME keyword of the INPUT command in Section 3.2.7.)
- Similarly, omitted items may be coded by another character defined in the corresponding column of the omit key. (See the OFNAME keyword of the INPUT command.)

- The path to and filename of any of these files may be longer than 80 characters. As the maximum length of any line in the command file is 80 characters, multiple lines should be used. It is important to continue up to and including the 80th column when specifying a long path and filename.

For example, suppose the data file **exampl06.dat** is in a folder named:

```
C:\PROGRAM FILES\ITEM RESPONSE THEORY\IRT_2002\MARCH20\BILOG-MG-VERSION1.2\EXAMPLES
```

The correct way to enter this information in the command file is to enclose the name and path in apostrophes, and continue until column 80 is reached. Then proceed in column 1 of the next line as shown below:

```
>GLOBAL DFNAME='C:\PROGRAM FILES\ITEM RESPONSE THEORY\IRT_2002\MARCH20\BILOG-MG
-VERSION1.2\EXAMPLES\EXAMPL06.DAT', NTEST=1, NVTEST=1, NPARM=2, SAVE;
```

If the data are stored in the same folder as the command file, it is sufficient to type

```
DFNAME='EXAMPL06.DAT'
```

Related topics

- **Data** menu: **Examinee Data** dialog box (see Section 2.4)
- **Data** menu: **Group-Level Data** dialog box
- **Data** menu: **Item Keys** dialog box
- GLOBAL command (see Section 3.2.5)
- INPUT command (see Section 3.2.7)

Output files

Through use of the keywords on the SAVE command, the following output files may be created.

- Ability score file
- Classical item statistics file
- DIF parameter file
- DRIFT parameter file
- Estimated covariance file
- Expected frequency file
- Item parameter file
- Marginal posterior probability file
- Test information file

Related topics

- SAVE command (see Section 3.2.13)
- **Save** menu (see Section 2.6)

Ability score file

Keyword: SCORE

This file is created during Phase 3 of the program if SCORE is specified in the SAVE command. It consists of the title records and two records per subtest for each respondent.

The format is as follows:

Records	Description
1 & 2	In 20A4/20A4 format, the title records of the BILOG-MG run that created the ability score file.
3+	Two records per subtest for each respondent, containing the following information

Columns	Format	Description
First record		
1 – 3	I3	group indicator
4 – 5	2X	blank filler
6 – 35	30A1	respondent identification
Second record		
1 – 6	F6.2	respondent case weight
7 – 7	A1	* if the subject is not calibrated; a blank otherwise
8 – 15	A8	subtest name
16 – 20	I5	number of attempts to items in the subtest
21 – 25	I5	number of correct responses to the subtest
26 – 35	F10.6	percent-correct score
36 – 47	F12.6	scale score estimate

48 – 59	F12.6	estimated standard error of scale score
60 – 60	A1	* if standard error of scale score is inestimable; a blank otherwise
61 – 70	F10.6	group fit probability, if requested
71 – 80	F10.6	marginal probability of response pattern if EAP scoring is chosen

Related topics

- SAVE command (see Section 3.2.13)
- **Save** menu (see Section 2.7)

Classical item statistics file

Keyword: ISTAT

This file contains all classical item statistics computed and printed by Phase 1 of the program. The following items are written to this external file in the same format as used in the result output from Phase 1, ***.ph1**:

- the title records in format (20A4/20A4)
- item facilities (percent correct)
- number of attempts and correct responses to each item
- item-subscore correlations

Related topics

- SAVE command (see Section 3.2.13)
- **Save** menu (see Section 2.7)

DIF parameter file

Keyword: DIF

Records	**Description**
1 & 2	In 20A4/20A4 format, the title records of the BILOG-MG run that created the DIF parameter file
3 +	Three sets of item records for each subtest. The first set contains the unadjusted item threshold parameters and s.e.s for each group.

	The second set contains adjusted threshold parameters and s.e.s for each group. The last set contains estimates of group differences in adjusted threshold parameters.

First set of item records:

Columns	**Format**	**Description**
First record		
1 – 8	`A8`	test name
8 – 10	`2X`	blank filler
11 - 18	`A8`	item name
19 – 20	`2X`	blank filler
21 - 220	`20(F10.5)`	unadjusted threshold parameters for groups
Second record		
1 – 20	`20X`	Blank filler
21 – 220	`20(F10.5)`	Estimated s.e. of unadjusted threshold parameters for groups

Second set of item records:

First record		
1 – 8	A8	test name
8 – 10	2X	blank filler
11 – 18	A8	item name
19 – 20	2X	blank filler
21 – 220	20(F10.5)	adjusted threshold parameters for groups
Second record		
1 – 20	20X	blank filler
21 – 220	20(F10.5)	s.e. of adjusted threshold parameter for group contrasts

Third set of item records:

First record		
1 – 8	A8	test name
8 – 10	2X	blank filler
11 – 18	A8	item name
19 – 20	2X	blank filler
21 – 210	20(F10.5)	group differences in threshold parameters for group contrasts
211 – 220	10X	blank filler
Second record		
1 – 20	20X	blank filler
21 – 210	20(F10.5)	s.e. of group differences for group contrasts
211 – 220	10X	blank filler

Related topics

- SAVE command (see Section 3.2.13)
- **Save** menu (see Section 2.7)

DRIFT parameter file

Keyword: DRIFT

This file is saved during Phase 2 if DRIFT is specified on the SAVE command. It consists of title records and two records for each item. The format is as follows:

Records	Description
1 & 2	In 20A4/20A4 format, the title records of the BILOG-MG run that created the DRIFT parameter file
3 +	Two records for each item, containing the following information

Columns	Format	Description
First record		
1 – 8	`A8`	item name
9 – 10	`2X`	blank filler
11 – 21	`F11.5`	Intercept
22 – 32	`F11.5`	linear coefficient
33 – 43	`F11.5`	quadratic coefficient
44 – 54	`F11.5`	cubic coefficient
55 – 65	`F11.5`	quartic coefficient
66 – 76	`F11.5`	quintic coefficient

Second record		
1 – 10	10X	blank filler
11 – 21	F11.5	estimated s.e. of intercept
22 – 32	F11.5	estimated s.e. of linear coefficient
33 – 43	F11.5	estimated s.e. of quadratic coefficient
44 – 54	F11.5	estimated s.e. of cubic coefficient
55 – 65	F11.5	estimated s.e. of quartic coefficient
66 – 76	F11.5	estimated s.e. of quintic coefficient

Related topics

- SAVE command (see Section 3.2.13)
- **Save** menu (see Section 2.7)

Estimated covariance file

Keyword: COVARIANCE

This file is created by Phase 2 of the program and passed to Phase 3, where item information indices are added if requested. It contains title records and the item parameter estimates at the conclusion of Phase 2 and the added item information indices at the conclusion of Phase 3. The format is as follows:

Records	Description
1 & 2	In 20A4/20A4 format, the title records of the BILOG-MG run that created the covariance file.
3+	Four records for each item, containing the following information:

Columns	Format	Description
First record		
1 – 8	`A8`	item name
9 – 16	`A8`	subtest name
17 – 21	`I5`	group indicator
22 – 33	`F12.6`	slope estimate
34 – 45	`F12.6`	threshold estimate
46 – 57	`F12.6`	lower asymptote estimate
58 – 69	`F12.6`	estimation error variance for slope
70 – 81	`F12.6`	estimation error covariance for slope and threshold
Second record		
1 – 17	`17X`	blank filler
18 – 29	`F12.6`	estimation error variance for threshold
30 – 41	`F12.6`	estimation error covariance for slope and asymptote
42 – 53	`F12.6`	estimation error covariance for threshold and asymptote
54 – 65	`F12.6`	estimation error variance for lower asymptote
66 – 81	`16X`	blank filler
Third record		population-independent indices
1 – 17	`17X`	blank filler
18 – 29	`F12.5`	value of maximum information
30 – 41	`F12.5`	estimated s.e. of value of maximum information
42 – 53	`F12.5`	point of maximum information
54 – 65	`F12.5`	estimated s.e. of point of maximum information

66 – 81	16X	blank filler
Fourth record		population-dependent indices
1 – 17	17X	blank filler
18 – 29	F12.5	value of maximum effectiveness (info*density)
30 – 41	F12.5	point of maximum effectiveness
42 – 53	F12.5	average information
54 – 65	F12.5	reliability index (s.d./(s.d. + 1/(ave.info)2))
66 – 81	16X	blank filler

Related topics

- SAVE command (see Section 3.2.13)
- **Save** menu (see Section 2.7)

Expected frequency file

Keyword: EXPECTED

This file is created by Phase 2 of the program. It contains expected sample sizes, expected number of correct responses, expected proportions of correct responses, standardized posterior residuals and model proportions of correct responses. These values are evaluated at each quadrature point and item. The format of each item and each of the quadrature points is as follows:

Records	Description
1 & 2	In 20A4/20A4 format, the title records of the BILOG-MG run that created the expected file
3 +	Seven records for each item, containing the following information

Column	Format	Description
First record		
1 – 8	`A8`	item name
9 – 10	`2X`	blank filler
11 – 15	`I5`	group indicator
16 – 17	`2X`	blank filler
18 – 27	`A10`	label “POINT”
28 – 82	`5(F10.5,1X)`	five values of quadrature points
Second record		
1 – 8	`A8`	item name
9 – 10	`2X`	blank filler
11 – 15	`I5`	group indicator
16 – 17	`2X`	blank filler
18 – 27	`A10`	label “WEIGHT”
28 – 82	`5(F10.5,1X)`	five values of quadrature weights
Third record		
1 – 8	`A8`	item name
9 – 10	`2X`	blank filler
11 – 15	`I5`	group indicator
16 – 17	`2X`	blank filler
18 – 27	`A10`	label “TRIED”
28 – 82	`5(F10.5,1X)`	five values of expected sample sizes

Fourth record		
1 – 8	A8	item name
9 – 10	2X	blank filler
11 – 15	I5	group indicator
16 – 17	2X	blank filler
18 – 27	A10	label “RIGHT”
28 – 82	5(F10.5,1X)	five values of expected number of correct responses
Fifth record		
1 – 8	A8	item name
9 – 10	2X	blank filler
11 – 15	I5	group indicator
16 – 17	2X	blank filler
18 – 27	A10	label “PROPORTION”
28 – 82	5(F10.5,1X)	five values of expected proportions of correct responses
Sixth record		
1 – 8	A8	item name
9 – 10	2X	blank filler
11 – 15	I5	group indicator
16 – 17	2X	blank filler
18 – 27	A10	label “s.e.”
28 – 82	5(F10.5,1X)	five values of standardized posterior residuals

Seventh record		
1 – 8	A8	item name
9 – 10	2X	blank filler
11 – 15	I5	group indicator
16 – 17	2X	blank filler
18 – 27	A10	label "MODEL PROP"
28 – 82	5(F10.5,1X)	five values of model proportions of correct responses

Remark:

If more than five quadrature points are used, each record is duplicated with the same format. If there is more than one group, the item information is presented for each group. Sets of records within an item are separated by single-dashed lines. Sets of records between items are separated by double-dashed lines.

Related topics

- SAVE command (see Section 3.2.13)
- **Save** menu (see Section 2.7)

Item parameter file

Keyword: PARM

This file is saved during Phase 2 of the program if PARM is specified in the SAVE command. The file contains the item parameter estimates and other information. The format is as follows:

Records	Description
1 & 2	In 20A4/20A4 format, the title records of the BILOG-MG run that created the item parameter file
3	In 2I4 format, the number of subtests and the total number of items appearing in this file
4	In 20I4 format, the numbers of items in the main and variant subtest on as many records as necessary.

5+	One record for each item in the main and variant subtests (if any), containing the following information

Columns	**Format**	**Description**
1 – 8	A8	item name
9 – 16	A8	subtest name
17–26	F10.5	intercept parameter
27 – 36	F10.5	intercept s. e.
37 – 46	F10.5	slope parameter
47 – 56	F10.5	slope s. e.
57 – 66	F10.5	threshold parameter
67 – 76	F10.5	threshold s. e.
77 – 86	F10.5	dispersion parameter (reciprocal of slope)
87 – 96	F10.5	dispersion s. e.
97 – 106	F10.5	lower asymptote parameter
107 – 116	F10.5	lower asymptote s.e.
117 – 126	F10.5	DRIFT parameter
127 – 136	F10.5	DRIFT s. e.
137 – 146	F10.5	unused columns
147 – 150	I4	location of item in INPUTstream
151	A1	answer key
152	I1	dummy values

Related topics

- SAVE command (see Section 3.2.13)
- **Save** menu (see Section 2.7)

Marginal posterior probability file

Keyword: POST

This file is created by Phase 2 of the program. It contains title records, the respondent's identification and group numbers, the case weight, and the marginal posterior probability of its response pattern. The format of each respondent's record is as follows:

Records	**Description**
1 & 2	In 20A4/20A4 format, the records of the BILOG-MG run that created the posterior file
3+	Two records for each response pattern, containing the following information

Columns	**Format**	**Description**
First record		
1 – 5	I5	group indicator
6 – 10	5X	blank filler
11 – 40	30A1	respondent's identification number
41 – 80	40X	blank filler
Second record		
1 – 8	A8	subtest name
9 – 10	2X	blank filler
11 – 20	F10.3	case weight
21 – 25	5X	blank filler

26 – 40	`F15.10`	marginal posterior probability of the response pattern
41 – 80	`40X`	blank filler

Related topics

- SAVE command (see Section 3.2.13)
- **Save** menu (see Section 2.7)

Test information file

Keyword: TSTAT

This file contains all summary item and test information computed and printed by Phase 3 of the program. The following items are written to this external file in the same format as used in the result output from Phase 3, ***.ph3**:

- the title records in format (20A4/20A4)
- correlations among subtest scale scores
- means and estimates of scale scores

The following items are written only if the appropriate INFO keyword on the SCORE command has been specified:

- test information and standard error curves
- table of item information indices, including the point and value of maximum information and the corresponding estimated standard errors for those indices.

Related topics

- SAVE command (see Section 3.2.13)
- **Save** menu (see Section 2.7)
- SCORE command (see Section 3.2.14)

4 IRT graphics

4.1 Introduction

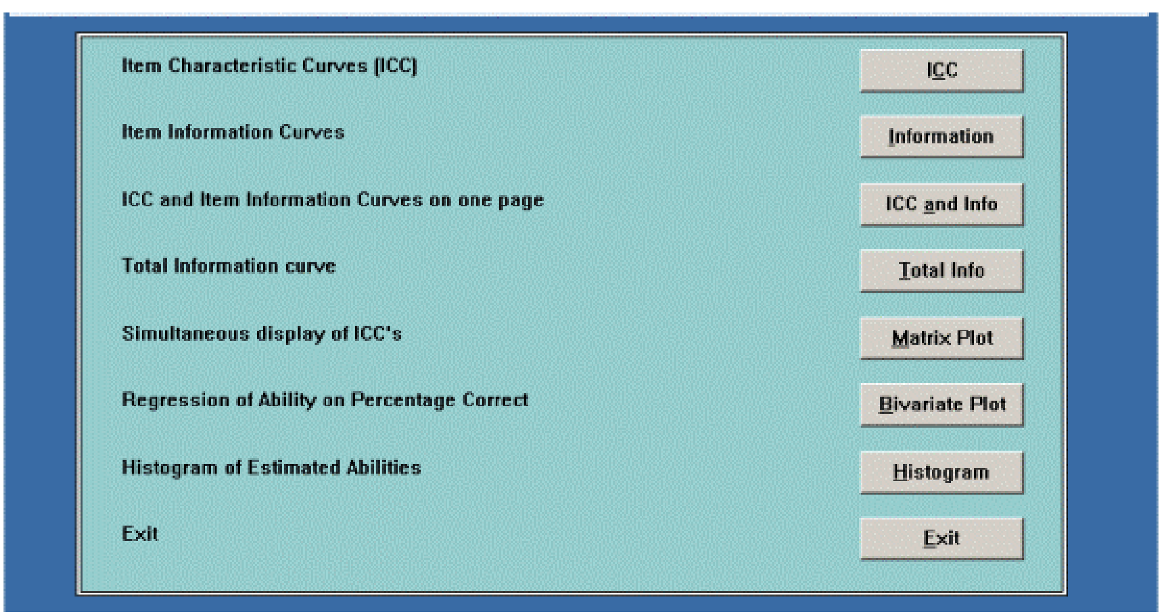
A new feature included with BILOG-MG is the IRT GRAPHICS procedure. Item characteristic curves, item and test information curves, and a histogram of the estimated abilities may be plotted. A matrix plot showing all item characteristic curves simultaneously can also be obtained. This feature is accessed via the **Run** menu on the main menu bar and becomes available once the analysis has been completed. The plots are based on the contents of the parameter files produced by the respective programs. In this chapter, an overview of the interface and options of this feature is given.

4.2 Main menu

The **Main** window of the IRT GRAPHICS program is used to access the following graphics:

- Item characteristic curves through the **ICC** option
- Item information curves through the **Information** option
- ICC and item information curves on the same page through the **ICC and Info** option
- Total information curve through the **Total Info** option
- Simultaneous display of all Item Characteristic Curves (ICCs) through the **Matrix Plot** option
- Regression of ability on the percentage correct through the **Bivariate Plot** option
- Histogram of estimated abilities through the **Histogram** option.

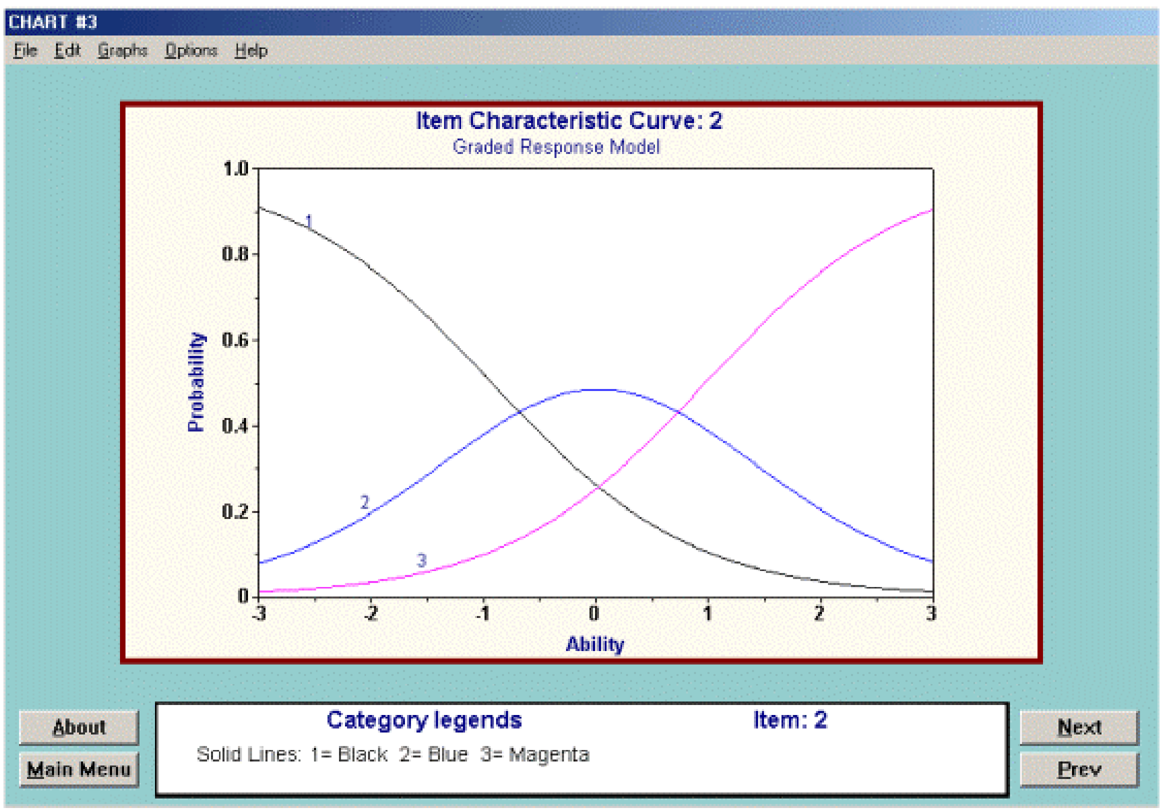
The graphs displayed may be selected, changed, saved to file, or printed using various options and dialog boxes described in Section 4.3. To exit the program, click the **Exit** option on the **Main** menu.



4.2.1 The ICC option

This option provides access to item characteristic curves for all the items in the test. In the image below, the ICC for item 2 is displayed.

- As a nominal model was fitted in this case, the high category is displayed in red and a message to this effect is displayed in the **Category Legends** box at the bottom of the window. This field contains the legend for all categories plotted.
- The **Next** button provides access to following items, while the **Prev** button allows the user to go back to previously viewed Item Characteristic Curves (ICC s).
- Use the **Main Menu** button at the bottom left of the window to return to the main menu.
- The graph can be selected, edited, saved, or printed using the **File**, **Edit**, **Graphs**, and **Options** menus on the main menu bar. For more on the options available, see Section 4.3.



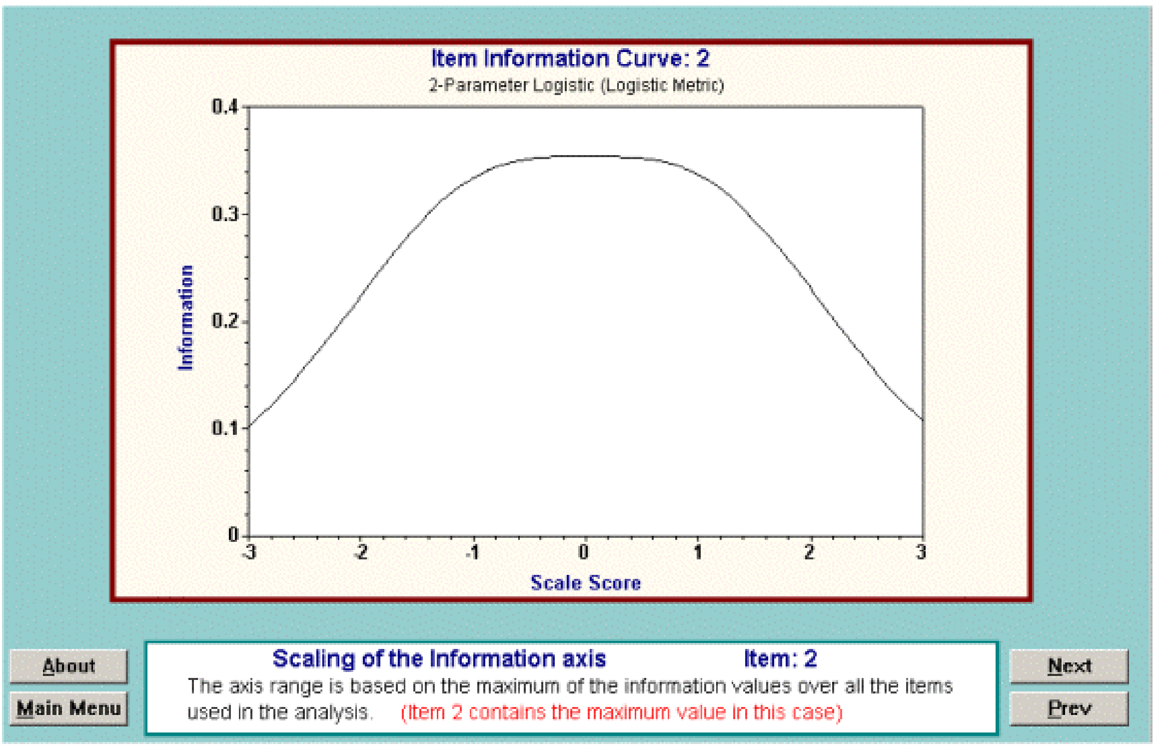
Related topics

- Manipulating and modifying graphs (see Section 4.3)
- Item characteristic curves (see Section 4.4)

4.2.2 The Information option

- This option provides access to item information curves for all the items in the test. In the image below, the item information curve for the second item is displayed.

- The **Scaling Information** box at the bottom of the window contains information on the scaling of the information axis. The item with the most information is indicated here for all items in a test.
- The **Next** button provides access to following items, while the **Prev** button allows the user to go back to previously viewed item information curves.
- Use the **Main Menu** button at the bottom left of the window to return to the main menu.
- The graph can be selected, edited, saved, or printed using the **File**, **Edit**, **Graphs**, and **Options** options on the main menu bar. For more on the options available, see Section 4.3.



Related topics

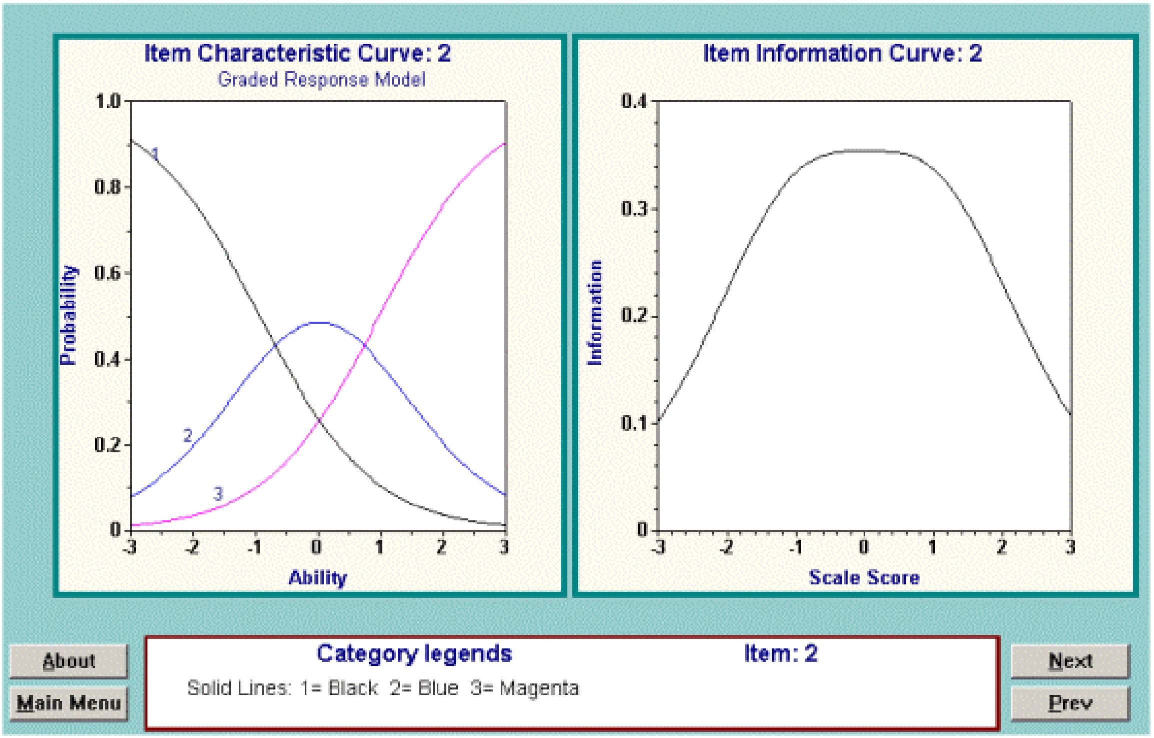
- Manipulating and modifying graphs (see Section 4.3)
- Item information curves (see Section 4.5)

4.2.3 The ICC and Info option

When this option is selected from the **Main** menu, the ICC and item information curve for an item are displayed simultaneously.

- As a nominal model was fitted in this case, the high category is displayed in red and a message to this effect is displayed in the **Category Legends** box at the bottom of the window. This field also contains information on the legend for all other categories plotted.
- The **Next** button provides access to following items, while the **Prev** button allows the user to go back to previously viewed item curves.
- Use the **Main Menu** button at the bottom left of the window to return to the main menu.

- The graph can be selected, edited, saved, or printed using the **File**, **Edit**, **Graphs**, and **Options** menus on the main menu bar. For more on the options available, see Section 4.3.



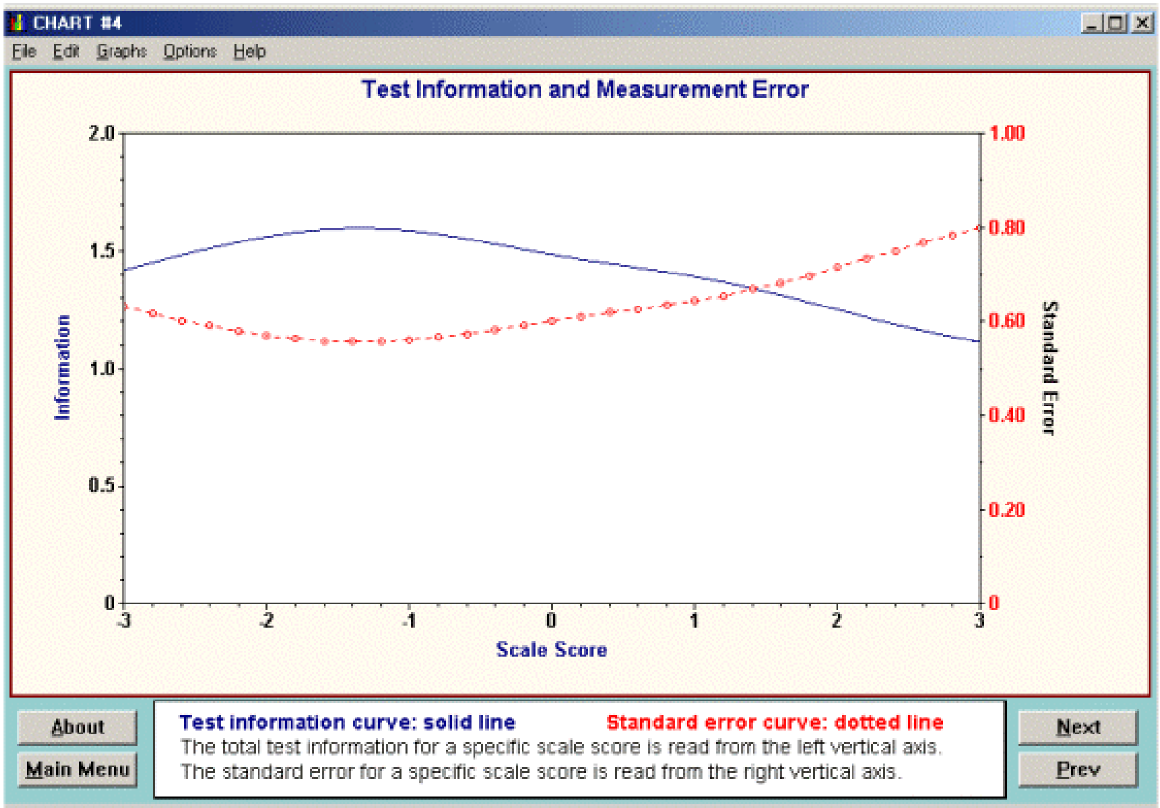
Related topics

- Manipulating and modifying graphs (see Section 4.3)
- Item characteristic curves (see Section 4.4)
- Item information curves (see Section 4.5)

4.2.4 The Total Info option

This option is used to access the test information and standard error curves.

- The total test information for a given scale score is read from the axis on the left of the graph and is plotted in blue.
- The axis to the right of the graph is used for reading the standard error estimate for a given scale score. The measurement error is shown in red.
- Use the **Main Menu** button at the bottom left of the window to return to the main menu.
- The **Next** and **Prev** buttons may be used to access similar plots for multiple groups (if any).
- The graph can be selected, edited, saved, or printed using the **File**, **Edit**, **Graphs**, and **Options** menus on the main menu bar. For more on the options available, see Section 4.3.



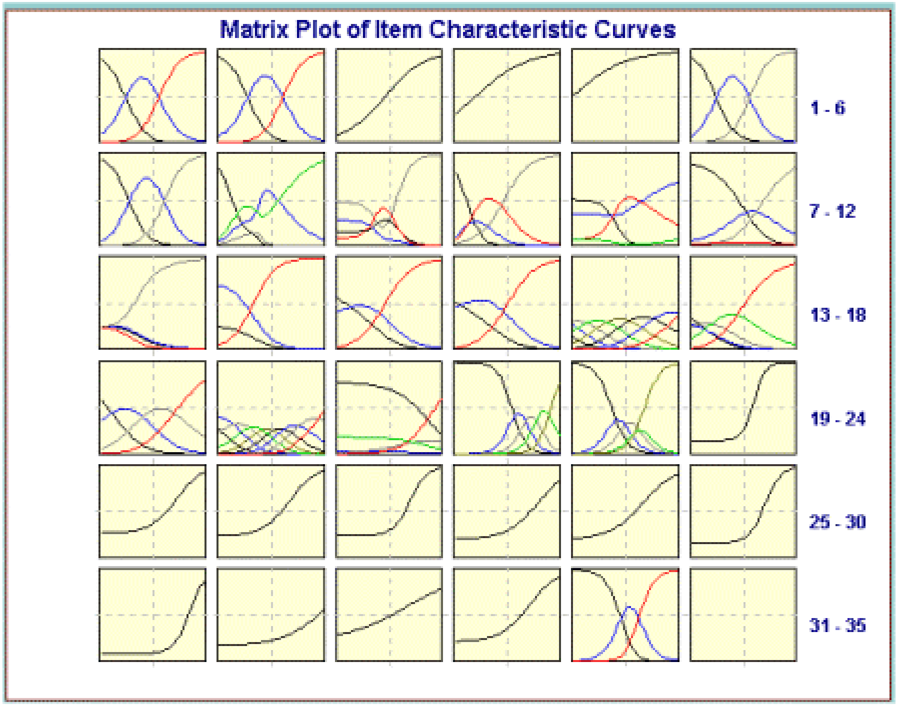
Related topics

- Manipulating and modifying graphs (see Section 4.3)
- Test information curves (see Section 4.6)

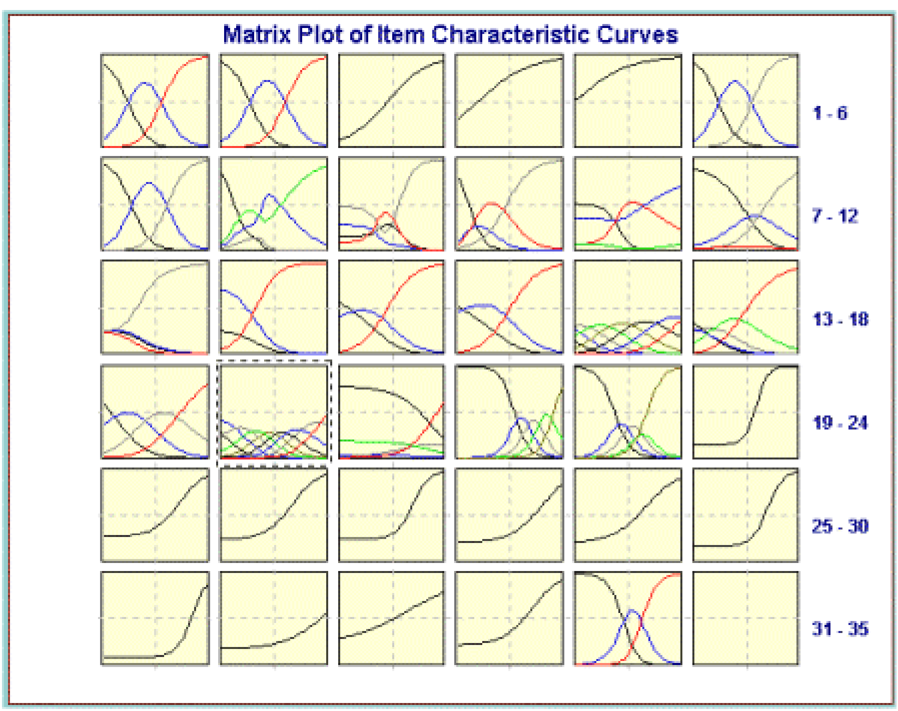
4.2.5 Matrix Plot option

This option provides an organized way of simultaneously looking at the item characteristic curves of up to 100 items.

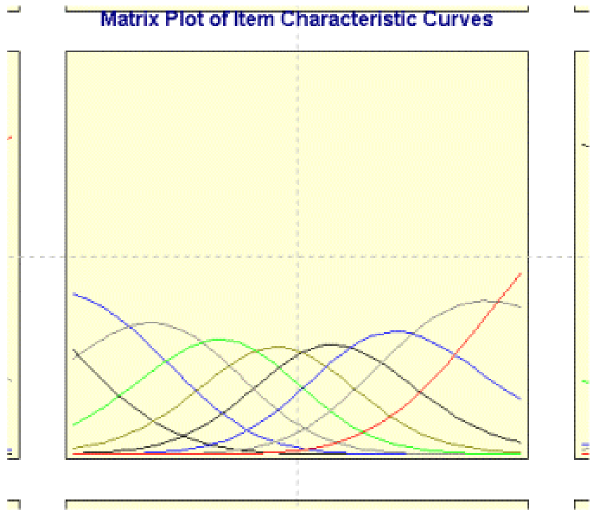
In the graph below, the ICCs of 35 items are plotted. As can be seen from the graph, models fitted to the items range from the 1PL model to the nominal, graded and multiple response models. Item 1 is shown in the top left corner of the combined graph, as indicated by the item numbers given to the right of the plots. The gray lines dividing each plot into four quadrants are drawn at a probability of 0.5 (on the y-axis) and ability of 0 (on the x-axis).



To take a closer look at item 20, to which a nominal response model was fitted, click and drag the right mouse button to select the area for zooming as shown below.



Releasing the mouse button produces a closer look at the graph for item 20 as shown below. Note that any part of the matrix of plots can be selected for zooming, and that the zoom option is also available for already enlarged areas of the matrix such as that shown below.



Note that the high category is shown in red. To reset the image, double-click the right mouse button.

- Up to 100 items can be simultaneously displayed. If the test contains more than 100 items, return to the **Main Menu** and click the **Matrix Plot** button again for the next set of items.
- The graphs can be selected, edited, saved, or printed using the **File**, **Edit**, **Graphs**, and **Options** menus on the main menu bar. For more on the options available, see Section 4.3.

Related topics

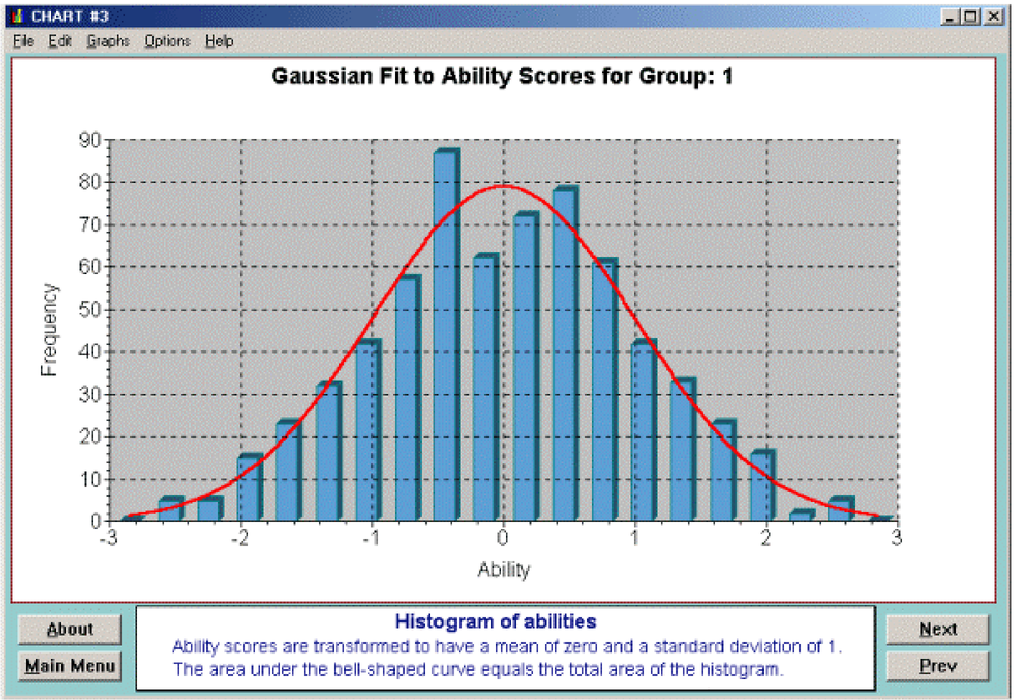
- Manipulating and modifying graphs (see Section 4.3)
- Item information curves (see Section 4.5)

4.2.6 The Histogram option

The **Histogram** option provides a histogram of the ability scores. This option is only available if scoring has been requested and the scores have been saved to an external file.

As indicated in the legend box at the bottom of the window, abilities are rescaled to a mean of 0 and standard deviation of 1. The area under the bell-shaped curve equals the total area of the histogram.

- Use the **Main Menu** button at the bottom left of the window to return to the main menu.
- The **Next** and **Prev** buttons may be used to access similar plots for multiple groups (if any).
- The graph can be selected, edited, saved, or printed using the **File**, **Edit**, **Graphs**, and **Options** options on the main menu bar.



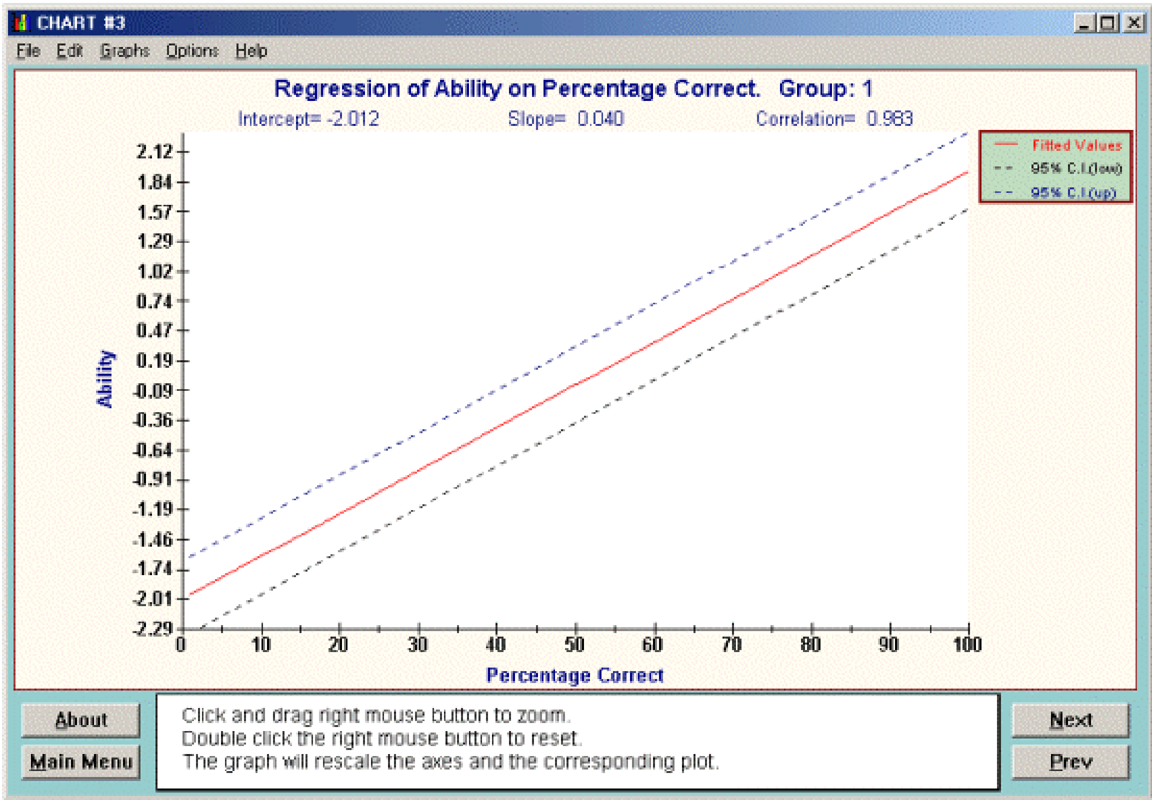
Related topics

- Manipulating and modifying graphs (see Section 4.3)

4.2.7 The Bivariate Plot option

The **Bivariate Plot** option provides a regression of ability on the percentage correct. This option is only available if scoring has been requested and the scores have been saved to an external file.

- As with the matrix plots, segments of the plot may be inspected by zooming in. This is done by clicking and dragging the mouse to select the area of interest.
- A 95% prediction interval for a new examinee is also shown on the plot.
- Use the **Main Menu** button at the bottom left of the window to return to the main menu.
- The graph can be selected, edited, saved, or printed using the **File**, **Edit**, **Graphs**, and **Options** menus on the main menu bar. For more on the options available, see Section 4.3.
- If information is available for multiple groups, bivariate plots are available by group and the **Next** and **Prev** buttons may be used to access the plots for following groups.



4.2.8 The Exit option

Use this option to exit the application.

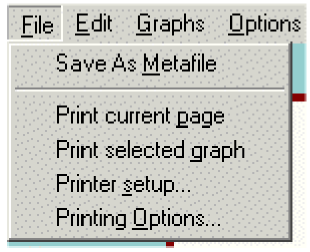
4.3 Manipulating and modifying graphs

Displayed graphs can be modified, saved and printed by using menus available on the main menu bar of the graph window.

4.3.1 File menu

The **File** menu controls the printing and saving of graphs.

- The **Save as Metafile** option is used to save the selected page or graph as a ***.wmf** (Windows Metafile) for later use in other applications.
- Note that an entire page, including legend boxes, may be printed using the **Print current page** option.
- Alternatively, the **Show Selectors** option on the **Options** menu may be used to select a graph, after which the **Print selected graph** option of the **File** menu may be used to print only the selected graph.
- The **Printer Setup** and **Printing Options** options provide access to the standard Windows printing controls.

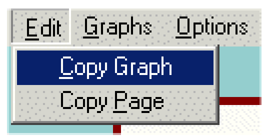


Related topics

- **Options** menu

4.3.2 Edit menu

The **Edit** menu is used for copying of graphs or entire pages to the Windows clipboard. To select a graph, the **Show Selectors** option on the **Options** menu may be used.



Related topics

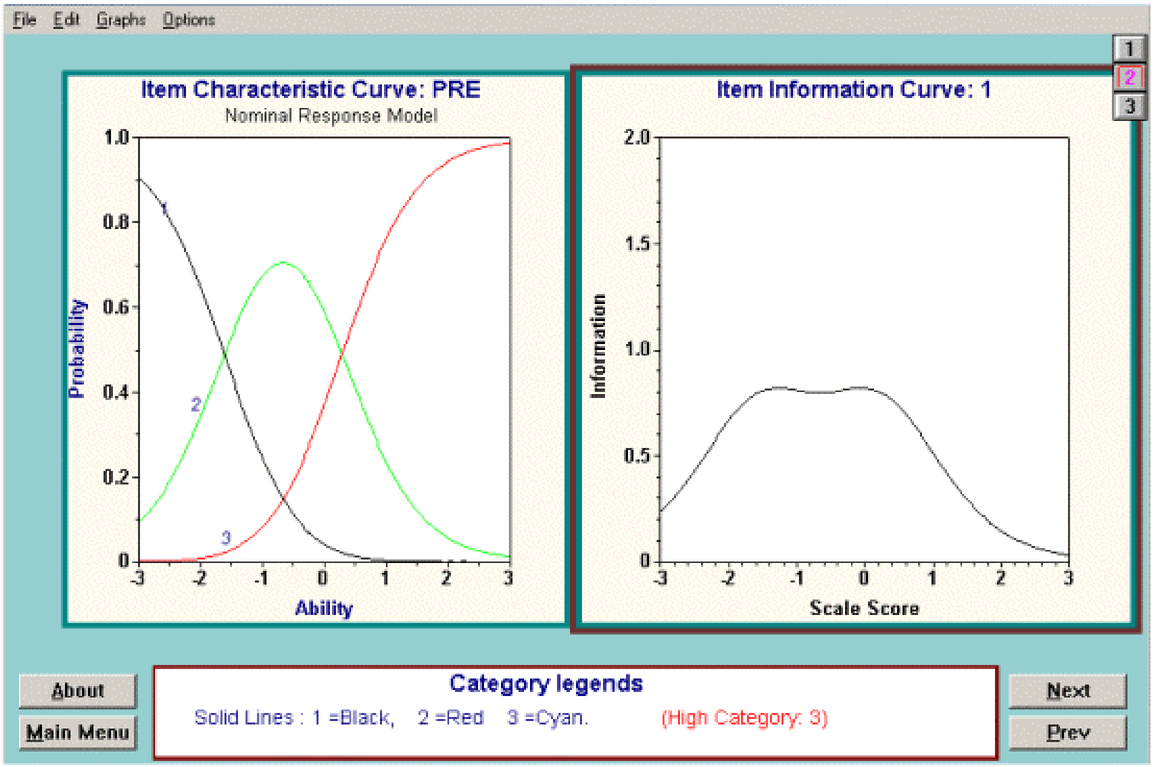
- **Options** menu

4.3.3 Options menu

The **Options** menu is used to enable graph selectors and to highlight a selected graph.

In the image below, both options have been enabled

and the selectors for the three areas of the graph below (the ICC, the item information curve and the **Category legends** box) are displayed at the right of the window. The second graph has been selected, and this entire section of the window is highlighted in dark red. This selected graph may now be saved or printed using options on the **File** menu.



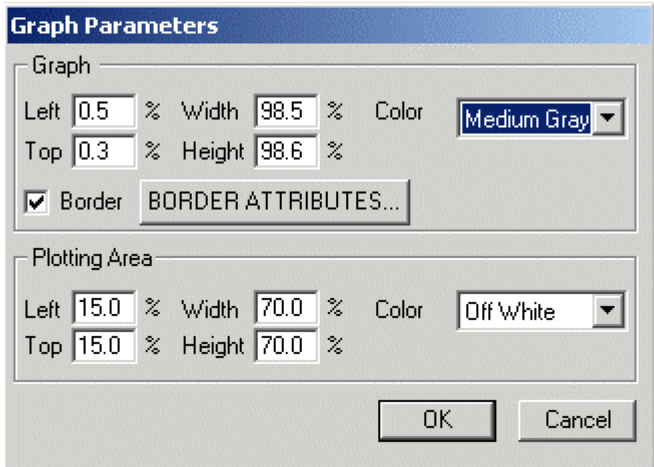
4.3.4 Graphs menu

The **Graphs** menu provides access to the **Parameters** and **Fill Page** options.

The **Fill Page** option is used to resize the graph to fill the entire window. The **Parameters** option is used to change attributes of the graph displayed and is associated with the **Graph Parameters** dialog box. This dialog box is used to change the position, size, and color of the currently selected graph and its plotting area.

The following functions are defined:

- The **Left**, **Top**, **Width,** and **Height** edit controls allow the user to specify a new position and size of the graph (relative to the page window) and of the plotting area (relative to the graph window).
- The **Color** drop-down list boxes are used to specify the graph window color and the color of the graph's plotting area.
- If the **Border** check box is checked, the graph will have a border around it.
- If the **Border** check box is checked, the **Border Attributes** button leads to another standard dialog box (the **Line Parameters** dialog box) that allows specification of the thickness, color, and style of the borderline.



In addition to the **Graphs Parameters** dialog box, a number of other dialog boxes may be used to change attributes of graphs. The dialog boxes accessible depend on the type of graph displayed. The dialog boxes are:

- **Axis Labels** dialog box
- **Text Parameters** dialog box
- **Bar Graph Parameters** dialog box
- **Legend Parameters** dialog box
- **Line Parameters** dialog box

The user may access any of these dialog boxes by double-clicking in the corresponding section of the graph. For example, double-clicking in the legend area of the graph will activate the **Legend Parameters** dialog box. Double-clicking on the title of the graph, on the other hand, will provide access to the **Text Parameters** dialog box.

4.3.5 Axis Labels dialog box

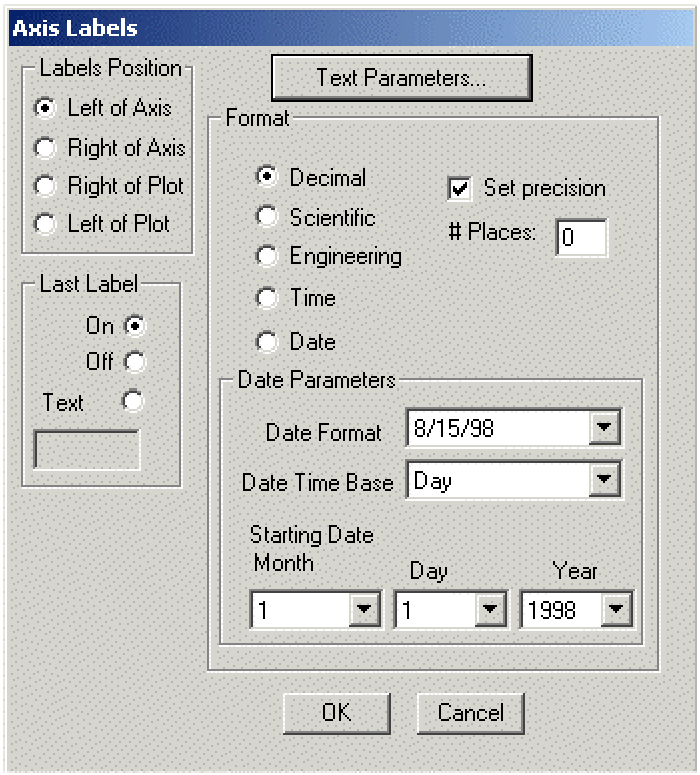
This dialog box is used for editing axis labels and is activated by double clicking on the axis of a displayed graph.

The following functions are defined:

- The **Labels Position** group box controls the position of the labels relative to the axis or plotting area.
- The **Last Label** group box allows manipulation of the last label drawing options. If **On** is selected, the last label is displayed like the others. If **Off** is selected, it is not displayed, If **Text** is selected, the text string entered in the edit box below will be displayed instead of the last numerical label.
- The format of the numerical labels can be specified using the radio buttons in the **Format** group box.
- The **Date Parameters** group box becomes active once the **Date** radio button is checked. The **Date Format** box selects the date format to use for labels, while the **Date Time Base**

box selects the time base (minute, hour, day, week, month, year) for the date calculations. The **Starting Date** drop-down list boxes specify the starting date that corresponds to the axis value of 0. All dates are calculated relative to this value.

- If the **Set Precision** check box is not checked, the labels' precision is determined automatically. If it is checked, the number entered into the **#Places** field specifies the number of digits after the decimal point.
- The **Text Parameters** button provides access to the **Text Parameters** dialog box (see Section 4.3.10) that controls the font, size, and color of labels.

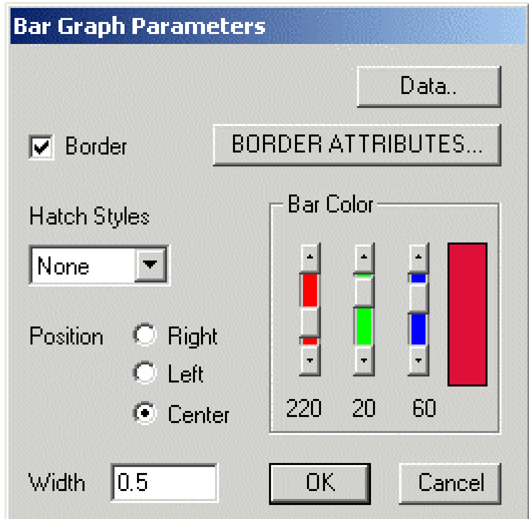


Related topics

- **Text Parameters** dialog box

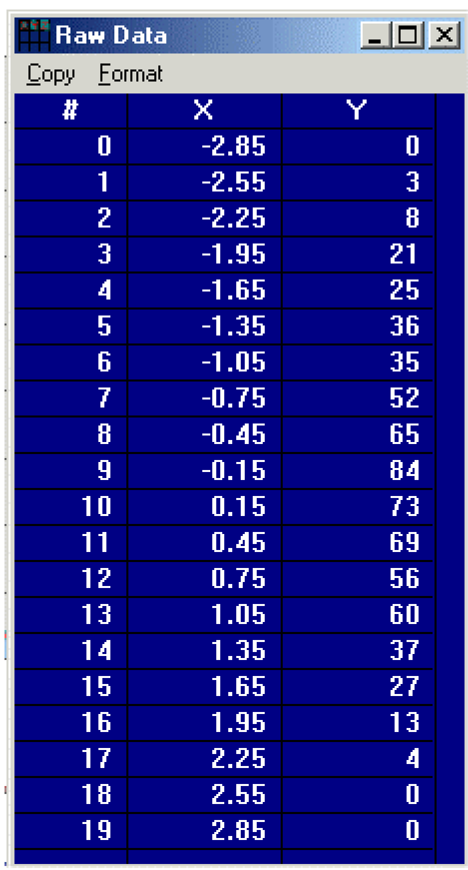
4.3.6 Bar Graph Parameters dialog box

This dialog box is used for editing the parameters of all bars in a regular bar graph, or a selected group member of grouped bar graphs. It is displayed when a bar in the histogram (**Histogram** option on the **Main** menu) is double-clicked.



It operates as follows:

- If the **Border** check box is checked, the bars have a border around them. In this case, the **Border Attributes** button leads to the **Line Parameters** dialog box that controls border thickness, color, and style.
- The **Data** button leads to the spreadsheet-style window for editing plotted data points (shown below)

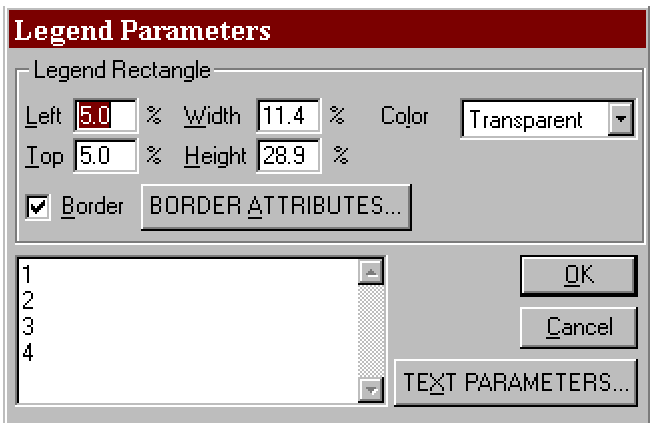


#	X	Y
0	-2.85	0
1	-2.55	3
2	-2.25	8
3	-1.95	21
4	-1.65	25
5	-1.35	36
6	-1.05	35
7	-0.75	52
8	-0.45	65
9	-0.15	84
10	0.15	73
11	0.45	69
12	0.75	56
13	1.05	60
14	1.35	37
15	1.65	27
16	1.95	13
17	2.25	4
18	2.55	0
19	2.85	0

- The **Hatch Style** drop-down list box allows the user to choose the hatch style for bars.
- The **Bar Color** scrolling bars control the bar RGB color.
- The **Position** radio buttons control the bar position relative to the independent variable values.
- The **Width** string field allows the user to enter the bar width in units of the independent variable.

4.3.7 Legend Parameters dialog box

This dialog box allows the editing of legends. It opens when the mouse button is double-clicked while the cursor is anywhere inside the legend box, except over a symbol representing a plotting object.



This dialog box operates as follows:

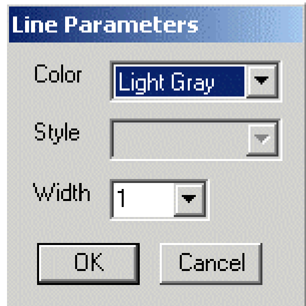
- The **Left**, **Top**, **Width,** and **Height** edit controls allow the user to specify a new position and size of the legend-bounding rectangle relative to the graph window.
- The **Color** drop-down menu specifies the legend rectangle background color.
- If the **Border** check box is checked, the rectangle will have a border. In this case, the **Border Attributes** button leads to the **Line Parameters** dialog box that controls border thickness, color, and style of the border line.
- The multi-line text box in the lower left corner lists and allows editing of each of the legend text strings.
- The **Text Parameters** button leads to the **Text Parameters** dialog box discussed earlier.

Related topics

- **Text Parameters** dialog box

4.3.8 Line Parameters dialog box

This dialog box is used for editing lines in the graph. It is accessed via the **Plot Parameters** dialog box, which is activated when a curve in a graph is double-clicked.



It has the following functions:

- The **Color** drop-down list box controls the line color.
- The **Style** drop-down list box, visible when activated, allows selection of a line style.
- The **Width** control specifies the line width, in window pixels.

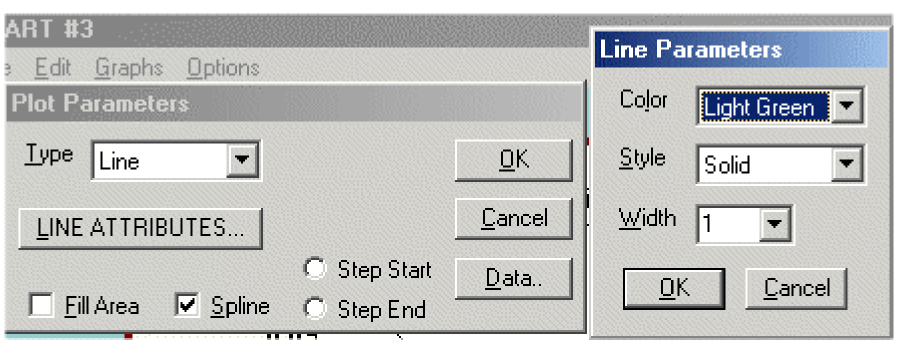
Related topics

- **Plot Parameters** dialog box

4.3.9 Plot Parameters dialog box

The **Plot Parameters** dialog box is accessed when a curve is double-clicked.

- The type of line to be displayed may be changed using the **Type** drop-down lost box.
- To fill the area under the curve, the **Fill Area** check box may be used.
- The type of curve fitted (spline or not) is controlled by the **Spline** check box.
- The **Data** button provides direct access to the data used to plot the curve.
- The **Line Attributes** button provides access to the **Line Parameters** dialog box (shown to the right of the **Plot Parameters** dialog box below). The **Line Parameters** dialog box is discussed elsewhere in this section.

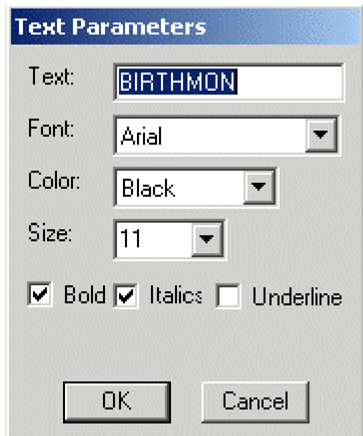


Related topics

- **Line Parameters** dialog box

4.3.10 Text Parameters dialog box

This dialog box is used for editing text strings, labels, titles, etc. It can be called from some of the other dialog boxes controlling graphic features. It may be activated by double clicking on any text in a displayed graph.



The following functions are defined:

- The **Text** edit control allows the user to edit the text string.
- The **Font** drop-down list box allows control of the typeface.
- The text color can be selected from the **Color** drop-down menu.
- The size of the fonts (in points) is controlled by the **Size** drop-down menu.
- The **Bold**, **Italic** and **Underline** check boxes control the text style.

4.4 Item characteristic curves

The item characteristic curve is a nonlinear function that portrays the regression of the item score on the trait or ability measured in a test. It shows the relationship between the probability of success on the item to the ability measured by the item set or test containing the item.

In the case of binary data, a single curve is used to portray this relationship, and the difficulty, discrimination and guessing parameters (where applicable) are indicated on the graph. In polytomous models such as the graded response model and nominal response model, a number of *item option* curves are plotted. Each curve shows the selection probability of a category of the item as a function of the ability.

For a description of the models for which item characteristic curves or item option curves may be obtained, see

Binary data:

- The one-parameter (1PL, Rasch) model (see Section 5.2)
- The two-parameter (2PL, Birnbaum) model

- The three-parameter (3PL, guessing) model

4.5 Item information curves

Item information functions are dependent on ability and provide valuable insight into the differences in the precision of measurement different ability levels. They are of particular interest in test construction, where these curves can be used to ensure the inclusion of different items that maximize the precision of measurement at different levels of θ in the test.

In the case of a 1PL model, the item information function is given by (Hambleton & Swaminathan, 1985, Table 6-1)

$$D^2\{1+\exp[-D(\theta-b_i)]\}^{-1}\{1-(1+\exp[-D(\theta-b_i)])^{-1}\}$$

The maximum value of the information is constant for the one-parameter model and is at the point b_i.

For a 2PL model, the item information function is given by (Hambleton & Swaminathan, 1985, Table 6-1)

$$D^2a_i^2\{1+\exp[-D(\theta-b_i)]\}^{-1}\{1-(1+\exp[-D(\theta-b_i)])^{-1}\}$$

with the maximum value directly proportional to the square of the item discrimination parameter, a. A larger value of a is associated with greater information. The maximum information is obtained at b_i.

For the three-parameter model, the information function is (Hambleton & Swaminathan, 1985, Table 6-1)

$$\frac{D^2a_i^2\{1+\exp[-D(\theta-b_i)]\}^{-1}\{1-(1+\exp[-D(\theta-b_i)])^{-1}\}}{\{1-c_i\}^2}$$

The maximum information is reached at

$$b_i+\frac{1}{Da_i}\ln\{1/2+1/2\sqrt{1+8c_i}\}$$

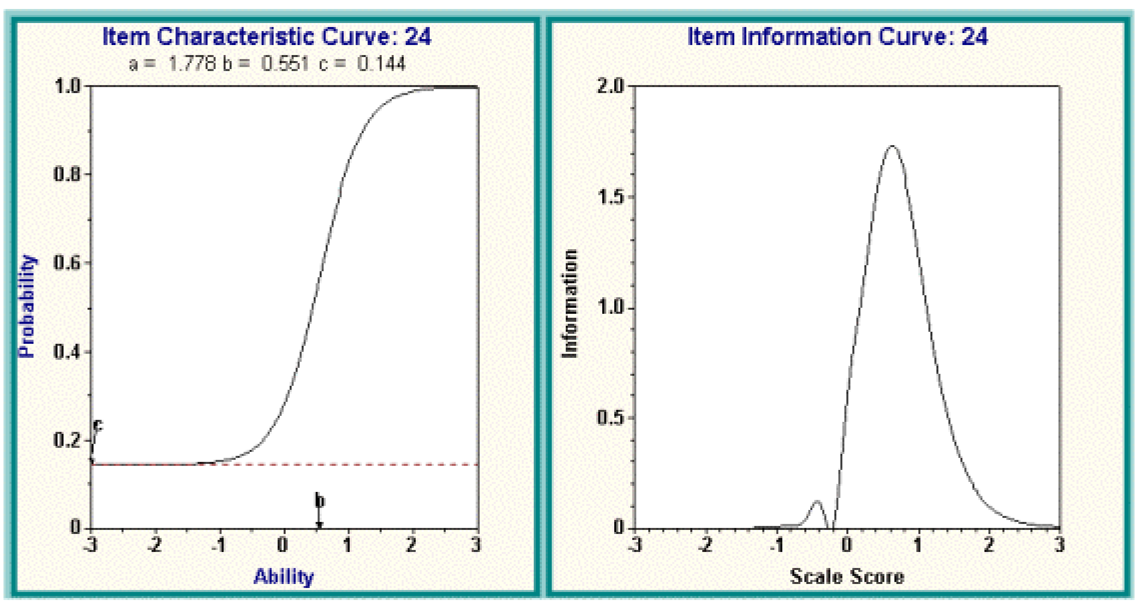
An increase in information is associated with a decrease in c_i. The maximum information is obtained when c_i = 0.

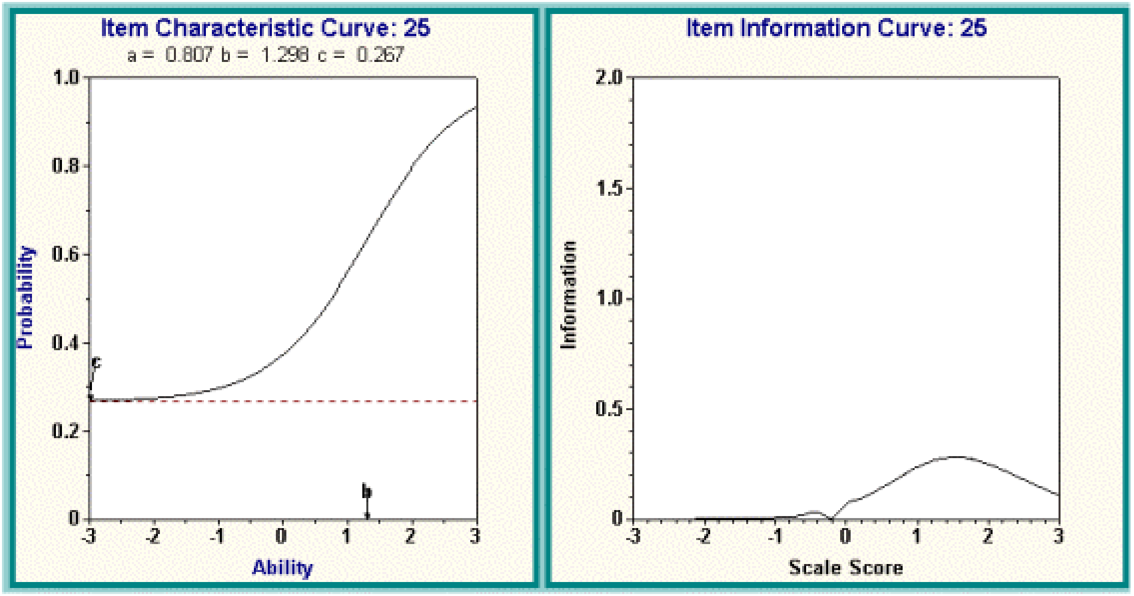
The slope of the item response function and the conditional variance at each ability level θ play an important role in terms of the information provided by an item. An increase in the slope, together with a decrease in the variance, leads to more information being obtained. This in turn provides a smaller standard error of measurement. By assessing these curves, items with large standard errors of measurement may be identified and discarded.

The contributions of both item and test information curves are summarized by Hambleton & Swaminathan (1985) as follows:

> “The item and test information functions provide viable alternatives to the classical concepts of reliability and standard error. The information functions are defined independently of any specific group of examinees and, moreover, represent the standard error of measurement at any chosen ability level. Thus, the precision of measurement can be determined at any level of ability that is of interest. Furthermore, through the information function, the test constructor can precisely assess the contribution of each item to the precision of the total test and hence choose items in a manner that is not contradictory with other aspects of test construction. ”

The item and item information curves for two items to which a 3PL model has been fitted are shown below. The discrimination parameter for item 24 is approximately twice that of item 25, and the effect of this can be seen in the corresponding item information curves. Both item information functions were plotted on the same scale. The item in the test with the most information determines the scale.





Related topics

- The **Information** option (see Section 4.2.2)
- The **ICC and Info** option (See Section 4.2.3)

4.6 Test information curves

The test information function summarizes the information function for a set of items or test. The contribution of each item in the test to the total information is additive, as can be seen from the definition of the test information function

$$I(\theta) = \sum_{i=1}^{n} \frac{P_i^{'}(\theta)^2}{P_i(\theta)Q_i(\theta)}$$

where $P_i(\theta)$ denotes the probability of an examinee responding correctly to item *i* given an ability of θ, and $Q_i(\theta) = 1 - P_i(\theta)$.

The function provides information for a set of items at each point on the ability scale and the amount of information is influenced by the *quality* and *number* of test items. As was the case for the item information function, the item slope and item variance play an important role. An increase in the slope and a decrease in the item variance both lead to more information being obtained. This in turn provides a smaller standard error of measurement. Also note that the contribution of each test item is independent of the other items in the test.

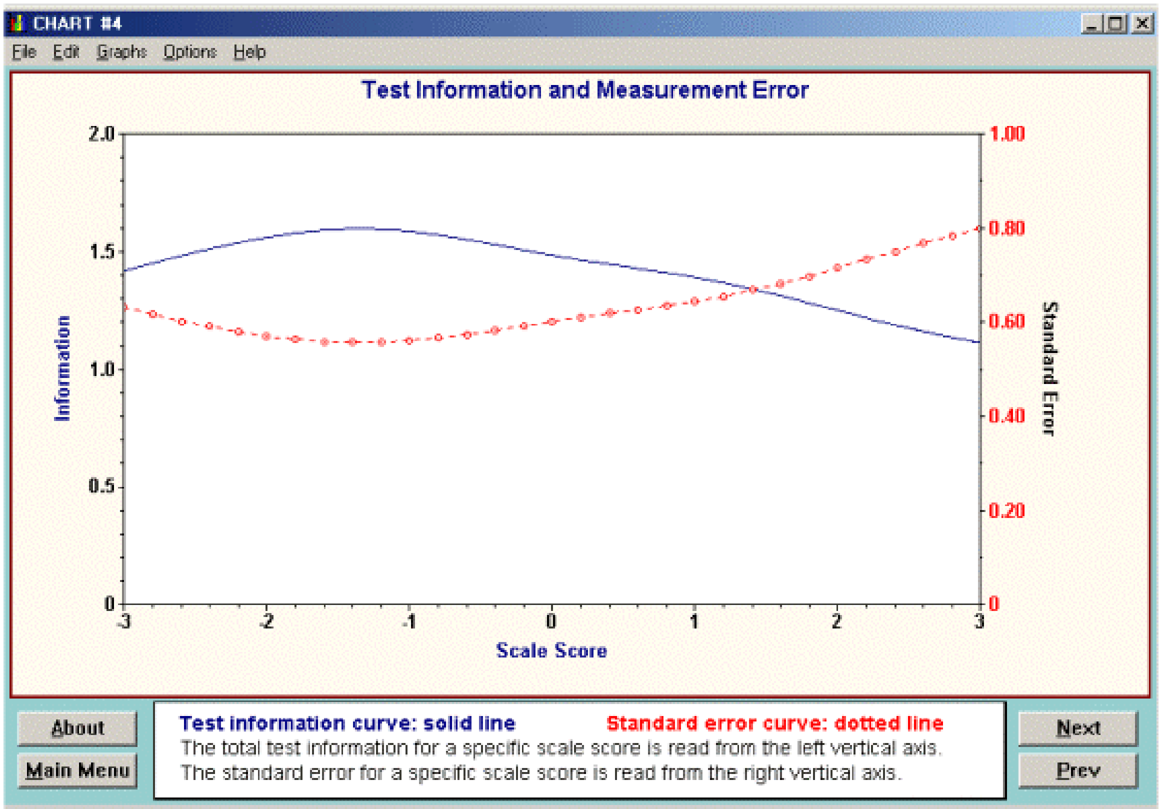
The amount of information provided by a set of test items at an ability level is inversely related to the error associated with ability estimates at the ability level. The standard error of the ability estimates at ability level θ can be written as

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}}.$$

An example of test information and measurement error curves is shown below. Note that the vertical axis to the left is used to find the information at a given ability while the vertical axis to the right serves a similar purpose for the standard error curve.

Related topics

- The **Total Info** option (see Section 4.2.4)



5 Estimation in BILOG-MG

5.1 Introduction

The central concept of item response theory is that of the item response model. These models are mathematical expressions describing the probability of a correct response to a test item as a function of the ability (or proficiency) of the respondent. For binary data, the response functions most often encountered in IRT applications are the normal ogive and the logistic models. These are discussed in Section 1.2.10. Multiple-group applications are considered in Section 1.2.

5.2 Item calibration

To make use of IRT for test scoring, the parameters of the model for each item of the test must be estimated. Estimating the item parameter and checking the fit of the models is referred to as item *calibration*. The calibration process requires item data from a sample of respondent who have been administered the test under exactly the same conditions as those in which the test will be used in practice. After a preliminary pilot study to select suitable items, the main item calibration can be performed on data obtained in the first operational use of the test. Replacement of items in subsequent administrations can then be carried out by one of the methods described in the section on test equating.

5.3 Marginal Maximum Likelihood estimation (MML)

An approach to item parameter estimation that applies to all types of item response models and is efficient for short and long tests is the method of marginal maximum likelihood (MML). (See Bock & Aitkin, 1981; Harwell, Baker & Zwarts, 1988). Except in special cases, the MML method assumes the *conditional independence* of responses to different items by persons of the same ability θ. Because the joint probability of independent events is the product of the probabilities of the separate events, this assumption makes it possible to calculate the probability of observing a particular pattern of item scores,

$$\mathrm{x} = (x_1, x_2, \ldots, x_n),$$

in the responses of a person with ability θ.

This probability may be expressed as

$$P(\mathrm{x} \mid \theta) = \prod_{j=1}^{n} [P_j(\theta)]^{x_j} [1 - P_j(\theta)]^{1-x_j} \qquad (5.1)$$

that is, as the continued product of $P_j(\theta)$ or $1 - P_j(\theta)$ according as the person response correctly or incorrectly to item j. This quantity is the probability of the pattern x, *conditional* on θ. It is to be distinguished from the probability of observing the pattern x from a person of unknown ability drawn at random from a population in which θ is distributed with a continuous density $g(\theta)$. The latter is the *unconditional* probability given by the definite integral,

$$P(\mathrm{x}) = \int_{-\infty}^{\infty} P(\mathrm{x} \mid \theta) g(\theta)\, d\theta. \qquad (5.2)$$

This quantity is also referred to as the *marginal* probability of x. Because the ability, θ, has been integrated out, this quantity is a function of the item parameters only.

In IRT applications, the integral in (5.2) cannot generally be expressed in closed form, but the marginal probability can be evaluated as accurately as required by the Gaussian quadrature formula

$$\bar{P}_x \approx \sum_{k=1}^{q} P(\mathrm{x} \mid X_k) A(X_k), \qquad (5.3)$$

where X is a quadrature point and $A(X_k)$ is a positive weight corresponding to the density function, $g(X)$. Tables giving quadrature points and corresponding weights are available for various choices of $g(\theta)$ (see Stroud & Sechrest, 1966). We recommend 2 x the square root of the number of items as the *maximum* number of quadrature points.

In the MML method, values for the item parameters are chosen so as to maximize the logarithm of the marginal maximum likelihood function, defined as

$$\log L_M = \sum_{l=1}^{S} r_l \log_e \bar{P}(x_l), \qquad (5.4)$$

where r_l is the frequency with which the pattern x_l is observed in a sample of N respondents, and S is the number of distinct patterns observed.

A necessary condition on the maximum of (5.4) for the 3PL model of item j is given by the likelihood equations

$$\sum_{k=1}^{q} \left(\frac{\bar{r}_{jk} - \overline{N}_{jk} P_j(X_k)}{P_j(X_k)[1 - P_j(X_k)]} \right) \frac{\partial P_j(X_k)}{\partial \begin{bmatrix} c_j \\ a_j \\ g_j \end{bmatrix}} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \qquad (5.5)$$

where

$$\bar{r}_{jk} = \sum_{l}^{S} r_l x_{lj} P(\mathrm{x}_l \mid X_k) A(X_k) / \bar{P}_{\mathrm{x}l}$$

and (5.6)

$$\bar{N}_k = \sum_{l}^{S} r_l P(\mathrm{x}_l \mid X_k) A(X_k) / \bar{P}_{\mathrm{x}l}$$

are, respectively, the posterior expectation of the number-correct and of the number of attempts at point X_k. (x_{lj} is the 0,1 score for item *j* in pattern *l*).

The so-called EM algorithm and Newton-Gauss (Fisher-scoring) methods are used to solve these implicit equations. Details may be found in Bock and Aitkin (1981) and Thissen (1982). Standard errors and correlations of the parameter estimators are obtained by inverting the information matrix in the Fisher-scoring solution.

Marginal Maximum A Posteriori estimation (MMAP)

MML estimation for the two- and three-parameter models is essentially a one-dimensional item factor analysis. As such, it is subject to so-called Heywood cases in which a unique variance goes to zero. The symptom of such a case is an indefinitely increasing slope during the EM and Newton iterations of the maximization.

Because all items are fallible to some degree, zero unique variance is untenable. It is therefore reasonable to avoid Heywood cases by placing a stochastic constraint on the item slopes to prevent them from becoming indefinitely large. This may be done by adopting a Bayes procedure called "marginal maximum a posteriori" (MMAP) estimation. In one form of this procedure, the slopes (which must be positive) are assumed to have a log normal distribution in the domain from which the items are drawn. Values for the item parameters are then chosen so as to maximize the logarithm of the product of the likelihood of the sample and the assumed log normal "prior" distribution of the slopes. The parameters of this log normal distribution for slopes can be either specified as *a priori*—the Bayes solution—or estimated from the data at hand—an empirical Bayes solution. This amounts to finding the maximum of the posterior distribution of the slopes, given the data.

For the three-parameter model, a similar procedure is needed to keep the lower asymptote parameter, g_j, in the open interval from 0 to 1. The beta distribution may be used for this purpose. The intercept parameter can also be constrained to a plausible region, although this is less important than constraining the slope and asymptote. (See Mislevy, 1986, and Tsutakawa & Lin, 1986, for details).

Estimation of the latent distribution

It is possible to estimate the *distribution of* θ by MML. To do so it is necessary to solve the indeterminacy of location and scale that is inherent in the item response models. This indeterminacy arises because, in the logit,

$$z_j = a_j(\theta - b_j),$$

any change in the origin of θ can be absorbed in b_j, and any change in the unit of θ can be absorbed in a_j. A widely accepted convention is to fix location by setting the mean of the latent distribution (of θ) to 0 and to fix scale by setting the standard deviation of the distribution to 1. The parameters are then said to be in the "0, 1" metric. To set the mean and standard deviation to some other values, *m* and *s*, say, it is only necessary to change b_j to

$$b_j^* = sb_j + m \qquad (5.7)$$

and

$$a_j^* = a_j / s. \qquad (5.8)$$

If the logit is parameterized as $z_j = a_j\theta + c_j$, then the change of c_j is

$$c_j^* = c_j - a_j m / s. \qquad (5.9)$$

The asymptote parameter, g_j, is not affected by these changes.

A convenient way to characterize an arbitrary latent distribution with finite mean and variance is to compute the probability density at a finite number of suitably chosen values of θ and to normalize the densities by dividing by their total. The result is a so-called "discrete distribution on a finite number of points" (Mislevy, 1984). These normalized densities can be used as the weights, $A(X_k)$ in quadrature formulas such as (5.3).

This discrete representation of the latent distribution can be readily estimated from the item response model. The expected frequency at point X_k, given the item data from a sample of *N* respondents, is $\overline{N}_k$, the expected number of attempts defined by (5.6) above.

The estimated densities are these values divided by their total:

$$A^*(X_k) = \overline{N}_k / \sum_h^q \overline{N}_h. \qquad (5.10)$$

They are called empirical or "posterior" weights, as distinguished from the theoretical or "prior" weights, $A(X_k)$, assumed before the data are in hand.

Testing the goodness-of-fit of the IRT model

If data from a large sample of respondents is available, the fit of the model may be tested, either for the test as a whole, or item by item. The method of examining fit depends upon the number of items in the test.

(i) Very short tests (10 or fewer items)

If nearly all of the 2^n possible response patterns for an n-item test appear in the data, the overall goodness-of-fit of the model can be tested directly. The distinct response patterns must be counted to obtain the pattern frequencies $r_1, r_2, \ldots, r_{2^n}$.

If a few of these frequencies are zero, ½ may be substituted for each and the sum of these substitutions subtracted from the largest frequency. Then the likelihood ratio χ^2 statistic for the test of fit is

$$G^2 = 2\sum_{l=1}^{2^n} r_l \log_e \frac{r_l}{N\bar{P}(x_l)}, \qquad (5.11)$$

where $\bar{P}(x_l)$ is the marginal probability of pattern x_l given by (5.3). This χ^2 has degrees of freedom $2^n - kn - 1$, where *k* is the number of item parameters in the model. Significantly large values of the statistic indicate a failure of fit of one or more of the response models for the *n* items.

(ii) Short tests (11 to 20 items)

No dependable, formal test of fit yet exists for all of this range. But useful information about the fit of individual items may be obtained by inspecting standardized differences between the *posterior* probability of correct response at selected values of θ and the probabilities at those points computed from the corresponding fitted response model. These differences are called "standardized posterior residuals".

In terms of quantities defined above, the posterior probability of a correct response is computed for item *j* at the point X_k is the ratio $\bar{r}_{jk} / \bar{N}_k$, the terms of which are defined above.

The corresponding standardized posterior residual can be expressed as follows:

$$\delta_{jk} = \frac{\sum_{l}^{S} W_{lk}[x_{lj} - P(X_k)]}{\{\sum_{l}^{S} W_{lk}[x_{lj} - P(X_k)]\}^{1/2}} \qquad (5.12)$$

where

$$W_{lk} = \frac{r_l P(\mathrm{x}_l \mid X_k)}{P(\mathrm{x}_l)}. \qquad (5.13)$$

Values of this residual greater than, say, 2.0 may be taken to indicate some failure of fit of the model at the corresponding point. In interpreting such deviates, it is advisable to take into consideration the posterior weight, $A^*(X_k)$, at the point, since a discrepancy in a region of θ with very little probability in the population will have little effect on the performance of the model.

As an overall index of fit, we suggest the population root-mean-square of the posterior deviates. Its formula is

$$RMS(\delta_j) = \left[\sum_k^q \overline{N}_k \delta_{jk}^2 / \sum_k^q \overline{N}_k \right]^{1/2}.$$

Unfortunately, the posterior residuals seem to be too highly correlated to be successfully combined into a χ^2 statistic for the item. Neither do they take into account the sampling variance of $P_j(X_k)$ due to estimation of its item parameters, but this source of variation is presumably small.

(iii) Long tests (more than 20 items)

If the test is sufficiently long, the respondents in a sample of size N can be assigned with good accuracy to intervals on the θ-continuum on the basis of their estimated value of θ. For this purpose, we use the EAP estimate with whatever prior is assumed for item calibration. 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 item parameters is based. The number of respondents in each interval who respond *correctly* to item *j* can be tallied from their item scores. Finally, a likelihood ratio χ^2 statistic may be used to compare the resulting frequencies of correct and incorrect responses in the intervals with those expected from the fitted model at the interval mean, $\overline{\theta}_h$:

$$G_j^2 = 2\sum_{h=1}^{n_g} \left[r_{hj} \log_e \frac{r_{hj}}{N_h P_j(\overline{\theta}_h)} + (N_h - r_{hj}) \log_e \frac{N_h - r_{hj}}{N_h [1 - P_j(\overline{\theta}_h)]} \right], \quad (5.14)$$

where n_g is the number of intervals, r_{hj} is the observed frequency of correct response to item *j* in interval *h*, N_h is the number of respondents assigned to that interval, and $P_j(\overline{\theta}_h)$ is the value of the fitted response function for item *j* at $\overline{\theta}_h$, the average ability of respondents in interval *h*.

Because neither the MML nor the MMAP method of fitting the response functions actually minimizes this χ^2, the residuals are not under linear constraints and there is no loss of degrees of freedom due to the fitting of the item parameters. The number of degrees of freedom is therefore equal to the number of intervals remaining after neighboring intervals are merged if necessary to avoid expected values less than 5.

To diagnose cases of poor fit, one can inspect a plot of r_{hj} / N_h compared to $P_j(\theta_h)$. Ninety-five percent tolerance intervals on these points are

$$\pm 2\sqrt{P_j(\bar{\theta}_h)[1 - P_j(\bar{\theta}_h) / N_h}.$$

When the number of items is small, the standardized posterior deviates should be plotted instead.

5.4 Test scoring

Unlike classical test theory, IRT does not in general base the estimate of the respondent's ability (or other attribute) on the number-correct score. The only exception is the one-parameter logistic model, in which the estimate is a non-linear function of that score. To distinguish IRT scores from their classical counterparts, we refer to them as "scale" scores.

The main advantages of scale scores are that they

- remain comparable when items are added to or deleted from the tests,
- weight the individual items optimally according to their discriminating powers,
- have more accurate standard errors,
- provide more flexible and robust adjustments for guessing than the classical corrections, and
- are on the same continuum as the item locations.

There are three types of IRT scale score estimation methods now in general use:

- Maximum likelihood estimation
- Bayes estimation
- Bayes modal estimation.

The three types of IRT scale score estimation methods are discussed in the sections to follow.

Maximum likelihood estimation

The maximum likelihood (ML) estimate of the scale score of respondent *i* is the value of θ that maximizes

$$\log L_i(\theta) = \sum_{j=1}^{n} \{x_{ij} \log_e P_j(\theta) + (1 - x_{ij}) \log_e [1 - P_j(\theta)]\}, \qquad (5.15)$$

where $P_j(\theta)$ is the fitted response function for item *j*.

The implicit likelihood equation to be solved is

$$\frac{\partial \log L_i(\theta)}{\partial \theta} = \sum_{j=1}^{n} \frac{x_{ij} - P_j(\theta)}{P_j(\theta)[1 - P_j(\theta)]} \cdot \frac{\partial P_j(\theta)}{\partial \theta} = 0.$$

The ML estimate, $\hat{\theta}$, is conveniently calculated by the Fisher-scoring method, which depends on the so-called "Fisher information",

$$I(\theta) = \sum_{j=1}^{n} a_j^2 P_j(\theta)[1 - P_j(\theta)], \tag{5.16}$$

in the case of the two-parameter logistic model. Similar formulas are available for the other models. The iterations of the Fisher-scoring solution are

$$\hat{\theta}_{t+1} = \hat{\theta}_t + I^{-1}(\hat{\theta}) \left(\frac{\partial \log L_i(\hat{\theta})}{\partial \theta} \right).$$

The standard error of the ML estimator is the square root reciprocal of the information at $\hat{\theta}$:

$$\text{S.E.}(\hat{\theta}) = \sqrt{1 / I(\hat{\theta})}. \tag{5.17}$$

Unlike the classical standard error of measurement, which is a constant, the IRT standard error varies across the scale-score continuum. It is typically smaller towards the center of the scale where more items are located and larger at the extremes where there are fewer items. A disadvantage of the ML estimate is that it is not defined for the response patterns in which all items are correct or all items are incorrect (and occasionally for other unfavorable patterns near the chance level when the three-parameter model is used). These problems do not arise in the other two methods of estimation.

Bayes estimation

The Bayes estimate is the mean of the posterior distribution of θ, given the observed response pattern x_i (Bock & Mislevy, 1982). It can be approximated as accurately as required by the Gaussian quadrature (see the section on MML estimation):

$$\overline{\theta_i} \cong \frac{\sum_{k=1}^{q} X_k P(\mathbf{x}_i \mid X_k) A(X_k)}{\sum_{k=1}^{q} P(\mathbf{x}_i \mid X_k) A(X_k)}.$$

This function of the response pattern $\mathbf{x}_i$ has also been called the *expected a posteriori* (EAP) estimator. A measure of its precision is the posterior standard deviation (PSD) approximated by

$$PSD(\overline{\theta}_i) \cong \frac{\sum_{k=1}^{q}(X_k - \overline{\theta}_i)^2 P(\mathbf{x}_i \mid X_k) A(X_k)}{\sum_{k=1}^{q} P(\mathbf{x}_i \mid X_k) A(X_k)}.$$

The weights, $A(X_k)$, in these formulas depend on the assumed distribution of θ. Theoretical weights, empirical weights, $A^*(X_k)$, or subjective weights are possibilities.

The EAP estimator exists for any answer pattern and has a smaller average error in the population than any other estimator, including the ML estimator. It is in general biased toward the population mean, but the bias is small within $\pm 3\sigma$ of the mean when the PSD is small (*e.g.*, less than 0.2σ). Although the sample mean of EAP estimates is an unbiased estimator of the mean of the latent population, the sample standard deviation is in general smaller than that of the latent population. This is not a serious problem if all the respondents are measured within the same PSD. But it could be a problem if respondents are compared using alternative test forms that have much different PSDs. The same problem occurs, of course, when number-right scores from alternative test forms with differing reliabilities are used to compare respondents. Tests administrators should avoid making comparisons between respondents who have taken alternative forms that differed appreciably in their psychometric properties. A further implication is that, if EAP estimates are used in computerized adaptive testing, the trials should not terminate after a fixed number of items, but should continue until a prespecified PSD is reached.

Bayes modal estimation

Similar to the Bayes estimator, but with a somewhat larger average error, is the Bayes modal or so-called *maximum a posteriori* (MAP) estimator. It is the value of θ that maximizes

$$P(\theta \mid x_i) = \sum_{j=1}^{n}\{x_{ij} \log_e P_i(\theta) + (1 - x_{ij}) \log_e [1 - P_i(\theta)]\} + \log_e g(\theta),$$

where $g(\theta)$ is the density function of a continuous population distribution of θ. The likelihood equation is

$$\sum_{j=1}^{n} \frac{x_{ij} - P_j(\theta)}{P_j(\theta)[1 - P_j(\theta)]} \cdot \frac{\partial P_j(\theta)}{\partial \theta} + \frac{\partial \log_e g(\theta)}{\partial \theta} = 0.$$

Analogous to the maximum likelihood estimate, the MAP estimate is calculated by Fisher scoring, employing the *posterior information*,

$$J(\theta) = I(\theta) + \frac{\partial^2 \log_e g(\theta)}{\partial \theta^2},$$

where the right-most term is the second derivative of the population log density of θ.

In the case of the 2PL model and a normal distribution of θ with variance σ^2, the posterior information is

$$I(\theta) = \sum_{j=1}^{n} a_j^2 P_j(\theta)[1 - P_j(\theta)] + \frac{1}{\sigma^2}.$$

The PSD of the MAP estimate, θ, is approximated by

$$\text{PSD}(\theta) = \sqrt{1 / I(\hat{\theta})}.$$

Like the EAP estimator, the MAP estimator exists for all response patterns, but is generally biased toward the population mean.

5.5 Test and item information

According to classical theory, the standard error of measurement (SEM) is a function only of the test reliability and the variance of the score distribution. But this is an oversimplification. Actually, the error standard deviation of a score on a test of finite length—whether the classical number-right score or an IRT scale score—also depends upon the level of the score itself.

When the maximum likelihood estimator is used to obtain an IRT scale score, the SEMs of the three logistic models expressed in the normal metric are as follows:

1PL:

$$\text{S.E.}_{(1)}(\hat{\theta}) = \left\{ 1 / D^2 a^2 \sum_{j=1}^{n} P_{(1)j}(\hat{\theta})[1 - P_{(1)j}(\hat{\theta})] \right\}^{\frac{1}{2}} \tag{5.18}$$

2PL:

$$\text{S.E.}_{(2)}(\hat{\theta}) = \left\{ 1 / D^2 \sum_{j=1}^{n} a_j^2 P_{(2)j}(\hat{\theta})[1 - P_{(2)j}(\hat{\theta})] \right\}^{\frac{1}{2}} \tag{5.19}$$

3PL:

$$\text{S.E.}_{(3)}(\hat{\theta}) = \left\{ 1/D^2 \sum_{j=1}^{n} a_j^2 \frac{1 - P_{(3)j}(\hat{\theta})}{P_{(3)j}(\hat{\theta})} \cdot \left(\frac{P_{(3)j}(\hat{\theta}) - g_j}{1 - g_j} \right)^2 \right\}^{\frac{1}{2}} \tag{5.20}$$

Although these formulas are more realistic than the classical standard error of measurement, they are nevertheless approximations. Strictly speaking, they are exact only as the number of items becomes indefinitely large. But in general, they are good approximations for tests with as few as ten or twenty items. Although they neglect the errors due to estimating the item parameters, these errors are inconsequential if the calibration sample is large.

Because the terms that are summed in the information functions (5.18), (5.19), and (5.20) can be regarded as the information functions of the *items*, they show how the SEM depends upon the item slopes, locations and lower asymptotes. By plotting the item information functions of the items against the test information, the test constructor can see which items are contributing most to increasing the test information in relevant regions of the scale, and thus to decreasing the SEM. The plots show where additional items are needed to improve the precision of measurement locally. Generally, the aim is to produce a test item function that is high and flat over the range of θ in which accurate measurement is required.

It is evident in the information functions for the logistic models that, as $P_j(\theta)$ goes to 1 or 0 (or to g_j for the 3PL model), the information goes to zero and the standard error to infinity. Thus, the ML estimator is effective only over a finite range. As a result, it is necessary to set some limit, perhaps ±5 standard deviations of the latent distribution, as upper and lower bounds of $\hat{\theta}$.

The posterior information for the Bayes modal (MAP) estimator has properties similar to those of the Fisher information of the ML estimator except that, when the prior is suitably chosen (*e.g.*, normal) the posterior information does not go to zero as θ becomes extreme. Rather, for a normal prior, the posterior information goes to $1/\sigma^2$, and the SEM goes to the population standard deviation, σ, which means that nothing is known about θ except that it is very large or very small, depending on the sign of $\hat{\theta}$.

The squared inverse posterior standard deviation (PSD) of the Bayes (EAP) estimator does not have the convenient additive property of the Fisher and posterior information. But because of the equivalence of the EAP and MAP estimators as the number of items becomes large, ML information analysis of items can be applied to the EAP estimation for most practical purposes of test construction.

5.5.1 Effects of guessing

Guessing in response to multiple-choice items has a deleterious effect on any estimator of ability, classic or IRT. For the three-parameter model, the average effect of guessing, and thus the size of the asymptote parameter, g_j, can be reduced by instructing the examinees to omit the item rather than make a blind guess. But when the three-parameter model is used in scoring, it does not distinguish between those examinees who omit and those who ignore the instructions, and guess.

Two methods of improving the accuracy of scale score estimation in the presence of mixed omitting and guessing have been proposed. One method is to assign to the omitted responses a probability equal to the asymptote parameter, g_j, or to $1/A$, where A is the number of alternatives of the multiple-choice item (Lord, 1980, p. 229). In effect, the omitted responses are replaced by guessed responses and scored fractionally correct.

The other method is to score omits as incorrect, but suppress the effects of guessing by giving reduced weight to unlikely correct responses in the response pattern. A technique of robust data analysis, called "biweighting", has been proposed for this purpose (Mislevy & Bock, 1982). Simulation studies have shown that such robustifying procedures improve the accuracy of estimating ability in the presence of chance successes in response to multiple-choice items.

5.5.2 Aggregate-level IRT models

In some forms of educational assessment, scores are required for populations of groups and students (schools, for example) rather than for individual students (Mislevy, 1983). In these applications, IRT scale scores for the groups can be estimated directly from matrix sampling data if the following conditions are met:

- The assessment instrument consists of 15 or more randomly parallel forms, each of which contain exactly one item from each content element to be measured.
- The forms are assigned in rotation to students in the groups being assessed and administered under identical conditions.

On these conditions, it may be reasonable to assume that the ability measured by each scale is normally distributed within the groups. In that case, the proportion of students in the groups who respond correctly to each item of a scaled element will be well approximated by a logistic model in which the ability parameter, θ, is the mean ability of the group. Because each item of the element appears on a different form, these responses will be experimentally independent.

An aggregate-level IRT model can therefore be used to analyze data for the groups summarized as the number of attempted responses, N_{hj}, and the number of correct responses, r_{hj}, to item j in group h. The probability of these response frequencies for the n items of the element, given the mean ability of the group, $\overline{\theta}_h$, is then

$$P(r_h \mid \overline{N}_h, \overline{\theta}_h) = \prod_{j=1}^{n} \frac{N_{hj}!}{(N_{hj} - r_{hj})! r_{hj}!} \cdot [\Psi_j(\overline{\theta}_h)]^{r_{hj}} [1 - \Psi_j(\overline{\theta}_h)]^{N_{hj} - r_{hj}}. \tag{5.21}$$

Using (5.21) in place of the individual MML estimator

$$P(\mathbf{x} \mid \theta) = \prod_{j=1}^{n} [P_j(\theta)]^{x_j} [1 - P_j(\theta)]^{1-x_j}$$

we can carry out MML estimation of item parameters for the aggregate-level IRT model in the same manner as estimation for the individual-level model. Scale scoring of the pattern of frequencies of attempts and correct is performed by a similar substitution in (5.15), (5.16) or (5.17). All other aspects of the IRT analysis are unchanged.

Unlike the individual-level analysis, the aggregate-level permits a rigorous test of fit of the response pattern for the group. Because the response frequencies for the items of a scaled element are binomially distributed and independent, a likelihood ratio or Pearsonian χ^2 test statistic may be computed to test the fit of the model within each group.

The starting values computed in the INPUTphase and used in item parameter estimation in the calibration phase in BILOG-MG are generally too high for aggregate-level models. The user should reduce these values by substituting other starting values in the TEST command.

5.5.3 Estimating the population distribution

For purposes of norming test results, it is necessary to estimate the distribution of test scores in the population of interest. This presents a problem for classical test theory for two reasons. First, the number-correct test scores contain both true score variation and measurement error variation; since the measurement error variance is a function of test length, the variance of the score distribution therefore depends on an arbitrary choice in test construction. Second, the shape of the test score distribution depends arbitrarily upon the distribution of item difficulties in the test; tests with severely skewed distributions of item difficulties will produce skewed distributions of test scores in the population.

Classical test theory sidesteps these problems by expressing norms as population percentiles, which are invariant with respect to the spread or shape of the score distribution. Further analysis of the test scores by statistical methods that assume a normal distribution may still be affected, however. IRT theory is more favorable in this respect in that the shape of the observed scale score distribution is relatively little influenced by the distribution of item difficulties. If the true score distribution is approximately normal, for example, the scale score distribution will be also. The variance of the latter is still increased by measurement error, but as is also true of test scores, the effect can be largely suppressed independent of test length by computing so-called “regressed” or “shrunken” estimates as a function of test reliability. The Bayes (EAP) regressed and Bayes modal (MAP) scores provided by the programs are regressed estimates.

IRT can handle this problem more rigorously, however, by estimating an inferred *latent* distribution of proficiency scores. The shape of the latent distribution is estimated directly from the patterns of correct and incorrect responses to the test items and does not involve the test scores. If there is only one sample group in the analysis, the location and dispersion of the latent distribution are indeterminant and must be set arbitrarily (*e.g.* to 0 and 1). If there are multiple sample groups in the analysis, locations and dispersions of their latent distributions can be set relative to a designated reference group or relative to arbitrarily set values in the combined groups. Multiple-group analysis is implemented in the BILOG-MG program.

5.5.4 Differential item functioning

Almost any population of potential test takers will consist of identifiable subpopulations—different age groups, the two sexes, urban or rural residents, education levels, ethnic and language groups, etc. Relevant information on group membership may be available from background questionnaires administered along with the test. If so, the data will allow investigation of whether persons in one such group experience differences in item difficulty or discriminating power relative to those in other groups when all groups have equal mean scores on the test as a whole. When this is the case, the test is said to exhibit *differential item functioning* (DIF). DIF is essentially item by group interaction in item difficulty or discriminating power. If at the same time, the groups show unequal mean test scores, the test is said to have *adverse impact* on the groups that perform more poorly. Adverse impact can, of course, also occur in the absence of DIF. Since DIF in effect alters the substantive meaning of the test score from one group to another, it is undesirable and should be eliminated if possible. An English language vocabulary test with words of Latin or Germanic origin, for example, will tend to show DIF with respect to English or Spanish as first language acquired. If only a few items of the test exhibit DIF, they usually can be removed without impairing measurement of the intended construct.

The problem for the data analyst is how to detect DIF in tests that may also show adverse impact. There are both classical and IRT approaches to this problem. The classical methods look for differences in item difficulty among persons from different background groups whose tests scores are equal or fall in a narrow score interval. A summary statistic for these differences over the scores or score intervals provides a measure of DIF; an associated statistical test establishes its presence. Based on a log-linear model of item by group interaction, a similar analysis can be carried out with the so-called Mantel-Haenszel statistic.

The IRT treatment of DIF is an example of multiple-group analysis in which item thresholds or discriminating power are estimated separately in each group and jointly with the group latent distributions, under the restriction that the means of the item thresholds must be equal in all groups. The item guessing parameters, if any, are also restricted to be equal among groups. IRT estimation of DIF effects includes standard errors that can be used to assess statistical significance of effects for individual items. In addition, a test of DIF in all items jointly is provided by comparison of the goodness-of-fit of the response model when different thresholds or discriminating power are assumed vs. the fit when a single set of thresholds or discriminating power is estimated in the combined data. The IRT method of analyzing DIF is in general more sensitive than its classical counterparts, especially with shorter tests, because IRT better defines the latent construct measured by the test. DIF in item difficulty is implemented in BILOG-MG.

5.5.5 Forms equating

Many testing programs must update their test at regular intervals to prevent overexposure and compromise of the item content. This creates the problem of keeping the reporting scores for successive forms comparable so that a person is neither advantaged nor disadvantaged by taking one form rather than another. Somehow, the reported results must allow for the differences in overall difficulty of the forms that inevitably occur when items are changed. Classical test theory solves this problem by *equivalent-groups* equating. This method requires the alternative forms to be assigned randomly to persons in some large sample. The randomization ensures that persons taking different forms will have essentially the same true score distribution (provided that the successor forms are of the same length as the preceding form and have similar distributions of item difficulties and discriminating powers). If these conditions are met, the test scores for the new forms can be expressed on the same scale as the old forms by assigning them to the corresponding points of their respective observed score distributions. This is the *equi-percentile* method of keeping the score reports comparable to one another through successive generations of test forms. If the distribution of item difficulties within the forms is more or less normally distributed and well centered for the population, the test score distributions will be approximately normal. In that case, a nearly equivalent equating can be obtained merely by standardizing the scores of the various forms—that is, by setting the mean and standard deviations of their respective distributions to any convenient fixed values. This method is called *linear* equating.

Since IRT scale scores are much more likely to approximate normal distribution than number-right scores, equi-percentile equating is less needed in IRT applications. Linear equating suffices, and it happens automatically if the origin and unit of the IRT scale is set so that the scale scores have a specified mean and standard deviation in the sample. In addition, IRT is unique in allowing the equating forms administered to *non-equivalent* groups—*i.e.*, groups with different true score distributions. This type of equating requires, however, that the test forms share a certain number of common "linking" items. Provided the linking items do not exhibit DIF with respect to the groups, *multiple-group* IRT analysis of all forms together automatically produces a single IRT scale on which the reported scores are comparable. The multiple-group procedure estimates separate latent distributions of the groups jointly with the item parameters of all the forms. The advantage of this method is that it does not require a separate administration of the forms to some group of persons for purposes of equivalent groups equating. Forms can, for example, be updated in the course of operational administrations of the test in which a certain proportion of items from the previous year's forms is carried over to the current year's forms. A random sample of examinees from the previous year's operational testing provides data for one of the groups, and a similar sample from the current year provides data for the other. The resulting scale scores are linearly equated to those of the previous year by setting the mean and standard deviations of the latent distribution of the first group to its previous year's values. Estimates of change in the mean and standard deviation between years are a by-product of the equating. If desired, non-equivalent groups equating can be carried back more than one year, provided linking items exist between at least adjacent pairs of forms. Multiple-groups forms equating is implemented in BILOG-MG.

5.5.6 Vertical equating

In school systems with a unified primary and secondary curriculum, there is often interest in monitoring individual children's growth in achievement from Kindergarten through eighth grade. A number of test publishers have produced articulated series of tests covering this range for subject matter such as reading, mathematics, language skills, and, more recently, science. The tests are scored on a single scale so that each child's gains in these subjects can be measured. The analytical procedure for placing results from the grade-specific test forms on a common scale for this purpose is referred to as *vertical equating*.

The most widely used classical method of vertical equating is the transformation of test scores into so-called *grade equivalents*. In essence, the number-correct scores for each year are scaled in such a way that the mean score for the age group is equal to the numerical values of the grades zero through eight. This convention permits a child's performance on any test in the series to be described in language similar to that used with the Binet mental age scale. One may say of a child whose reading score exceeds the grade mean, for example, that he or she is "reading above grade level".

For IRT, vertical equating is merely another application of non-equivalent groups equating in which the children administered particular grade-specific tests correspond to the groups. As in the equating of updated forms mentioned above, linking items between at least consecutive forms in the series are required. They must be provided in each subject matter included in the graded series. (Note that grade-equivalent scaling does not require linking items.)

The two methods produce quite different scales. Grade equivalents are of course linear in school grade. They treat the average gain between first and second grade, for example, as if it were equal to that between seventh and eighth. On this scale, the amount of variation between children's scores appears to increase as the cohort moves through the grades, and there is a corresponding positive correlation between a child's average score level over the years and the child's average gain. In other words, children who begin at a lower level appear to gain less overall than those who begin at a higher level. This so-called "fan-spread effect" is regularly seen in all subject matters.

On an IRT vertically equated scale, average gains are generally greatest at the earlier grade levels and decrease with increasing grade. Within grade, standard deviations are fairly uniform, and the correlation between children's average score levels and their gains are small, or even slightly negative in some subject matters.

Unfortunately, there is no objective basis for deciding which of these scales better represents a child's true course of growth in knowledge and skills during the school years. Different IRT models assuming other transformations of the proficiency scale could be made to fit the item response data equally well and yet exhibit much different appearing relationships between grade level and average score or average gain. Extrinsic considerations would have to be brought to bear on the question to determine a preferred scale. For example, if one wished to compare annual average gains in test performance of children in different classrooms when assignment to classrooms is non-random, the scale that showed zero correlation between level and gain would be most advantageous. IRT vertical equating comes much closer to this ideal than grade equivalents, but might require some further transformation, possibly subject matter and site specific, to attain complete independence of level and gain. (See Bock, Wolfe & Fisher, 1996, for a discussion of this topic).

5.5.7 Construct definition

The discussion up to this point assumes that all items in the test measure the same underlying construct. When it is not clear that the item set is homogeneous in this sense, steps must be taken to explore the construct dimensionality of the set. The classical approach to this problem is to perform, in a large sample of test data, a multiple factor analysis of the matrix of tetrachoric correlations between all pairs of items. The more familiar Pearson product-moment correlation of item responses assigned different numerical values if correct or incorrect (phi coefficient) is not generally satisfactory for this purpose because variation in item difficulties introduces spurious factors in the results. Random guessing on multiple-choice items has a similar effect that must also be allowed for. Tetrachoric correlations with corrections for guessing are largely free of these problems, but they have others of their own. One of these is computational instability that appears when the correlations have large positive or negative values and the item difficulties are very low or high; in these cases, it is often necessary to replace the correlation in the matrix with an attributed default value. The other problem is that factor analysis of tetrachoric correlation matrices almost always produces a certain number of small, unreal factors that are meaningless and must be discarded.

IRT improves on this procedure by a method of *full information* item factor analysis that operates directly on the patterns of correct and incorrect responses without intervening computation of correlation coefficients. In effect, this method fits a multidimensional item response model to the patterns in the sample data. Full information item factor analysis is robust in the presence of omitted or not-presented items and is free of the artifacts of the tetrachoric method. It also provides a statistical test of the number of factors that can be detected in the data.

The objective of both classical and IRT item factor analysis is the identification of items with similar profiles of factor loadings—an indication that they arise from the same cognitive or affective sources underlying the responses of persons taking the test. Objective methods of rotating the factor structure, such as orthogonal varimax rotation and non-orthogonal promax rotation, are especially effective in picking out clusters of items that identify these implicit constructs. The presence of significant multiple factors in the data means that there are corresponding dimensions of variation in the population of persons. In some cases, actual subgroups in the population associated with particular factors can be identified by including demographic variables in the analysis. Alternatively, they may be found by conventional multiple regression analysis of factor scores for the persons, which are also provided by IRT full information item factor analysis.

5.5.8 Analysis and scoring of rated responses

When tests contain items or exercises that cannot be scored mechanically, the responses are often rated on a graded scale that indicates quality or degree of correctness. For individually administered intelligence tests the grading is done by the test administrator at the time the response is recorded. For group administered open-ended exercises and essay questions, written responses are graded later by trained raters. In both cases, the additional information conveyed, beyond that provided by correct-incorrect scoring, provides better justification of the considerable cost of graded scoring.

In addition to problems that may arise in preparing the rating protocol and training the raters for graded scoring, analysis of the resulting data presents other difficulties not encountered with correct-incorrect scoring. How to combine the ratings into an overall score in a rational way is not at all clear in classical test theory—especially so if the test also includes multiple-choice items. The classical approach never goes much beyond mere assignment of arbitrary numerical values to the scale categories and summing these values to obtain the test score. The arbitrariness of this method, and the fact that items with different numbers of rating categories receive different weights in the sum, has always proved troublesome.

In this respect, IRT methods are a very considerable advance. Item response models now exist that express the probability of a response falling in a given graded category as function of 1) the respondent's position on the IRT scale, 2) parameters for the spacing of the categories, and 3) the difficulty and discriminating power of the item. Models for items with different numbers of rating categories and models for dichotomously scored responses can be mixed in any order when analyzing items or scoring tests; arbitrary assignments of score points are not required. The IRT test scoring based on these models makes use of the information in the pattern of ratings in a way that is internally consistent in the data and minimizes measurement error. The IRT approach to graded data allows tests to have more interesting and varied item formats and makes them accessible to IRT methods of construction and forms equating.

5.5.9 Matrix sampling

Testing at the state and national level plays a part not only in counseling or qualification of individual students, but also in evaluating the effectiveness of instructional programs, schools, or school systems. The objective is to compare instructional programs and schools with respect to their strength and weaknesses in promoting student achievement in various categories of the curriculum. Testing used in this way is referred to as *assessment* to distinguish it from student-oriented achievement testing. Educational assessment is typically carried out in large-scale surveys, often on a sampling basis rather than a total census of schools and students. The sampling approach consists of drawing a probability sample of schools and, within these schools, testing a random sample of students. To minimize the burden on schools and students alike there is an attempt to test as many curricular categories as possible in a limited amount of time, usually one class period. This is accomplished by assigning randomly to the selected students one of 20 or 30 different test forms each containing only a small number of the items representing a category. Usually the categories are main topic areas within subject matters. The total sampling design can be laid out as a table in which the rows correspond to schools and students tested and the columns correspond to items sampled for the test forms and the categories within forms. This arrangement is referred to as a *matrix sample*.

In the original conception of a matrix-sampled assessment, the score to be assigned to programs, schools, states, or demographic groups is the average aggregated percent-correct for the items in each subject topic. As an aid to interpretation of differences between groups or between assessment years, statistical theory provides formulas for standard errors for these average scores under the assumptions of matrix sampling. This treatment of the data is within the framework of the number-correct score concept of classical test theory, although no explicit scores for individual students are computed.

A problem with average-percent correct reporting occurs, however, if the assessment aims at monitoring trends in average achievement over successive years. When the time comes to update the items of the assessment instrument, new items substituted for old inevitably introduce changes in average scores at higher levels of aggregation—changes which may be larger than the expected differences between year or programs or schools. Although scores on the successive instruments can be made comparable by equivalent groups equating, very large sample groups are required to bring the equating errors below the size of the smallest difference that would have policy implications.

IRT nonequivalent groups equating, which can be done in the full operational samples, is much more cost effective in this situation. It requires only that a certain proportion of items from the previous assessment be carried over into the update to serve as links between successive forms. Typically, one-third of the items are retained as links. A large random sample of cases from the two assessments are then analyzed in a multiple-group IRT calibration that estimates the latent distributions for the two assessment samples jointly with the new set of item parameters. The link items serve to set the origin and unit of scale equal to those of the previous assessment.

Paralleling the average percent-correct approach, IRT can also estimate scores at the group-level without intervening score estimation for individual students. This can be done in one of two ways. If the interest is only in comparing mean scores among schools or higher level aggregates, these quantities can be estimated directly from counts of the number of times each item is presented to a student in the group, and of these, the proportion correct. The group means are estimated on a scale that is standardized by setting the mean of the estimated group means weighted by the numbers of persons testing in the respective groups, and the standard deviation of the estimated group means calculated in a similar weighted form. Standard errors with respect to the sampling of students within schools are available for the estimated school means, and the higher-order aggregate means.

If it is also of interest, however, to know something about the distribution of student achievement within the aggregate groups, multiple-group IRT analysis can be used to estimate the latent distributions within the groups directly, without estimating scores for individual respondents. The procedure is more efficient for a definite form of latent distribution, such as the normal or other distribution that depends on a relatively small number of parameters. If a completely general form is assumed, a nonparametric procedure, possibly involving computer simulations, may be necessary.

5.5.10 Estimating domain scores

Both classical test theory and item response theory have to contend with the arbitrary nature of test scores as measurements. As mentioned above, the classical number-correct score, and even the length-independent percent-correct score, depend arbitrarily upon the difficulties of the items selected for the test. The IRT scale score, although relatively free of that problem, is nevertheless expressed on a scale of arbitrary origin and unit. The earliest and still most widely used method of removing this arbitrariness is to scale the scores relative to their distribution in some large sample of persons taking the test. This is most commonly done by expressing the scores as percentiles of the distribution or as standardized scores, *i.e.*, subtracting the mean of the distribution from the observed score and dividing by the distribution standard deviation. This approach to reporting test scores is called *norm* referencing; it assumes that comparisons between persons is the object of the testing, which undeniably it is in selection testing.

In the context of qualification testing, however, a more relevant objective is whether a person taking the test shows evidence of having learned or mastered a satisfactory proportion of the knowledge and skills required for qualification. Similarly, in program evaluation the objective is whether a sufficient proportion of students in a program has reached a satisfactory level of learning or mastery. Reporting test results in these terms is referred to as *domain* referencing, or in a somewhat similar usage, *criterion* referencing. For domain referencing to be realizable in practice, some reasonably large pool of items or exercises must exist to define the domain operationally. Particular tests containing items or exercises from the pool may then be selected for purposes of estimating domain scores.

The classical method of domain score estimation is to assume that items of the test are a random sample of the pool. In that case, the test percent-correct directly estimates the domain percent-correct, and its standard error can be computed from the test's generalizability coefficient. IRT can improve upon this estimate if response models for items in the pool have been calibrated in data from a relevant population of examinees. With this information available, the items selected for a particular test do not need to be a random sample of the pool. They need only be link items in tests calibrated by non-equivalent groups equating. In that case, one estimates the domain score by first estimating the person's IRT scale score, then substituting the score in the model for each test item to compute the person's corresponding probability of correct response: the IRT estimated domain score is the sum of these probabilities divided by the number of items on the test. Domain scores estimated in this way are more accurate than classical estimates because they take into account the varying difficulty and discriminating power of the items making up the test. These methods of estimation can be carried out with multidimensional as well as unidimensional response models. Domain scores are implemented in the BILOG-MG program.

5.5.11 Adaptive testing

Adaptive testing is a method of test administration in which items are chosen that are maximally informative for each individual examinee. Among items with acceptable discriminating power, those selected are at a level of difficulty that affords the examinee a roughly 50 percent probability of correct response. This corresponds to minimum *a priori* knowledge of the response, and thus maximum information gain from its observation.

The two main forms of adaptive test administration are *two-stage* testing and *sequential* item testing. In the two-stage method, which is suitable for group administration, a brief first-stage test is administered in order to obtain a rough provisional estimate of each examinee's proficiency level. At a later time, a longer second-stage test form is administered at a level of difficulty adapted to the provisional score of each examinee. In sequential adaptive testing, usually carried out by computer, a new provisional estimate of the examinee's proficiency is calculated after each item presentation, and a most informative next item is chosen based on that estimate. The presentation sequence begins with an item of median difficulty in the population from which the examinee is drawn. Depending on whether the response to that item is correct or incorrect, the second item chosen is harder or easier. The presentations continue in this manner until the successive provisional estimates of proficiency narrow-in on a final value with acceptably small measurement error. Unlike two-stage testing, this method of administration requires the adaptive process to be carried out during the testing session. For this reason computer administration is possible only if the items are machine scorable.

When IRT scale scores are used to obtain the provisional estimates of proficiency in computerized adaptive testing, the presented items must be calibrated beforehand in data obtained non-adaptively. Once the system is in operation, however, items required for routine updating can be calibrated "on line". For this purpose, new items that are not part of the adaptive process must be presented to examinees at random, usually in the early presentations. Responses to all items in the sequence are then saved and assembled from all testing sites and sessions. A special type of IRT calibration called *variant item* analysis is applied in which parameters are estimated for the new "variant" items only; parameters of the old items are kept at the values used in the adaptive testing. Because IRT calibration as well as scoring can be carried out on different arbitrary subsets of item presented to respondents, the parameters of the variant items are correctly estimated in the calibration even though the old items have been presented non-randomly in the adaptive process. Variant item analysis is implemented in the BILOG-MG program.

With different examinees presented items of differing difficulty in adaptive testing, the number-correct score is not appropriate for comparing proficiency levels among examinees. For this reason, no treatment of adaptive testing appeared within classical test theory, and hardly any discussion of the topic arose until item response theory made it possible to estimate comparable scores from arbitrary item subsets. That development, combined with the availability of computer terminals and microcomputers, has made sequential testing a practical possibility. Significant applications of computerized adaptive testing have followed, particularly in the area of selection testing. Apart from its logistical and operational convenience, the primary benefit of this method of test administration is in reducing testing time. As little as one-third of the time required for a non-adaptive test suffices for a fully adaptive sequential test of equal precision.

6 BILOG-MG examples

6.1 Conventional single-group IRT analysis

This example illustrates how the BILOG-MG program can be used for traditional IRT analyses. The data are responses to 15 multiple-choice mathematics items that were administered to a sample of eighth-grade students. The answer key and the omitted response key are in files called **exampl01.key** and **exampl01.omt**, respectively (defined on the INPUT command).

The data lines, of which the first few lines are shown below, contain 15 item responses. This is the simplest form in which raw data can be read from file: there is one line of data for each examinee, and the response to item 1, for example, can always be found in column 6. All items are used on the single subtest. Item responses start in column 6 as reflected in the format statement (4A1,1X,15A1).

```
1 242311431435242
2 243323413213131
3 142212441212312
4 341211323253521
```

Exampl01.key contains a single line:

```
KEY  341421323441413
```

With such a short test (15 items), item chi-squares are not reliable. For illustration purposes the minimum number of items needed for chi-square computations has been reduced from the default of 20 to the number of items in this test, using the CHI keyword on the CALIB command. With the item chi-squares computed, the PLOT=1 specification can now be used to plot all the item response functions.

Note that the ICC s produced with the IRTPLOT program in the Windows version of BILOG-MG display the χ^2-test statistics, degrees of freedom, and probability, as well as the observed response probabilities only for those items that have a significance level below the value specified with the PLOT keyword.

The scoring phase includes an information analysis (INFO=2) with expected information indices for a normal population (POP). Rescaling of the scores and item parameters to mean 0 and standard deviation 1 in the estimated latent distribution has been requested (RSC=4). Printing of the students' scores on the screen is suppressed (NOPRINT), because that information is saved in the **exampl01.sco** file.

```
EXAMPL01.BLM - TRADITIONAL IRT ANALYSIS OF A FIFTEEN-ITEM PRETEST FROM
               A TWO-STAGE TEST OF MATHEMATICS AT THE EIGHTH-GRADE LEVEL
>GLOBAL  DFNAME='EXAMPL01.DAT', NPARM=3, SAVE;
>SAVE    PARM='EXAMPL01.PAR', SCORE='EXAMPL01.SCO';
>LENGTH  NITEMS=15;
>INPUT   NTOTAL=15, NALT=5, NIDCHAR=4,
         KFNAME='EXAMPL01.KEY', OFNAME='EXAMPL01.OMT';
>ITEMS   INAMES=(MATH01(1)MATH15);
```

```
>TEST1   TNAME='PRETEST', INUMBER=(1(1)15);
(4A1,1X,15A1)
>CALIB   NQPT=31, CYCLES=25, NEWTON=10, CRIT=0.001, ACCEL=0.0, CHI=15, PLOT=1;
>SCORE   NOPRINT, RSCTYPE=4, INFO=2, POP;
```

Phase 1 output

This is a standard 3-parameter, one-form, single-group analysis of a 15 item test. The Phase 1 classical item statistics for the first 5 items are as follows.

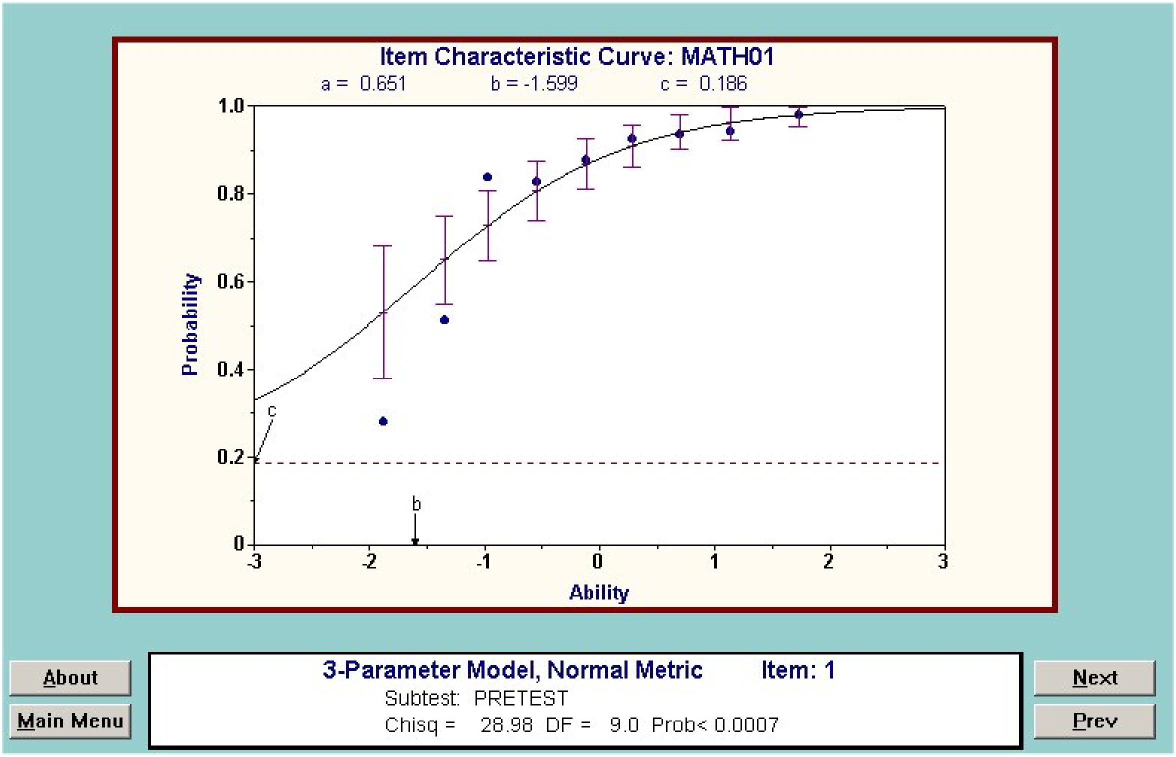
```
ITEM STATISTICS FOR SUBTEST PRETEST
                                                          ITEM*TEST CORRELATION
ITEM  NAME        #TRIED    #RIGHT   PCT    LOGIT/1.7  PEARSON  BISERIAL
 ------------------------------------------------------------------------
    1   MATH01      1000.0      844.0    84.4    -0.99      0.274    0.415
    2   MATH02      1000.0      972.0    97.2    -2.09      0.112    0.285
    3   MATH03      1000.0      696.0    69.6    -0.49      0.356    0.468
    4   MATH04      1000.0      503.0    50.3    -0.01      0.442    0.553
    5   MATH05      1000.0      594.0    59.4    -0.22      0.477    0.603
 ------------------------------------------------------------------------
```

Phase 2 output

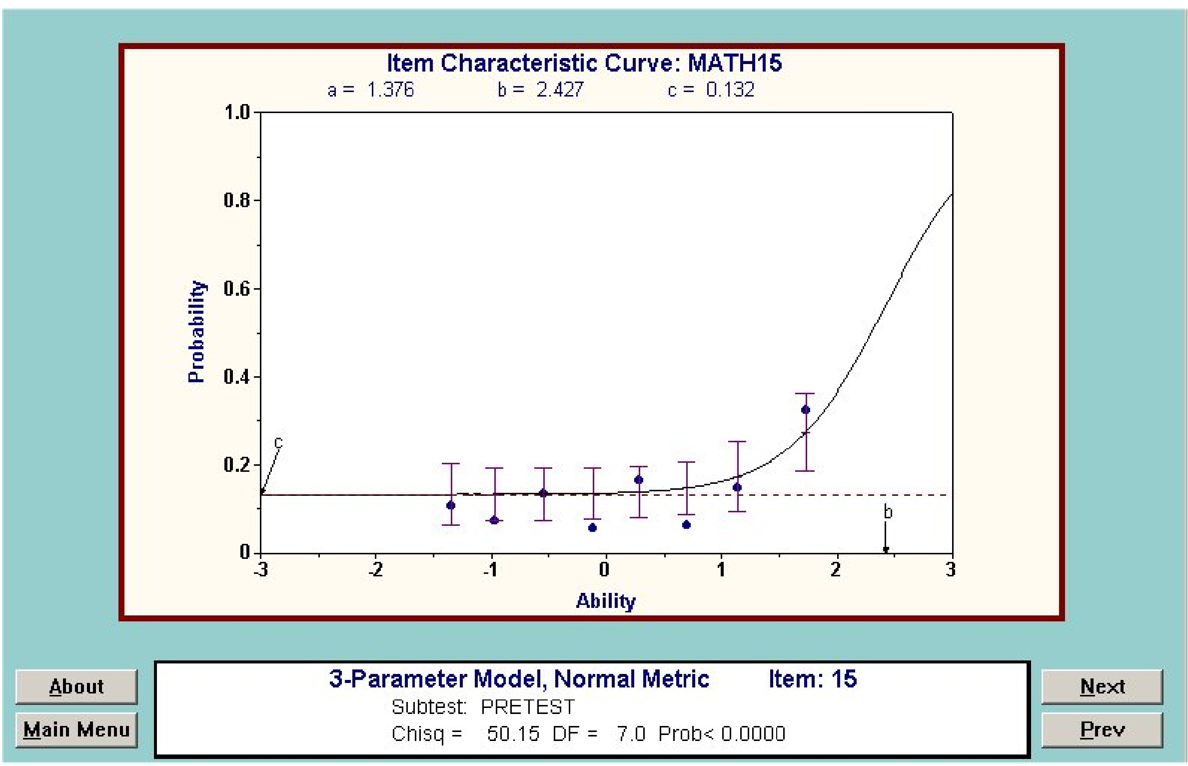
No new features are illustrated in the Phase 2 analysis, except that the plot criterion has been set to include all items.

```
>CALIB   NQPT=31, CYCLES=25, NEWTON=10, CRIT=0.001, ACCEL=0.0, CHI=15, PLOT=1;
```

The first and last item response function plots are shown below. The first item is extremely easy and the last extremely difficult. These plots were produced using the IRTGRAPH procedure, which is accessed via the **Plot** option on the **Run** menu after completion of the analysis. Note that the Phase 2 output file also contains similar line plots.



Item Characteristic Curve: MATH01
a = 0.651
b = -1.599
c = 0.186
1.0
0.8
0.6
0.4
0.2
0
Probability
-3
-2
-1
0
1
2
3
Ability
c
b
About
Main Menu
3-Parameter Model, Normal Metric
Item: 1
Subtest: PRETEST
Chisq = 28.98 DF = 9.0 Prob< 0.0007
Next
Prev



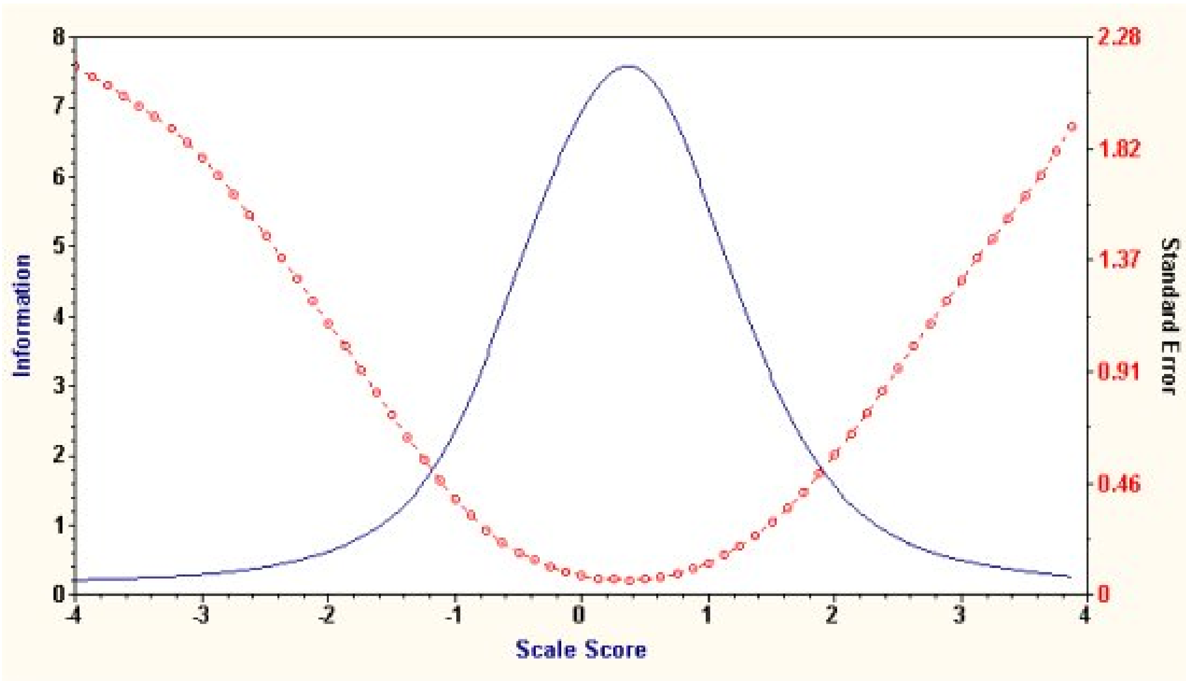
Item Characteristic Curve: MATH15
a = 1.376
b = 2.427
c = 0.132
1.0
0.8
0.6
0.4
0.2
0
Probability
-3
-2
-1
0
1
2
3
Ability
c
b
About
Main Menu
3-Parameter Model, Normal Metric
Item: 15
Subtest: PRETEST
Chisq = 50.15 DF = 7.0 Prob< 0.0000
Next
Prev

Phase 3 output

With this short, wide range test, ten quadrature points are sufficient for scoring. The item parameters are rescaled so that the scores have mean zero and standard deviation one in the latent distribution estimated from the full sample of 1000 examinees. Population characteristics of the score information, including the IRT estimate of test reliability (equal to [score variance–1/average information] / score variance) are shown with the information plot.

```
>SCORE, NOPRINT, RSCTYPE = 4, INFO = 2, POP;

                QUAD   RESCALING  CONSTANTS
 TEST    NAME  POINTS    SCALE    LOCATION
 -------------------------------------------
    1  PRETEST    10     1.000     0.000
 -------------------------------------------

ITEM INFORMATION STATISTICS FOR TEST PRETEST  FORM    1
FOR A NORMAL POPULATION WITH MEAN =    0.000  AND S.D. =    1.000

            MAXIMUM        POINT OF         MAXIMUM          AVERAGE
         INFORMATION   MAX INFORMATION    EFFECTIVENESS      INFORMATION
ITEM       STANDARD        STANDARD         POINT OF          INDEX OF
            ERROR *         ERROR *     MAX EFFECTIVENESS *   RELIABILITY
-------------------------------------------------------------------------
MATH01|     0.2142   |     -1.3703   |         0.0587   |         0.1206
      |     0.0579*  |      0.2226*  |        -0.5284*  |         0.1076*

(Similar output omitted)

MATH15|     1.0608   |      2.5110   |         0.0375   |         0.0476
      |     0.8638*  |      0.3388*  |        -1.9279*  |         0.0454*
-------------------------------------------------------------------------
```

Using the **Plot** option on the **Run** menu to access the IRTGRAPH program, the following plot of test information is obtained:



6.2 Differential item functioning

This example is based on an example in Thissen, Steinberg & Wainer (1993). The data are drawn from a 100 word spelling test administered by tape recorder to psychology students at a large university. The words for the test were randomly selected from a popular word book for secretaries. Students were asked to write the words as used in a sentence on the tape recording. Responses were scored 1 if spelled correctly and 0 if spelled incorrectly. Because the items are scored 1,0, an answer key is not required. A complete description of these data are given in Section 2.14.1.

The groups in this example are the two sexes and this is indicated by the NGROUP keyword on the INPUT command. The same four items are presented to both groups on a single test form. The format statement following the second GROUP command describes the position and order of data in **exampl02.dat**. The group indicator is found in column 3 of the data records and is read in integer format. A form indicator is not required in the data records because there is only one form. The data have been sorted into answer patterns, and the frequencies are found in columns 10-11 of the data (2A1). These frequencies serve as case weights in the analysis. The TYPE=2 and NWGHT=3 keywords describe this type of data. The value assigned to the keyword NWGHT requests the use of weighting in both the statistics and calibration (by default, no weights would be applied).

A 1-parameter logistic model is requested using the NPARM keyword on the GLOBAL command. The LOGISTIC option on the GLOBAL command indicates that the natural metric of the logistic response function will be assumed in all calculations. If this keyword is not present, the logit is, by default, multiplied by 1.7 to obtain the metric of the normal response function.

The SAVE option on the GLOBAL command indicates that a SAVE command will follow directly after the GLOBAL command. On the SAVE command, the item parameter estimates are saved to an external file **exampl02.par** and the DIF analysis results are written to an external file **exampl02.dif**.

The total number of unique items is described using the NTOTAL keyword on the INPUT command while the NITEMS keyword on the LENGTH command is set to 4 to indicate that all 4 items are to be used in the single subtest.

The ITEMS command lists the four items in the order that they will be read from the data records. The INAMES and INUMBERS keywords assign each item a name and a corresponding number. Because there is only one form, the NFORM keyword is not required in the INPUT command and a FORM command is not required. Because examinees in both groups are presented all the items listed in the ITEMS command, the TEST and GROUP commands need contain only the test name and the group names, respectively.

A DIF analysis is requested through the use of the DIF option on the INPUT command.

The REFERENCE=1 keyword on the CALIB command designates males as the reference group. The convergence criterion is set to 0.005 instead of the default 0.01 using the CRIT keyword. When NGROUP >1, 20 quadrature points will be used for each group. Setting the NQPT keyword to 10 implies that 10 points will be used for each group, as fewer points are needed when the number of items is small.

No SCORE command is included in the command file, as DIF models cannot be scored.

```
EXAMPL02.BLM - MALE VS FEMALE DIFFERENTIAL ITEM FUNCTIONING
          SPELLING, GIRDER ITEM 4, OTHER 3 ITEMS 1-3
>GLOBAL   NPARM=1, LOGISTIC, SAVE, NWGHT=3, DFNAME='EXAMPL02.DAT';
>SAVE     PARM='EXAMPL02.PAR', DIF='EXAMPL02.DIF';
>LENGTH   NITEMS=4;
>INPUT    NTOTAL=4, NGROUPS=2, DIF, NIDCHAR=2, TYPE=2;
>ITEMS    INAMES=(SP1(1)SP4), INUMBERS=(1(1)4);
>TEST     TNAME=SPELL;
>GROUP1   GNAME=MALES;
>GROUP2   GNAME=FEMALES;
(2A1,I1,T10,F2.0,T5,4A1)
>CALIB    NQPT=10, CYCLES=15, CRIT=0.005, NEWTON=2, REFERENCE=1, PLOT=1;
```

Phase 1 output

The title and additional comments (if the optional COMMENT command has been used) are echoed to the output file. Immediately after that, Phase 1 commands and specifications of the analysis are given. Under FILE ASSIGNMENT, relevant information as read in from the GLOBAL, SAVE, LENGTH, and TEST command are listed.

```
EXAMPLE 02: MALE VS FEMALE DIFFERENTIAL ITEM FUNCTIONING
 SPELLING, GIRDER ITEM 4, OTHER 3 ITEMS 1-3
>GLOBAL  NPARM=1,LOGISTIC,SAVE,NWGHT=3, DFNAME='EXAMPL02.DAT';
```

```
FILE ASSIGNMENT AND DISPOSITION
===============================

SUBJECT DATA INPUT FILE     EXAMPL02.DAT
BILOG-MG MASTER DATA FILE   MF.DAT         WILL BE CREATED FROM DATA FILE
CALIBRATION DATA FILE       CF.DAT         WILL BE CREATED FROM DATA FILE
ITEM PARAMETERS FILE        IF.DAT         WILL BE CREATED THIS RUN
CASE SCALE-SCORE FILE       SF.DAT
CASE WEIGHTING                             FOR SUBJECT STATISTICS AND
                                           ITEM CALIBRATION
ITEM RESPONSE MODEL                        1 PARAMETER LOGISTIC
                                           LOGIT  METRIC (I.E., D = 1.0)

>SAVE    PARM='EXAMPL02.PAR',DIF='EXAMPL02.DIF';

BILOG-MG SAVE FILES

[OUTPUT FILES]
ITEM PARAMETERS FILE        EXAMPL02.PAR
DIF PARAMETER FILE          EXAMPL02.DIF

>LENGTH  NITEMS=4;



TEST LENGTH SPECIFICATIONS
==========================

MAIN TEST LENGTHS:          4

>INPUT   NTOTAL=4,NGROUP=2,DIF,NIDCHAR=2,TYPE=2;
```

Specification of INPUTrelated keywords are echoed in the next section. The data are entered as item-score patterns (right = 1, wrong = 0) and frequencies (case weights).

```
DATA INPUT SPECIFICATIONS
=========================

NUMBER OF FORMAT LINES                           1
NUMBER OF ITEMS IN INPUT STREAM                  4
NUMBER OF RESPONSE ALTERNATIVES               1000
NUMBER OF SUBJECT ID CHARACTERS                  2
NUMBER OF GROUPS                                 2
NUMBER OF TEST FORMS                             1
TYPE OF DATA                      SINGLE-SUBJECT DATA, CASE WEIGHTS
MAXIMUM SAMPLE SIZE FOR ITEM CALIBRATION  10000000
ALL SUBJECTS INCLUDED IN RUN

>ITEMS   INAMES=(SP1(1)SP4),INUMBERS=(1(1)4);

TEST SPECIFICATIONS
===================

>TEST    TNAME=SPELL;
```

The following lines indicate the assignment of items to the single subtest, utilizing the information on both the TEST and ITEMS commands.

```
TEST NUMBER:   1     TEST NAME: SPELL
NUMBER OF ITEMS:     4

  ITEM   ITEM         ITEM   ITEM         ITEM   ITEM         ITEM   ITEM
 NUMBER  NAME        NUMBER  NAME        NUMBER  NAME        NUMBER  NAME
 ----------------------------------------------------------------------
     1  SP1             2  SP2             3  SP3             4  SP4
 ----------------------------------------------------------------------
```

Information on the forms and groups is given next. The definition of the male and female groups, and the use of the same four items for both groups are reflected below. It is also noted that a DIF model is to be employed in this analysis.

```
FORM SPECIFICATIONS
===================
ITEMS READ ACCORDING TO SPECIFICATIONS ON THE ITEMS COMMAND

>GROUP1  GNAME=MALES;
>GROUP2  GNAME=FEMALES;

MULTIPLE GROUP SPECIFICATIONS
=============================
DIFFERENTIAL ITEM FUNCTIONING MODEL IS EMPLOYED.

GROUP NUMBER:  1     GROUP NAME: MALES
TEST NUMBER:   1     TEST NAME: SPELL
NUMBER OF ITEMS:     4

        ITEM   ITEM
       NUMBER  NAME
      ------------------
            1  SP1
            2  SP2
            3  SP3
            4  SP4
      ------------------

GROUP NUMBER:  2     GROUP NAME: FEMALES
TEST NUMBER:   1     TEST NAME: SPELL
NUMBER OF ITEMS:     4

        ITEM   ITEM
       NUMBER  NAME
      ------------------
            1  SP1
            2  SP2
            3  SP3
            4  SP4
      ------------------
```

Following is the format statement used in reading the data and the answer, omit, and not-present keys (if any). Data for this example are item scores and they are complete; keys are not required.

The case ID is read in the first 2 columns (2A1), followed by the group indicator (I1). After reading the weights (F2.0), the 4 item responses are read (4A1).

```
 FORMAT FOR DATA INPUT IS:
(2A1,I1,T10,F2.0,T5,4A1)
```

The first two cases are echoed to the output file so that the user can verify the input.

```
 OBSERVATION #     1  WEIGHT:    22.0000  ID :  1

 SUBTEST #:   1    SPELL
   GROUP #:   1    MALES

     TRIED    RIGHT
     4.000    0.000

 ITEM       1     2     3     4
 TRIED    1.0   1.0   1.0   1.0
 RIGHT    0.0   0.0   0.0   0.0



 OBSERVATION #     2  WEIGHT:    10.0000  ID :  2

 SUBTEST #:   1    SPELL
   GROUP #:   1    MALES

     TRIED    RIGHT
     4.000    1.000

 ITEM       1     2     3     4
 TRIED    1.0   1.0   1.0   1.0
 RIGHT    0.0   0.0   0.0   1.0
```

Classical item statistics for the total sample and each group sample follow. #TRIED designates the number of examinees responding to the item. For completeness, both the Pearson and biserial item-test correlations are shown. The latter has smaller bias when the percent right is extreme. The item statistics are given by group and then for the total group.

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

Pearson

The point biserial correlation $r_{PB,j}$ for item j is a computationally simplified Pearson's r between the dichotomously scored item j and the total score x. It is computed as

$$r_{PB,j} = \frac{(\mu_j - \mu_x)}{\sigma_x}\sqrt{\frac{p_j}{q_j}}$$

where μ_j is the mean total score among examinees who have responded correctly to item *j*, μ_x is the mean total score for all examinees, p_j is the item difficulty index for item *j*, q_j is $1-p_j$, and σ_x is the standard deviation of the total scores for all examinees.

Polyserial correlation

The polyserial correlation $r_{P,j}$ 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*, σ_j is the standard deviation of items scores *y* for item *j*, and $r_{PP,j}$ is the point-polyserial correlation.

The biserial correlation estimates the relationship between the total score and the hypothetical score on the continuous scale underlying the (dichotomous) item. The biserial correlation also assumes a normal distribution of the hypothetical scores. The reason for reporting these correlations separately for each group is that the appearance of large discrepancies between groups for a given item would suggest that the assumption of a common slope is untenable. Note that, if a biserial correlation more negative than –0.15 is detected by the program during this phase of the analysis, the item in question will be assumed miskeyed and will be omitted in the Phase 2 analysis.

```
ITEM STATISTICS FOR GROUP:   1    MALES
                                                  ITEM*TEST CORRELATION
ITEM   NAME      #TRIED     #RIGHT    PCT     LOGIT    PEARSON BISERIAL
--------------------------------------------------------------------------
   1   SP1         285.0      215.0   0.754    -1.12      0.243    0.332
   2   SP2         285.0      181.0   0.635    -0.55      0.351    0.450
   3   SP3         285.0       91.0   0.319     0.76      0.364    0.474
   4   SP4         285.0      179.0   0.628    -0.52      0.360    0.461
-------------------------------------------------------------------------
```

```
ITEM STATISTICS FOR GROUP:   2    FEMALES
                                                  ITEM*TEST CORRELATION
ITEM   NAME      #TRIED    #RIGHT    PCT     LOGIT    PEARSON  BISERIAL
--------------------------------------------------------------------------
    1   SP1         374.0      305.0   0.816    -1.49      0.254    0.370
    2   SP2         374.0      230.0   0.615    -0.47      0.295    0.376
    3   SP3         374.0      109.0   0.291     0.89      0.231    0.307
```

```
   4   SP4         374.0     171.0   0.457     0.17     0.306    0.385
----------------------------------------------------------------------

 ITEM STATISTICS FOR MULTIPLE GROUPS      SPELL
                                                  ITEM*TEST CORRELATION
ITEM   NAME     #TRIED    #RIGHT    PCT     LOGIT    PEARSON  BISERIAL
----------------------------------------------------------------------
   1   SP1         659.0     520.0   0.789    -1.32     0.239    0.337
   2   SP2         659.0     411.0   0.624    -0.51     0.320    0.409
   3   SP3         659.0     200.0   0.303     0.83     0.291    0.383
   4   SP4         659.0     350.0   0.531    -0.12     0.324    0.406
----------------------------------------------------------------------
```

Phase 2 output

During calibration, a logistic item response function is fitted to each item of each subscale. In this example, a 1-parameter logistic response function is fitted (NPARM =1 on GLOBAL).

Echoing of the Phase 2 commands and specification of the analysis starts the listing of Phase 2 output.

```
>CALIB    NQPT=10,CYCLES=15,CRIT=0.005,NEWTON=2,REFERENCE=1;
```

Under CALIBRATION PARAMETERS, the definitions of calibration related keywords for this analysis are given:

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

 MAXIMUM NUMBER OF EM CYCLES:                15
 MAXIMUM NUMBER OF NEWTON CYCLES:             2
 CONVERGENCE CRITERION:                      0.0050
 ACCELERATION CONSTANT:                      0.5000

 LATENT DISTRIBUTION:                      EMPIRICAL PRIOR FOR EACH GROUP
                                           ESTIMATED CONCURRENTLY
                                           WITH ITEM PARAMETERS
                                           REFERENCE GROUP:   1
PLOT EMPIRICAL VS. FITTED ICC's:           YES, FOR ITEMS WITH FIT PROBABILITY
                                           LESS THAN 1.00000
 DATA HANDLING:                            DATA ON SCRATCH FILE
 CONSTRAINT DISTRIBUTION ON SLOPES:        NO
 CONSTRAINT DISTRIBUTION ON THRESHOLDS:    NO
```

MML estimation is used when tests of three or more items are specified. The solution assumes that the respondents are drawn randomly from a population or populations of abilities, which is assumed to have a normal distribution. The empirical distribution of ability is represented as a discrete distribution on a finite number of points. The quadrature points and weights used for MML estimation of the item parameters for the two groups are printed next.

```
 METHOD OF SOLUTION:
 EM CYCLES (MAXIMUM OF    15)
 FOLLOWED BY NEWTON-RAPHSON STEPS (MAXIMUM OF   2)
```

```
 QUADRATURE POINTS AND PRIOR WEIGHTS: GROUP   1    MALES
               1           2           3           4           5
 POINT    -0.4000E+01 -0.3111E+01 -0.2222E+01 -0.1333E+01 -0.4444E+00
 WEIGHT    0.1190E-03  0.2805E-02  0.3002E-01  0.1458E+00  0.3213E+00

               6           7           8           9          10
 POINT     0.4444E+00  0.1333E+01  0.2222E+01  0.3111E+01  0.4000E+01
 WEIGHT    0.3213E+00  0.1458E+00  0.3002E-01  0.2805E-02  0.1190E-03

 QUADRATURE POINTS AND PRIOR WEIGHTS: GROUP   2    FEMALES
               1           2           3           4           5
 POINT    -0.4000E+01 -0.3111E+01 -0.2222E+01 -0.1333E+01 -0.4444E+00
 WEIGHT    0.1190E-03  0.2805E-02  0.3002E-01  0.1458E+00  0.3213E+00

               6           7           8           9          10
 POINT     0.4444E+00  0.1333E+01  0.2222E+01  0.3111E+01  0.4000E+01
 WEIGHT    0.3213E+00  0.1458E+00  0.3002E-01  0.2805E-02  0.1190E-03
```

The MML solution employs both the EM method and Newton-Gauss iterations to solve the marginal likelihood equations. On the CALIB command, a maximum of 15 EM cycles and 2 Newton-Gauss iterations were requested. Results for each iteration are displayed so that the extent of convergence can be judged.

In the case of nested models on the same data, the –2 log likelihood values at convergence can be used to evaluate the fit of the models. Refitting this example, for instance as a single-group analysis will allow the comparison of non-DIF and DIF models for these data. In that way, it can be determined whether differential item functioning effects are present.

```
 [E-M CYCLES]

 -2 LOG LIKELIHOOD =       3152.375

 CYCLE      1;   LARGEST CHANGE=   0.17572
 -2 LOG LIKELIHOOD =       3128.806

...

 CYCLE      8;   LARGEST CHANGE=   0.00486
```

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 on the item parameter estimates.

```
 [FULL NEWTON CYCLES]
 -2 LOG LIKELIHOOD:         3110.3990
 CYCLE     9;   LARGEST CHANGE=   0.00416
```

In Phase 2, when there is a single group, the unit and origin of the scale on which the parameters are expressed is based on the assumption that the latent ability distribution has zero mean and unit variance (the so-called “0,1” metric). In the case of multiple groups, the program provides the option of setting the mean and standard deviation of one group to 0,1 as shown here. The user may set the mean and standard deviation of the combined estimated distribution of the groups to

0 and 1 by setting the REFERENCE keyword on the CALIB command to zero. The parameter estimates can be rescaled in Phase 3 according to scale conventions selected by the user (using the RSCTYPE, SCALE and LOCATION keywords on the SCORE command). In a DIF model, no scoring is done, so use of the REFERENCE=0 specification is not pursued here. Estimated item parameters for the two groups are given next. The INTERCEPT column contains the estimates of the item intercepts, which are defined as the product of each item's slope and threshold. This is followed by the slope or discrimination parameters and the item threshold or location parameters. The LOADING column represents the one-factor item factor loadings given by the expression

$$\frac{slope}{\sqrt{1.0+slope^2}}.$$

For a 1PL model, no asymptotes or guessing parameters are estimated. In a 1PL model, all slopes are equal. In DIF analyses, the assumption is made that slopes are equal over the groups. This implies that items will discriminate equally well in all groups. Note that, in this example, the slopes of all items for both groups are constrained to 1.285.

```
GROUP   1    MALES   ;  ITEM PARAMETERS AFTER CYCLE   9

 ITEM      INTERCEPT    SLOPE    THRESHOLD   LOADING   ASYMPTOTE    CHISQ     DF
             S.E.       S.E.       S.E.       S.E.       S.E.      (PROB)
 -------------------------------------------------------------------------------
 SP1       |   1.489  |   1.285  |  -1.159  |   0.789  |   0.000  |     2.7   7.0
           |   0.168* |   0.096* |   0.130* |   0.059* |   0.000* | (0.9146)
           |          |          |          |          |          |
 SP2       |   0.749  |   1.285  |  -0.583  |   0.789  |   0.000  |    20.0   7.0
           |   0.152* |   0.096* |   0.119* |   0.059* |   0.000* | (0.0056)
           |          |          |          |          |          |
 SP3       |  -1.008  |   1.285  |   0.784  |   0.789  |   0.000  |    36.2   5.0
           |   0.109* |   0.096* |   0.085* |   0.059* |   0.000* | (0.0000)
           |          |          |          |          |          |
 SP4       |   0.709  |   1.285  |  -0.552  |   0.789  |   0.000  |    26.6   7.0
           |   0.150* |   0.096* |   0.117* |   0.059* |   0.000* | (0.0004)
 -------------------------------------------------------------------------------
                                                           * STANDARD ERROR
  LARGEST CHANGE =    0.004158                          95.3  12.0
                                                          (0.0000)
GROUP   2    FEMALES ;  ITEM PARAMETERS AFTER CYCLE   9

 ITEM      INTERCEPT    SLOPE    THRESHOLD   LOADING   ASYMPTOTE    CHISQ     DF
              S.E.       S.E.       S.E.       S.E.       S.E.      (PROB)
 -------------------------------------------------------------------------------
 SP1       |   1.887  |   1.285  |  -1.468  |   0.789  |   0.000  |    11.3   7.0
           |   0.168* |   0.096* |   0.131* |   0.059* |   0.000* | (0.1261)
           |          |          |          |          |          |
 SP2       |   0.617  |   1.285  |  -0.480  |   0.789  |   0.000  |    34.6   7.0
           |   0.136* |   0.096* |   0.106* |   0.059* |   0.000* | (0.0000)
           |          |          |          |          |          |
 SP3       |  -1.113  |   1.285  |   0.866  |   0.789  |   0.000  |    24.5   7.0
           |   0.144* |   0.096* |   0.112* |   0.059* |   0.000* | (0.0009)
           |          |          |          |          |          |
 SP4       |  -0.203  |   1.285  |   0.158  |   0.789  |   0.000  |    43.0   7.0
           |   0.133* |   0.096* |   0.104* |   0.059* |   0.000* | (0.0000)
 -------------------------------------------------------------------------------
                                                           * STANDARD ERROR
```

```
    LARGEST CHANGE =    0.004158                              95.3  12.0
                                                            (0.0000)
NOTE: ITEM FIT CHI-SQUARES AND THEIR SUMS MAY BE UNRELIABLE
FOR TESTS WITH LESS THAN 20 ITEMS
```

The item parameter estimates for each group are followed by the averages for the group thresholds. The mean threshold of the female group (Group 2) is 0.146 above that of the male or reference group. DIF is item by group interaction under the constraint that the mean thresholds of the groups are equal. The threshold adjustment sets the mean of the reference group's threshold to 1, and the mean threshold for the females is accordingly adjusted to 0.148. The unadjusted and adjusted mean thresholds for the two groups form the next section of the Phase 2 output file.

```
PARAMETER       MEAN  STN DEV
-----------------------------------
GROUP:   1  NUMBER OF ITEMS:     4
THRESHOLD     -0.377    0.823
GROUP:   2  NUMBER OF ITEMS:     4
THRESHOLD     -0.231    0.991
-----------------------------------

THRESHOLD MEANS
GROUP       ADJUSTMENT
------------------------
    1         0.000
    2         0.146
------------------------

MODEL FOR GROUP DIFFERENTIAL ITEM FUNCTIONING:
ADJUSTED THRESHOLD VALUES
  ITEM             GROUP
                1          2
----------------------------------
   SP1      |  -1.159  |  -1.614
            |   0.130* |   0.131*
            |          |
   SP2      |  -0.583  |  -0.626
            |   0.119* |   0.106*
            |          |
   SP3      |   0.784  |   0.720
            |   0.085* |   0.112*
            |          |
   SP4      |  -0.552  |   0.012
            |   0.117* |   0.104*
----------------------------------
*STANDARD ERROR
```

The adjusted threshold values are followed by the group differences of the constrained values. The standard errors for the differences are computed as

$$s.e._{G2-G1} = \sqrt{\mathrm{var}(G2) + \mathrm{var}(G1)}.$$

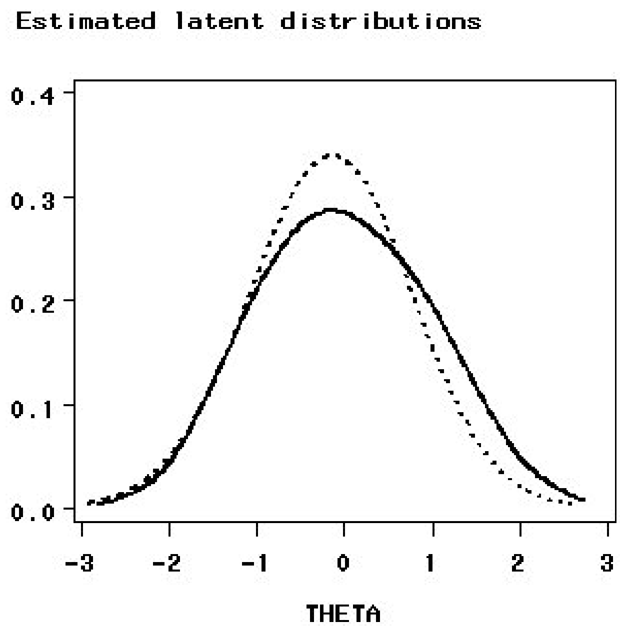
```
   ITEM        GROUP
                2 - 1
-----------------------
   SP1      |  -0.455
            |   0.185*
            |
```

```
  SP2      |  -0.043
           |   0.159*
           |
  SP3      |  -0.065
           |   0.141*
           |
  SP4      |   0.564
           |   0.156*
----------------------
 *STANDARD ERROR
```

The estimated latent distributions of the groups are given next; with the origin and unit of scale set so that the mean of the reference group is 0 and the standard deviation is 1.

```
 GROUP:  1    MALES     QUADRATURE POINTS, POSTERIOR WEIGHTS, MEAN AND S.D.:

                  1           2           3           4           5
 POINT       -0.3578E+01 -0.2788E+01 -0.1998E+01 -0.1208E+01 -0.4180E+00
 POSTERIOR    0.1972E-03  0.4485E-02  0.4394E-01  0.1737E+00  0.2780E+00

                  6           7           8           9          10
 POINT        0.3720E+00  0.1162E+01  0.1952E+01  0.2742E+01  0.3532E+01
 POSTERIOR    0.2647E+00  0.1724E+00  0.5526E-01  0.7020E-02  0.3483E-03

 MEAN         0.00000
 S.E.         0.00000

 S.D.         1.00000
 S.E.         0.00000



 GROUP:  2    FEMALES   QUADRATURE POINTS, POSTERIOR WEIGHTS, MEAN AND S.D.:

                  1           2           3           4           5
 POINT       -0.3724E+01 -0.2934E+01 -0.2144E+01 -0.1354E+01 -0.5642E+00
 POSTERIOR    0.2099E-03  0.4246E-02  0.3608E-01  0.1456E+00  0.3067E+00

                  6           7           8           9          10
 POINT        0.2258E+00  0.1016E+01  0.1806E+01  0.2596E+01  0.3386E+01
 POSTERIOR    0.3161E+00  0.1525E+00  0.3473E-01  0.3598E-02  0.1624E-03

 MEAN        -0.16191
 S.E.         0.06907

 S.D.         0.89707
 S.E.         0.00845
```

A plot of the two estimated latent distributions are shown below. The solid line represents the estimated distribution of the male group.



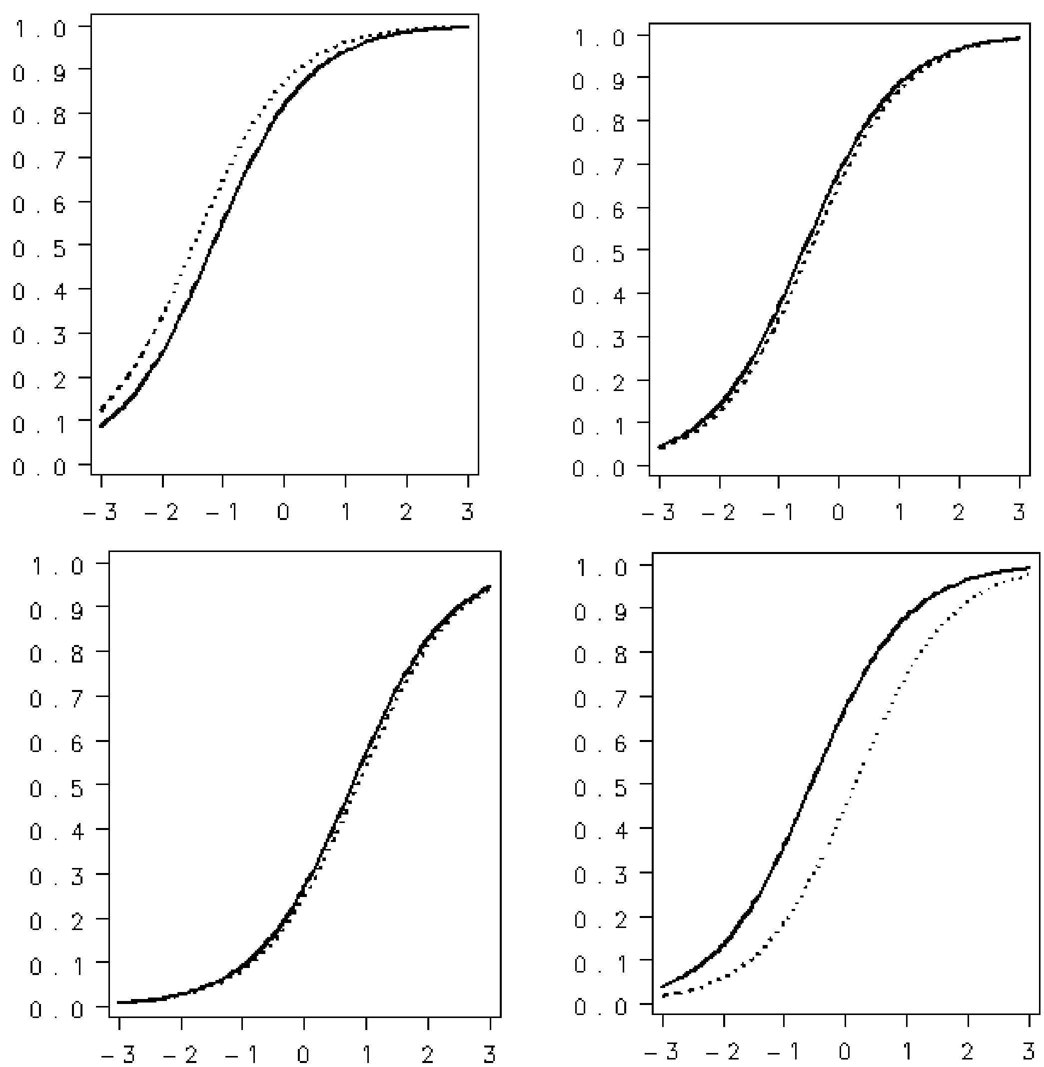
BILOG-MG is also capable of producing graphic representations of a number of item and test characteristics. Using the PLOT keyword on the CALIB command, it is possible to obtain plots of the item-response functions with a significance level below the value assigned to the PLOT keyword. By default, PLOT=0 and no plots are produced. On the other hand, setting PLOT to 1.0 will lead to the display of all item response functions in the output file. One such plot, for the fourth item administered to the female group, is shown below.

The plot also shows 95% tolerance intervals for the observed percent correct among respondents in corresponding EAP groups, assuming the percent-correct predicted by the model is correct. Note that similar plots may be obtained using the IRTGRAPH program accessible from the **Run** menu.

```
GROUP:   2   FEMALES
SUBTEST:  SPELL

ITEM:  SP4         CHISQ =    43.0  DF =   7.0  PROB< 0.0000
      1.00+-------------------------------------------------------------+
          |                                                             |
          |                                                             |
      0.90|                                                             |
          |                                                          ...|
          |                                                       ...   |
      0.80|                                                    ...      |
          |                                           X      ..         |
          |                                           |    ..           |
      0.70|                                           |  ..             |
          |                                           |..               |
          |                                          .|                 |
      0.60|                                        .. |                 |
          |                                      ..   |                 |
          |                                     .                       |
      0.50|                             |    ..                         |
          |                             |  ..                           |
```

```
         |                             |..                            |
     0.40|                             |                              |
         |                           ..|                              |
         |                         ..  |                              |
     0.30|                |      ..    X                              |
         |                |   ...                                     |
         |                | ..                                        |
     0.20|               .|.                                          |
         |            ... |                                           |
         |        ....    |                                           |
     0.10|  ......        X                                           |
         |..                                                          |
         |                                                            |
     0.00|                                                            |
         +--+-----+-----+-----+-----+-----+-----+-----+-----+-----+---+
THETA    -1.86 -1.48 -1.11 -0.73 -0.36  0.02  0.40  0.77  1.15  1.53
```

By saving the estimated parameter estimates to an external file, the estimates can also be used in external packages to produce additional plots. Below, the item response functions for both groups are plotted by item.



6.3 Differential item functioning

The data from example 2 are analyzed here as a single group. Thus no NGROUP keyword is provided on the INPUT command and, by default, the program assumes there is only one group. No GROUP commands follow the TEST command, and the group indicator has been removed from the variable format statement.

The acceleration factor on the CALIB command has been set to its default value of 0.5 (ACCEL=0.5). The difference in the log likelihoods from the two-group and single-group solutions can be examined to determine if differential item functioning effects are present. The item parameter file obtained in the previous section is specified in the GLOBAL command to provide starting values for parameter estimation in Phase 2.

```
EXAMPL03.BLM - MALE VS. FEMALE DIFFERENTIAL ITEM FUNCTIONING
               SPELLING, GIRDER ITEM 4, OTHER 3 ITEMS 1-3
>GLOBAL  NPARM=1, NWGHT=3, LOGISTIC, IFNAME='EXAMPL02.PAR',
         DFNAME='EXAMPL02.DAT';
>LENGTH  NITEMS=4;
>INPUT   NTOTAL=4, NIDCHAR=2, TYPE=2;
>ITEMS   INAME=(SP01,SP02,SP03,SP04), INUMBERS=(1(1)4);
>TEST    TNAME=SPELL;
(2A1,T10,F2.0,T5,4A1)
>CALIB   EMPIRICAL, NQPT=31, CRIT=0.005, ACCEL=0.5;
```

Phase 1 output

```
EXAMPLE 02: MALE VS. FEMALE DIFFERENTIAL ITEM FUNCTIONING
SPELLING, GIRDER ITEM 4, OTHER 3 ITEMS 1-3
```

The Phase 1 output for this example is the same as that obtained in Section 6.2, except that classical item statistics are computed only for the total sample.

Phase 2 output

The main interest in this example is the comparison of the log likelihood of the fit of the DIF and non-DIF models. The difference, 3138.4122 – 3110.3990 = 28.0132, distributed as χ^2 on four degrees of freedom, indicates significantly better fit of the DIF model.

```
-2 LOG LIKELIHOOD:          3138.4122
 CYCLE     4;   LARGEST CHANGE=   0.00439

  SUBTEST SPELL   ;  ITEM PARAMETERS AFTER CYCLE   4

ITEM      INTERCEPT    SLOPE    THRESHOLD   LOADING   ASYMPTOTE    CHISQ     DF
              S.E.       S.E.       S.E.       S.E.       S.E.      (PROB)
 ------------------------------------------------------------------------------
 SP01     |   1.688  |   1.234  |  -1.368  |   0.777  |   0.000  |    15.4   3.0
          |   0.124* |   0.094* |   0.101* |   0.059* |   0.000* | (0.0015)
```

```
         |           |           |           |           |           |
SP02     |   0.662   |   1.234   |  -0.536   |   0.777   |   0.000   |    33.6   3.0
         |   0.105*  |   0.094*  |   0.085*  |   0.059*  |   0.000*  | (0.0000)
         |           |           |           |           |           |
SP03     |  -1.069   |   1.234   |   0.866   |   0.777   |   0.000   |    17.9   3.0
         |   0.111*  |   0.094*  |   0.090*  |   0.059*  |   0.000*  | (0.0005)
         |           |           |           |           |           |
SP04     |   0.169   |   1.234   |  -0.137   |   0.777   |   0.000   |    32.2   3.0
         |   0.102*  |   0.094*  |   0.082*  |   0.059*  |   0.000*  | (0.0000)
--------------------------------------------------------------------------------
                                                            * STANDARD ERROR

    LARGEST CHANGE =    0.004551                              99.0  12.0
                                                            (0.0000)

NOTE: ITEM FIT CHI-SQUARES AND THEIR SUMS MAY BE UNRELIABLE
FOR TESTS WITH LESS THAN 20 ITEMS
```

6.4 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)
          |          |          |          |          |
T39       |  -4.691  |   0.021  | 225.362  |   0.673  |   0.000
          |   0.988* |   0.004* |   5.524* |   0.127* |   0.000*
          |          |          |          |          |
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
  STANDARD                                                              INFOR-
  ERROR                                                                 MATION
------------------------------------------------------------------
1.48|   *                         +++++                            * | .1341
    |                            +     ++                            |
1.41|                          ++        +                           | 6.7774
    |    *                                                        *  |
1.33|                         +           +                          | 6.4207
    |                        +             +                         |
1.26|     *                 +                                    *   | 6.0640
    |                                       +                        |
1.19|                      +                                    *    | 5.7073
    |      *                                 +                       |
1.11|                     +                                    *     | 5.3506
    |       *                                 +                      |
1.04|                    +                                    *      | 4.9939
    |        *                                 +                     |
0.96|                                                        *       | 4.6372
    |         *         +                       +           *        |
0.89|                                                                | 4.2805
    |          *       +                         +         *         |
0.82|                                                     *          | 3.9237
    |           *     +                           +                  |
0.74|                                                    *           | 3.5670
    |            *                                 +    *            |
0.67|             *  +                                 *             | 3.2103
    |              *                                + *              |
0.59|               +                                +               | 2.8536
    |                *                             **                |
0.52|              +  *                           *   +              | 2.4969
    |                  **                       **     +             |
0.44|             +      **                   **                     | 2.1402
    |            +         ****           ****          +            |
0.37|                          ***********               +           | 1.7835
    |           +                                         +          |
0.30|          +                                           +         | 1.4268
    |         +                                             +        |
0.22|        +                                               +       | 1.0701
    |       +                                                 ++     |
0.15|     ++                                                    +    | 0.7134
    |   ++                                                       +++ |
0.07|+++                                                            +| 0.3567
    |                                                                |
0.00|                                                                | 0.0000
    -+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
   -4.00   -3.00   -2.00   -1.00    0.00    1.00    2.00    3.00    4.00
```

6.5 Vertical equating

Two hundred students at each of three grade levels, grades four, six, and eight, were given grade-appropriate versions of a 20-item arithmetic examination. Items 19 and 20 appear in the grade 4 and 6 forms; items 37 and 38 appear in the grade 6 and 8 forms. Because each item is assigned a unique column in the data records, a FORM command is not required.

The data file contains, the answer key, not-presented key and raw data. Two lines of information are given per examinee as shown below. The answer key contains 56 entries, each equal to 1. If an item has not been presented, its presence in the data will be indicated with ‘ ‘.

```
KEY   11111111111111111111111111111111111111111111111111111111
NOT
001 1 11111112221211222212
002 1 21121211121111121212
003 1 11112112211222212212
004 1 11111112121111111211
005 1 21111112221212121222
```

No items are assigned to the TEST command using the INAMES or INUMBERS keywords. By default, it is assumed that all items are assigned to the test. Although the test name (TNAME=MATH) is not enclosed in single quotes, the group names are as these names contain blanks as part of the name.

The distributions of ability are assumed to be normal at each grade level (NORMAL on the CALIB command). Grade 6 serves as the reference group in the calibration of the items (REFERENCE=2). EAP estimates of ability are calculated using the information in the posterior distributions from Phase 2. The ability estimates are rescaled to a mean of 0 and standard deviation of 1 by specifying RSCTYPE=3 on the SCORE command.

```
EXAMPL05.BLM - VERTICAL EQUATING OF TEST FORMS OVER THREE GRADE LEVELS

 >GLOBAL  DFNAME='EXAMPL05.DAT', NPARM=2, SAVE;
 >SAVE    SCORE='EXAMPL05.SCO', PARM='EXAMPL05.PAR';
 >LENGTH  NITEMS=56;
 >INPUT   NTOT=56, NGROUPS=3, NIDCH=3,
          KFNAME='EXAMPL05.DAT', NFNAME='EXAMPL05.DAT';
 >ITEMS   INUM=(1(1)56), INAME=(M01(1)M56);
 >TEST    TNAME=MATH;
 >GROUP1  GNAME='GRADE 4', LENGTH=20, INUM=(1(1)20);
 >GROUP2  GNAME='GRADE 6', LENGTH=20, INUM=(19(1)38);
 >GROUP3  GNAME='GRADE 8', LENGTH=20, INUM=(37(1)56);
 (3A1,1X,I1,1X,56A1)
 >CALIB   NQPT=51, NORMAL, CYCLE=30, TPRIOR, REFERENCE=2;
 >SCORE   METHOD=2, IDIST=3, NOPRINT, RSCTYPE=3;
```

Phase 1 output

In this example, items assigned to the three groups of examinees are selected from the following set. The items are selected in such a way that two items are common to groups 1 and 2 and two other items are common to groups 2 and 3. The groups, corresponding to school grades four, six, and eight are non-equivalent and require separate classical item statistics. The fact that classical item statistics are not invariant with respect to sampling from different populations is illustrated by the different results for common items in different groups.

```
 MULTIPLE GROUP SPECIFICATIONS
 =============================


 MULTIPLE GROUPS ARE DEFINED,
 BUT NEITHER DIF MODEL NOR PARAMETER DRIFT MODEL IS EMPLOYED.
```

```
 GROUP NUMBER:  1    GROUP NAME: GRADE 4
 TEST NUMBER:   1    TEST NAME: MATH
 NUMBER OF ITEMS:   20

         ITEM   ITEM
        NUMBER  NAME
       ------------------
             1  M01
             2  M02
             …
            20  M20
       ------------------

 GROUP NUMBER:  2    GROUP NAME: GRADE 6
 TEST NUMBER:   1    TEST NAME: MATH
 NUMBER OF ITEMS:   20

         ITEM   ITEM
        NUMBER  NAME
       ------------------
            19  M19
            20  M20
            …
            38  M38
       ------------------

 GROUP NUMBER:  3    GROUP NAME: GRADE 8
 TEST NUMBER:   1    TEST NAME: MATH
 NUMBER OF ITEMS:   20

         ITEM   ITEM
        NUMBER  NAME
       ------------------
            37  M37
             …
            56  M56
       ------------------

       600 OBSERVATIONS READ FROM FILE:   EXAMPL05.DAT
       600 OBSERVATIONS WRITTEN TO FILE:  MF.DAT

           SUBTEST   1    MATH
             GROUP   1    GRADE 4     200 OBSERVATIONS
             GROUP   2    GRADE 6     200 OBSERVATIONS
             GROUP   3    GRADE 8     200 OBSERVATIONS
```

Item statistics for the first 5 items of each subtest are shown below. Similar output is produced for grades 6 to 8, and for multiple groups MATH which, in this case, contains the statistics for all the grades.

```
SUBTEST    1    MATH
 ITEM STATISTICS FOR GROUP:  1    GRADE 4
                                                    ITEM*TEST CORRELATION
ITEM   NAME     #TRIED    #RIGHT    PCT     LOGIT/1.7  PEARSON  BISERIAL
 ------------------------------------------------------------------------
    1   M01      200.0     138.0   0.690    -0.47      0.470    0.616
...
   19   M19      200.0      95.0   0.475     0.06      0.520    0.652
```

```
  20   M20      200.0      67.0   0.335     0.40     0.475    0.615
 ----------------------------------------------------------------------

 ITEM STATISTICS FOR GROUP:  2    GRADE 6
                                                ITEM*TEST CORRELATION
 ITEM   NAME     #TRIED    #RIGHT    PCT     LOGIT/1.7  PEARSON  BISERIAL
 ----------------------------------------------------------------------
   19   M19         200.0     138.0   0.690    -0.47      0.431    0.565
   20   M20         200.0     106.0   0.530    -0.07      0.512    0.643
...
   37   M37         200.0     104.0   0.520    -0.05      0.379    0.475
   38   M38         200.0      62.0   0.310     0.47      0.497    0.651
 ------------------------------------------------------------------------

 ITEM STATISTICS FOR GROUP:  3    GRADE 8
                                                ITEM*TEST CORRELATION
ITEM   NAME      #TRIED    #RIGHT    PCT     LOGIT/1.7  PEARSON  BISERIAL
 ------------------------------------------------------------------------
   37   M37         200.0     135.0   0.675    -0.43      0.420    0.546
   38   M38         200.0      96.0   0.480     0.05      0.594    0.745
...
   55   M55         200.0      90.0   0.450     0.12      0.471    0.592
   56   M56         200.0     111.0   0.555    -0.13      0.529    0.665
 ------------------------------------------------------------------------
```

Phase 2 output

In vertical equating over a range of age levels, the ability distributions of the groups may be widely spaced. For that reason, it is desirable to use a large number of quadrature points – in this case, 51.

The origins and unit of the ability distribution can be fixed in the calibration either by setting the mean and standard deviation of a reference group to zero and one, respectively, or, similarly, setting the mean and standard deviation of the combined groups. In this example, group 2 is selected as the reference group.

```
>CALIB   NQPT=51, NORMAL, CYCLE=30, TPRIOR, REFERENCE=2;

 CALIBRATION PARAMETERS
 ======================
 MAXIMUM NUMBER OF EM CYCLES:               30
 MAXIMUM NUMBER OF NEWTON CYCLES:            2
 CONVERGENCE CRITERION:                     0.0100
 ACCELERATION CONSTANT:                     1.0000
 LATENT DISTRIBUTION:                     NORMAL PRIOR FOR EACH GROUP
                                          GROUP MEANS AND SDS
                                          ESTIMATED CONCURRENTLY
                                          WITH ITEM PARAMETERS
                                          REFERENCE GROUP:   2
 PLOT EMPIRICAL VS. FITTED ICC'S:         NO
 DATA HANDLING:                           DATA ON SCRATCH FILE
…
       MEANS AND STANDARD DEVIATIONS:     PROGRAM DEFAULTS
                                          ITEM CONSTRAINTS IF PRESENT
                                            WILL BE UPDATED EACH CYCLE
```

The iterative estimation procedures typically converge more slowly in nonequivalent group data than in one group or equivalent groups data. The last few iterations are shown here along with some of the resulting parameter estimates. The means of the prior distributions on item thresholds and slopes are also listed.

```
CYCLE     19;   LARGEST CHANGE=   0.02538
 -2 LOG LIKELIHOOD =      13246.111
 UPDATED PRIOR ON LOG SLOPES; MEAN & SD =       -0.23806     0.50000
 UPDATED PRIOR ON THRESHOLDS; MEAN & SD =        0.08303     2.00000

 CYCLE     20;   LARGEST CHANGE=   0.00812

 [NEWTON CYCLES]
UPDATED PRIOR ON LOG SLOPES; MEAN & SD =       -0.23533     0.50000
UPDATED PRIOR ON THRESHOLDS; MEAN & SD =        0.08308     2.00000
 -2 LOG LIKELIHOOD:        13245.9542

 CYCLE    21;   LARGEST CHANGE=   0.00699

 INTERVAL COUNTS FOR COMPUTATION OF ITEM CHI-SQUARES
-------------------------------------------------------------------------
    19.    32.    56.    83.    93.   109.    82.    60.    66.
-------------------------------------------------------------------------

 INTERVAL AVERAGE THETAS
-------------------------------------------------------------------------
   -2.695 -1.942 -1.448 -0.866 -0.356  0.145  0.607  1.193  1.989
-------------------------------------------------------------------------

 SUBTEST MATH    ;  ITEM PARAMETERS AFTER CYCLE  21
ITEM  INTERCEPT    SLOPE    THRESHOLD   LOADING   ASYMPTOTE    CHISQ  DF
            S.E.       S.E.       S.E.       S.E.       S.E.      (PROB)
-------------------------------------------------------------------------
M01 |   1.218  |   0.805  |  -1.512  |   0.627  |   0.000  |     3.2  5.0
    |   0.194* |   0.155* |   0.163* |   0.121* |   0.000* | (0.6741)
    |          |          |          |          |          |
M02 |   1.149  |   0.707  |  -1.626  |   0.577  |   0.000  |     4.4  6.0
    |   0.169* |   0.129* |   0.186* |   0.105* |   0.000* | (0.6249)
    |          |          |          |          |          |
[Similar output omitted]
M55 |  -0.584  |   0.707  |   0.826  |   0.577  |   0.000  |     3.9  6.0
    |   0.129* |   0.126* |   0.143* |   0.103* |   0.000* | (0.6847)
    |          |          |          |          |          |
M56 |  -0.319  |   0.849  |   0.376  |   0.647  |   0.000  |     1.1  5.0
    |   0.127* |   0.144* |   0.125* |   0.110* |   0.000* | (0.9547)
-------------------------------------------------------------------------
                                                        * STANDARD ERROR

 LARGEST CHANGE =    0.007897                          188.0 296.0
                                                          (1.0000)
 PARAMETER       MEAN  STN DEV
 -----------------------------------
 SLOPE          0.802    0.138
 LOG(SLOPE)    -0.235    0.172
 THRESHOLD      0.083    0.775
```

The within-group latent distributions are assumed normal. Their means and standard deviations are estimated relative to the reference group. In these data, the means increase over the grades (-0.722, 0.000, 0.569), but the standard deviations are relatively constant (1.069, 1.00, 1.126).

```
 GROUP:  1    GRADE 4   QUADRATURE POINTS, POSTERIOR WEIGHTS, MEAN AND S.D.:
                 1           2           3           4           5
 POINT       -0.4275E+01 -0.4105E+01 -0.3935E+01 -0.3765E+01 -0.3594E+01
 POSTERIOR    0.4299E-03  0.7062E-03  0.1119E-02  0.1717E-01  0.2558E-02

[Similar output omitted]

                47          48          49          50          51
 POINT        0.3552E+01  0.3722E+01  0.3892E+01  0.4062E+01  0.4232E+01
 POSTERIOR    0.1899E-04  0.9879E-05  0.3535E-05  0.1816E-05  0.9055E-06

 MEAN        -0.72298
 S.E.         0.11260

 S.D.         1.06880
 S.E.         0.12631

GROUP:  2    GRADE 6   QUADRATURE POINTS, POSTERIOR WEIGHTS, MEAN AND S.D.:
                 1           2           3           4           5
 POINT       -0.4275E+01 -0.4105E+01 -0.3935E+01 -0.3765E+01 -0.3594E+01
 POSTERIOR    0.1136E-04  0.2278E-04  0.4596E-04  0.8712E-04  0.1599E-03

[Similar output omitted]

                47          48          49          50          51
 POINT        0.3552E+01  0.3722E+01  0.3892E+01  0.4062E+01  0.4232E+01
 POSTERIOR    0.1172E-03  0.6346E-04  0.3291E-04  0.1689E-04  0.8409E-05

 MEAN         0.00000
 S.E.         0.00000

 S.D.         1.00000
 S.E.         0.00000

 GROUP:  3    GRADE 8   QUADRATURE POINTS, POSTERIOR WEIGHTS, MEAN AND S.D.:
                 1           2           3           4           5
 POINT       -0.4275E+01 -0.4105E+01 -0.3935E+01 -0.3765E+01 -0.3594E+01
 POSTERIOR    0.4219E-05  0.7809E-05  0.1793E-04  0.3292E-04  0.5918E-04

[Similar output omitted]

                47          48          49          50          51
 POINT        0.3552E+01  0.3722E+01  0.3892E+01  0.4062E+01  0.4232E+01
 POSTERIOR    0.1837E-02  0.1230E-02  0.8192E-03  0.5316E-03  0.3268E-03

 MEAN         0.56861
 S.E.         0.11855

 S.D.         1.12577
 S.E.         0.14026
```

Phase 3 output

With nonequivalent groups, Bayes (EAP) and Bayes Modal (MAP) estimation of test scores should be carried out with respect to the Phase 2 latent distribution to which the examinee belongs. Specify IDIST=3 on the SCORE command.

```
>SCORE   METHOD=2,IDIST=3,NOPRINT, RSCTYPE=3;

PARAMETERS FOR SCORING, RESCALING, AND TEST AND ITEM INFORMATION
METHOD OF SCORING SUBJECTS:                EXPECTATION A POSTERIORI
                                           (EAP; BAYES ESTIMATION)
TYPE OF PRIOR:                             EMPIRICAL, FROM ITEM CALIBRATION
TYPE OF RESCALING:                         IN THE SAMPLE DISTRIBUTION
REFERENCE GROUP FOR RESCALING:             GROUP:  2


                      QUAD
TEST    NAME   GROUP POINTS
---------------------------
   1  MATH       1    51
   1  MATH       2    51
   1  MATH       3    51
---------------------------

                 RESCALING CONSTANTS
TEST    NAME       SCALE   LOCATION
------------------------------------
   1  MATH         1.000     0.000
------------------------------------
```

In this example, the scores are rescaled so that their mean and standard deviation in the total sample are zero and one, respectively. The parameter estimates are rescaled accordingly.

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

          RESCALING  CONSTANTS
 TEST       SCALE    LOCATION
 MATH         1.066     0.003

GROUP   SUBJECT IDENTIFICATION                                  MARGINAL
WEIGHT   TEST       TRIED  RIGHT  PERCENT     ABILITY      S.E.     PROB
-------------------------------------------------------------------------
 1      1                                |                      |
 1.00   MATH         20     11    55.00 |   -0.3055    0.3598 |  0.000000
 1      1                                |                      |
 1.00   MATH         20     13    65.00 |   -0.0653    0.3620 |  0.000000
-------------------------------------------------------------------------
```

```
TEST    MATH    ;  RESCALED ITEM PARAMETERS

ITEM       INTERCEPT   SLOPE    THRESHOLD   LOADING    ASYMPTOTE
              S.E.      S.E.       S.E.       S.E.        S.E.
---------------------------------------------------------------
M01       |   1.216  |   0.755  |  -1.610  |   0.627  |   0.000
          |   0.194* |   0.145* |   0.173* |   0.121* |   0.000*
          |          |          |          |          |
M02       |   1.148  |   0.663  |  -1.732  |   0.577  |   0.000
          |   0.169* |   0.121* |   0.198* |   0.105* |   0.000*
          |          |          |          |          |
[Similar output omitted]
          |          |          |          |          |
M55       |  -0.566  |   0.670  |   0.845  |   0.581  |   0.000
          |   0.127* |   0.120* |   0.151* |   0.104* |   0.000*
          |          |          |          |          |
M56       |  -0.298  |   0.805  |   0.370  |   0.651  |   0.000
          |   0.125* |   0.136* |   0.132* |   0.110* |   0.000*
---------------------------------------------------------------

PARAMETER       MEAN  STN DEV
-----------------------------
SLOPE          0.752    0.130
LOG(SLOPE)    -0.299    0.172
THRESHOLD      0.092    0.806

MEAN & SD OF SCORE ESTIMATES AFTER RESCALING

GROUP       MEAN       SD
-----------------------------
   1      -0.776     1.067
   2       0.000     1.000
   3       0.608     1.118
-----------------------------

MEAN & SD OF LATENT DISTRIBUTIONS AFTER RESCALING

GROUP       MEAN       SD
-----------------------------
   1      -0.776     1.149
   2       0.000     1.074
   3       0.608     1.201
-----------------------------
```

6.6 Multiple matrix sampling data

This example illustrates the use of the TYPE=3 specification on the INPUT command to analyze aggregate-level, multiple-matrix sampling data. The data in **exampl06.dat** are numbers tried and numbers correct for items from eight forms of a matrix sampled assessment instrument. The groups are selected 8th grade students from 32 public schools. The first record for each school contains the data for the items of a Number Concepts scale, NUMCON, and the second record contains the data for items of an Algebra Concepts scale, ALGCON. Data for the first two schools are shown below.

```
SCHOOL  1 NUM   1  0  3  2  2  1  4  4  3  2  2  1  4  3  4  1
SCHOOL  1 ALG   1  0  3  1  2  0  3  2  3  2  2  1  4  1  4  0
```

```
SCHOOL  2 NUM   5  3  4  4  3  2  3  3  2  2  4  3  4  3  5  3
SCHOOL  2 ALG   5  2  4  2  3  2  3  2  2  2  4  2  4  2  5  3
```

An answer key is not required for aggregate-level data in number-tried, number-right summary form. Note the format statement for reading the two sets of eight number-tried, number-right observations. For more information on how to set up the variable format statement for this type of data, see 3.2.16.

The items are multiple-choice and fairly difficult, so the 3PL model is needed. Because aggregate-level data are always more informative than individual-level item responses, it is worthwhile in the CALIB command to increase the number of quadrature points (NQPT), to set a stricter criterion for convergence (CRIT), and to increase the CYCLES limit. A prior on the thresholds (TPRIOR) and a ridge constant of 0.8 (RIDGE) are required for convergence with the exceptionally difficult ALGCON subtest. Aggregate-level data typically have smaller slopes in the 0,1 metric than do person-level data. For this reason, the mean of the prior for the log slopes has been set to 0.5 by use of the READPRIOR option of the CALIB command and the following PRIOR commands.

The aggregate scores for the schools are estimated by the EAP method using the empirical distributions from Phase 2. The number of quadrature points is set the same as in Phase 2.

The scores are rescaled to a mean of 250 and a standard deviation of 50 in the latent distribution of schools (IDIST=3, LOCATION=250, SCALE=50). The fit of the data to the group-level model is tested for each school (FIT). The NUMCON items have fairly homogeneous slopes and might be favorable for a one-parameter model.

```
EXAMPL06.BLM - MULTIPLE-MATRIX SAMPLING DATA
               AGGREGATE-LEVEL MODEL
>GLOBAL   NPARM=3, NTEST=2, DFNAME='EXAMPL06.DAT';
>LENGTH   NITEMS=(8,8);
>INPUT    NTOTAL=16, NALT=5, NIDCHAR=9, TYPE=3;
>ITEMS    INUM=(1(1)16), INAMES=(N1(1)N8,A1(1)A8);
>TEST1    TNAME=NUMCON, INUM=(1(1)8);
>TEST2    TNAME=ALGCON, INUM=(9(1)16);
(9A1,T15,8(2F3.0)/T15,8(2F3.0))
>CALIB    NQPT=51, CYCLES=50, NEWTON=10, CRIT=0.005, TPRIOR,
          READPRIOR, NOFLOAT, RIDGE=(2,0.8,2.0), CHI=8, PLOT=1;
>PRIORS1  SMU=(0.5(0)8);
>PRIORS2  SMU=(0.5(0)8);
>SCORE    NQPT=(12,12), IDIST=3, RSCTYPE=4,
          LOCATION=(250.0,250.0), SCALE=(50.0,50.0), FIT;
```

Phase 1 output

Group-level data consist of number-tried and number-right frequencies for each item in each group. The program reads them as values rather than characters and conversion to item scores is not required.

```
OBSERVATION #     1  WEIGHT:     1.0000  ID : SCHOOL  1

 SUBTEST #:   1    NUMCON
   GROUP #:   1

     TRIED    RIGHT
    23.000   14.000
 ITEM       1     2     3     4     5     6     7     8
 TRIED    1.0   3.0   2.0   4.0   3.0   2.0   4.0   4.0
 RIGHT    0.0   2.0   1.0   4.0   2.0   1.0   3.0   1.0

 SUBTEST #:   2    ALGCON
   GROUP #:   1

     TRIED    RIGHT
    22.000    7.000
 ITEM       1     2     3     4     5     6     7     8
 TRIED    1.0   3.0   2.0   3.0   3.0   2.0   4.0   4.0
 RIGHT    0.0   1.0   0.0   2.0   2.0   1.0   1.0   0.0

 OBSERVATION #     2  WEIGHT:     1.0000  ID : SCHOOL  2

 SUBTEST #:   1    NUMCON
   GROUP #:   1

     TRIED    RIGHT
    30.000   23.000
 ITEM       1     2     3     4     5     6     7     8
 TRIED    5.0   4.0   3.0   3.0   2.0   4.0   4.0   5.0
 RIGHT    3.0   4.0   2.0   3.0   2.0   3.0   3.0   3.0

 SUBTEST #:   2    ALGCON
   GROUP #:   1

     TRIED    RIGHT
    30.000   17.000
 ITEM       1     2     3     4     5     6     7     8
 TRIED    5.0   4.0   3.0   3.0   2.0   4.0   4.0   5.0
 RIGHT    2.0   2.0   2.0   2.0   2.0   2.0   2.0   3.0
```

Classical item statistics are computed for each subtest. Biserial correlations cannot be computed with group-level data.

```
ITEM STATISTICS FOR SUBTEST NUMCON
ITEM*TEST CORRELATION
ITEM   NAME        #TRIED    #RIGHT   PCT    LOGIT/1.7  PEARSON  BISERIAL
-------------------------------------------------------------------------
    1   N1          260.0     160.0    61.5    -0.28      0.637    0.000
    2   N2          268.0     162.0    60.4    -0.25      0.682    0.000
    3   N3          260.0     163.0    62.7    -0.31      0.663    0.000
    4   N4          261.0     137.0    52.5    -0.06      0.637    0.000
    5   N5          271.0     129.0    47.6     0.06      0.699    0.000
    6   N6          271.0     154.0    56.8    -0.16      0.656    0.000
    7   N7          270.0     157.0    58.1    -0.19      0.656    0.000
    8   N8          266.0     170.0    63.9    -0.34      0.781    0.000
-------------------------------------------------------------------------
```

```
 ITEM STATISTICS FOR SUBTEST ALGCON

                                                            ITEM*TEST CORRELATION
ITEM   NAME        #TRIED    #RIGHT   PCT    LOGIT/1.7  PEARSON  BISERIAL
-------------------------------------------------------------------------
    1   A1          259.0     120.0    46.3     0.09      0.636    0.000
    2   A2          267.0      81.0    30.3     0.49      0.606    0.000
    3   A3          241.0      94.0    39.0     0.26      0.669    0.000
    4   A4          245.0     121.0    49.4     0.01      0.687    0.000
    5   A5          263.0      96.0    36.5     0.33      0.669    0.000
    6   A6          263.0     166.0    63.1    -0.32      0.746    0.000
    7   A7          267.0      71.0    26.6     0.60      0.667    0.000
    8   A8          262.0      90.0    34.4     0.38      0.683    0.000
-------------------------------------------------------------------------
```

Phase 2 output

The set-up for group-level item calibration differs somewhat from examinee-level analysis: more quadrature points and more iterations for the solution are required. Prior distributions for all parameters are necessary, the means should be kept fixed (default = NOFLOAT), and the mean of the priors for slopes should be set lower than the examinee-level default.

```
>PRIORS1 SMU = (0.5000(0)8);

CONSTRAINT DISTRIBUTIONS ON ITEM PARAMETERS
  (THRESHOLDS, NORMAL; SLOPES, LOG-NORMAL; GUESSING, BETA)

                 THRESHOLDS         SLOPES         ASYMPTOTES
 ITEM           MU      SIGMA      MU      SIGMA     ALPHA        BETA
 ---------------------------------------------------------------------
 N1          0.000    2.000    0.500    1.649       5.00      17.00
 N2          0.000    2.000    0.500    1.649       5.00      17.00
 N3          0.000    2.000    0.500    1.649       5.00      17.00
 N4          0.000    2.000    0.500    1.649       5.00      17.00
 N5          0.000    2.000    0.500    1.649       5.00      17.00
 N6          0.000    2.000    0.500    1.649       5.00      17.00
 N7          0.000    2.000    0.500    1.649       5.00      17.00
 N8          0.000    2.000    0.500    1.649       5.00      17.00
 ---------------------------------------------------------------------
```

Group-level item parameter estimates for the first 3 items in subtest NUMCON are as follows.

```
SUBTEST NUMCON  ;  ITEM PARAMETERS AFTER CYCLE   12

ITEM      INTERCEPT    SLOPE    THRESHOLD   LOADING   ASYMPTOTE    CHISQ     DF
              S.E.       S.E.       S.E.       S.E.       S.E.      (PROB)
 -------------------------------------------------------------------------------
 N1       |   0.030  |   0.190  |  -0.156  |   0.186  |   0.232  |      5.7   6.0
          |   0.194* |   0.066* |   1.026* |   0.065* |   0.094* | (0.4521)
          |          |          |          |          |          |
 N2       |   0.046  |   0.279  |  -0.163  |   0.268  |   0.218  |      3.8   6.0
          |   0.222* |   0.107* |   0.801* |   0.103* |   0.093* | (0.7025)
          |          |          |          |          |          |
 N3       |   0.126  |   0.313  |  -0.404  |   0.299  |   0.212  |      3.2   5.0
          |   0.224* |   0.120* |   0.735* |   0.115* |   0.091* | (0.6638)
-------------------------------------------------------------------------------
                                                * STANDARD ERROR
```

```
    LARGEST CHANGE =    0.003146                             42.8  53.0
                                                             (0.8397)
NOTE: ITEM FIT CHI-SQUARES AND THEIR SUMS MAY BE UNRELIABLE
FOR TESTS WITH LESS THAN 20 ITEMS

PARAMETER       MEAN  STN DEV
-----------------------------------
ASYMPTOTE      0.210    0.041
SLOPE          0.306    0.099
LOG(SLOPE)    -1.223    0.290
THRESHOLD      2.241    1.515
```

Phase 3 output

Computing scores at the group-level is essentially the same as at the examinee level. Note that the selection of EAP estimations based on the empirical latent distribution from Phase 2 overrides the choice here of number of quadrature points. Because of the small number of items, the standard deviation of the estimated scores is considerably smaller than that of the latent distribution. Portions of the Phase 3 output are listed below.

```
>SCORE NQPT = (12, 12),IDIST = 3,RSCTYPE = 4,
LOCATION = (250.0000, 250.0000), SCALE = (50.0000, 50.0000),FIT;

PARAMETERS FOR SCORING, RESCALING, AND TEST AND ITEM INFORMATION
METHOD OF SCORING SUBJECTS:                EXPECTATION A POSTERIORI
                                           (EAP; BAYES ESTIMATION)
TYPE OF PRIOR:                             EMPIRICAL, FROM ITEM CALIBRATION
SUBJECT FIT PROBABILITIES:                 YES
TYPE OF RESCALING:                         IN THE ESTIMATED LATENT
                                            DISTRIBUTION
REFERENCE GROUP FOR RESCALING:             GROUP:  1

              QUAD   RESCALING  CONSTANTS
TEST    NAME  POINTS    SCALE    LOCATION
-------------------------------------------
   1  NUMCON     51    50.000   250.000
   2  ALGCON     51    50.000   250.000
--------------------------------------------
```

The scores are rescaled so that the mean and standard deviation of the Phase 3 latent distribution are 250 and 50, respectively. Scores for all 32 schools are computed and printed. Because the data are binomial rather than binary, a χ^2 index of fit on 8 degrees of freedom can be calculated for each school. The corresponding probabilities are shown in the output.

```
RESCALING WITH RESPECT TO LATENT DISTRIBUTION
--------------------------------------------------


         RESCALING  CONSTANTS
TEST       SCALE    LOCATION
NUMCON      58.462   251.342
ALGCON      56.462   251.127
```

```
GROUP   SUBJECT IDENTIFICATION                              FIT  MARGINAL
WEIGHT  TEST  TRIED  RIGHT  PERCENT     ABILITY      S.E.   PROB   PROB
-------------------------------------------------------------------------
1  SCHOOL  1                        |                      |
1.00   NUMCON    23     14    60.87 |  246.5104   43.5894 | 0.1539 0.0000
1  SCHOOL  1                        |                      |
1.00   ALGCON    22      7    31.82 |  243.1547   47.4683 | 0.3197 0.0000
 [Similar output omitted]
1  SCHOOL 32                        |                      |
1.00   NUMCON   181    100    55.25 |  221.6762   21.9655 | 0.0166 0.0000
1  SCHOOL 32                        |                      |
1.00   ALGCON   179     77    43.02 |  273.1747   21.8821 | 0.5242 0.0000
-------------------------------------------------------------------------

MEAN & SD OF SCORE ESTIMATES AFTER RESCALING:   250.000    31.149
MEAN & SD OF LATENT DISTRIBUTION AFTER RESCALING:   250.000    50.000
```

6.7 Analysis of variant items

In this example, responses to 50 items are read from those of 100 items in the data file using the format statement

```
(10A1,T38,25A1,1X,25A1).
```

The first few lines of the data file are shown below. In contrast to previous examples, each position in the item response fields for each examinee corresponds to the same item. In the earlier examples, the association between response and item depended on the group/form membership of an examinee.

The answer key (KFNAME keyword on the INPUT command) is given first, and is given in the first two lines of the raw data file in the same format as the item responses.

```
KEY        00000000000000000000000000000000000000000000000000000000000000000000000…
0102111900 00000401020100002001101002024030005001000000000233004002014062000012000100…
0104112200 10101200210100000100010230110030013000000100103021014000002042001012001000…
0105121900 11012041110200000010002230131010122101000000013123000002001042101012001300…
```

From the 50, 20 are selected as Main Test items and 4 as Variant Test items. This is indicated by setting NITEMS to 24 and NVARIANT to 4 on the LENGTH command. Items for the main test are selected by name in the TESTM command; items for the variant test are selected by name in the TESTV command. The item names correspond to the sequence numbers in the original set of 100 items. Here the short form of naming and numbering is used – the set of items forms an arithmetic progression of integer or decimal numbers allowing use of the short form (*first (increment) last*). A similar abbreviation may be used for consecutive item names (INAMES keyword on the ITEMS command).

The analysis is performed on a sample of 200 students randomly drawn from the original sample of 660 (SAMPLE=200 on the INPUT command). The EAP scale scores of Phase 3 are computed from the responses to items in the main test.

```
EXAMPL07.BLM - ANALYSIS OF VARIANT ITEMS IN A SPELLING TEST OF RANDOMLY
               SELECTED WORDS; SUBJECTS: 660 UNDERGRADUATE STUDENTS; 2PL MODEL.
```

```
>GLOBAL  DFNAME='EXAMPL07.DAT', NTEST=1, NVTEST=1, NPARM=2, SAVE;
>SAVE    PARM='EXAMPL07.PAR', SCORE='EXAMPL07.SCO';
>LENGTH  NITEM=24, NVARIANT=4;
>INPUT   NTOTAL=50, KFNAME='EXAMPL07.DAT', SAMPLE=200, NIDCHAR=10;
>ITEMS   INUMBERS=(1(1)50), INAME=(I26(1)I75);
>TESTM   TNAME=MAINTEST, INAMES=(I26,I27,I28,I29,I31,I33,I34,
         I35,I36,I38,I39,I47,I48,I49,I50,I54,I60,I64,I68,I72);
>TESTV   TNAME=VARIANT, INAMES=(I53,I59,I69,I73);
(10A1,T38,25A1,1X,25A1)
>CALIB   NQPT=31, CRIT=.005, CYCLES=10, NEWTON=2, FLOAT, ACCEL=0.5;
>SCORE   METHOD=2, NOPRINT;
```

Phase 1 output

Phase 1 lists the test specifications and the assignment of items to the main test and the variants.

```
>ITEMS  INUMBERS=(1(1)50), INAME=(I26(1)I75);


 TEST SPECIFICATIONS
 ===================


 >TESTM   TNAME=MAINTEST,
          INAMES=(I26,I27,I28,I29,I31,I33,I34,
          I35,I36,I38,I39,I47,I48,I49,I50,I54,I60,I64,I68,I72);


 TEST NUMBER:   1    TEST NAME: MAINTEST
 NUMBER OF ITEMS:   20

      ITEM   ITEM        ITEM   ITEM        ITEM   ITEM        ITEM   ITEM
    NUMBER   NAME      NUMBER   NAME      NUMBER   NAME      NUMBER   NAME
  ----------------------------------------------------------------------
         1   I26            9   I34           23   I48           43   I68
         2   I27           10   I35           24   I49           47   I72
         3   I28           11   I36           25   I50
         4   I29           13   I38           29   I54
         6   I31           14   I39           35   I60
         8   I33           22   I47           39   I64
  ----------------------------------------------------------------------


 >TESTV   TNAME=VARIANT,
          INAMES=(I53,I59,I69,I73);


 TEST NUMBER:   2    TEST NAME: VARIANT
 NUMBER OF ITEMS:    4

      ITEM   ITEM        ITEM   ITEM        ITEM   ITEM        ITEM   ITEM
    NUMBER   NAME      NUMBER   NAME      NUMBER   NAME      NUMBER   NAME
  ----------------------------------------------------------------------
        28   I53           34   I59           44   I69           48   I73
  ----------------------------------------------------------------------
```

Responses of 660 examinees are read from the data records, but only 200 randomly sampled cases are included in the Phase 1 and Phase 2 analysis. The classical item statistics are shown separately for main and variant items. The test scores for the item-test correlations are based on the test scores from the main test items only.

```
       660 OBSERVATIONS READ FROM FILE:   EXAMPL07.DAT
       660 OBSERVATIONS WRITTEN TO FILE:  MF.DAT

 REPORT ON SUBJECT SAMPLING:
 LEVEL OF SAMPLING =     0.3030
       660 SUBJECTS READ FROM FILE:    MF.DAT
       200 SUBJECTS WRITTEN TO FILE:   CF.DAT

 ITEM STATISTICS FOR SUBTEST MAINTEST
                                                         ITEM*TEST CORRELATION
ITEM  NAME        #TRIED    #RIGHT   PCT    LOGIT/1.7  PEARSON  BISERIAL
-----------------------------------------------------------------------
    1   I26         200.0     134.0    67.0    -0.42      0.188    0.244
    2   I27         200.0     102.0    51.0    -0.02      0.421    0.527
    3   I28         200.0      78.0    39.0     0.26      0.294    0.374
    4   I29         200.0     147.0    73.5    -0.60      0.444    0.598
...
 -----------------------------------------------------------------------

 ITEM STATISTICS FOR SUBTEST VARIANT
                                                          ITEM*TEST CORRELATION
ITEM   NAME        #TRIED    #RIGHT   PCT    LOGIT/1.7  PEARSON  BISERIAL
------------------------------------------------------------------------
    1   I53         200.0     139.0    69.5    -0.48      0.454    0.596
    2   I59         200.0     135.0    67.5    -0.43      0.456    0.594
    3   I69         200.0      53.0    26.5     0.60      0.379    0.510
    4   I73         200.0      50.0    25.0     0.65      0.069    0.094
------------------------------------------------------------------------
```

Phase 2 output

Calibration of the main test items is computed as in the other examples. Without altering the item parameter estimates of those items, parameter estimates for the variants are computed with respect to the latent dimension determined by the main items.

```
SUBTEST MAINTEST;  ITEM PARAMETERS AFTER CYCLE   6

ITEM INTERCEPT    SLOPE    THRESHOLD   LOADING   ASYMPTOTE    CHISQ   DF
          S.E.       S.E.       S.E.       S.E.       S.E.     (PROB)
------------------------------------------------------------------------
I26|   0.451  |   0.360  |  -1.254  |   0.339  |   0.000  |      8.6   8.0
   |   0.648* |   0.088* |   1.775* |   0.083* |   0.000* | (0.3784)
   |          |          |          |          |          |
I27|   0.028  |   0.753  |  -0.037  |   0.602  |   0.000  |      3.2   6.0
   |   0.691* |   0.152* |   0.918* |   0.121* |   0.000* | (0.7857)
   |          |          |          |          |          |
 (Similar output omitted)
I72|  -0.018  |   0.726  |   0.025  |   0.587  |   0.000  |     13.6   6.0
   |   0.684* |   0.149* |   0.942* |   0.121* |   0.000* | (0.0347)
------------------------------------------------------------------------
                                                        * STANDARD ERROR

     LARGEST CHANGE =    0.002542                          106.4 131.0
                                                         (0.9439)
```

```
PARAMETER       MEAN  STN DEV
 ----------------------------------
 SLOPE           0.616    0.220
 LOG(SLOPE)     -0.548    0.368
 THRESHOLD       0.143    1.256

                        *****************************
                         CALIBRATION OF VARIANT ITEMS
                                   VARIANT
                        *****************************

 -2 LOG LIKELIHOOD =       4545.542

ITEM INTERCEPT   SLOPE    THRESHOLD   LOADING   ASYMPTOTE    CHISQ     DF
        S.E.       S.E.        S.E.       S.E.        S.E.      (PROB)
-------------------------------------------------------------------------
I53|   0.587  |   0.613  |  -0.957  |   0.523  |   0.000  |      0.0   0.0
   |   0.104* |   0.111* |   0.202* |   0.094* |   0.000* | (1.0000)
   |          |          |          |          |          |
I59|   0.519  |   0.603  |  -0.860  |   0.517  |   0.000  |      0.0   0.0
   |   0.101* |   0.109* |   0.195* |   0.093* |   0.000* | (1.0000)
   |          |          |          |          |          |
I69|  -0.702  |   0.549  |   1.280  |   0.481  |   0.000  |      0.0   0.0
   |   0.107* |   0.109* |   0.263* |   0.095* |   0.000* | (1.0000)
   |          |          |          |          |          |
I73|  -0.668  |   0.231  |   2.886  |   0.225  |   0.000  |      0.0   0.0
   |   0.098* |   0.064* |   0.857* |   0.062* |   0.000* | (1.0000)
-------------------------------------------------------------------------
```

Phase 3 output

In Phase 3, scores for all 660 examinees are computed from the main test item response and saved to an external file. Printing of the scores is suppressed, except for the first three cases. The latent distribution estimated from all 660 cases is computed and printed. Scores are based on the unrescaled Phase 2 parameters, which are then saved to an external file.

```
 >SCORE   METHOD=2,NOPRINT;

 SCORES WILL NOT BE COMPUTED FOR VARIANT ITEM SUBTESTS
 PARAMETERS FOR SCORING, RESCALING, AND TEST AND ITEM INFORMATION

 METHOD OF SCORING SUBJECTS:                 EXPECTATION A POSTERIORI
                                             (EAP; BAYES ESTIMATION)
 TYPE OF PRIOR:                              NORMAL
 SCORES WRITTEN TO FILE                      EXAMPL07.SCO

GROUP   SUBJECT IDENTIFICATION                                   MARGINAL
WEIGHT   TEST       TRIED  RIGHT  PERCENT     ABILITY      S.E.     PROB
-------------------------------------------------------------------------
 1  0102111900                            |                      |
 1.00   MAINTEST     20      8     40.00 |   -0.4065    0.3645 |  0.000000
 1  0104112200                            |                      |
 1.00   MAINTEST     20      8     40.00 |   -0.4091    0.3641 |  0.000000
 1  0105121900                            |                      |
 1.00   MAINTEST     20      3     15.00 |   -1.2316    0.4637 |  0.000000
-------------------------------------------------------------------------
```

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

CORRELATIONS AMONG TEST SCORES

                MAINTEST
MAINTEST       1.0000

MEANS, STANDARD DEVIATIONS, AND VARIANCES OF SCORE ESTIMATES

TEST:               MAINTEST
MEAN:              0.0915
S.D.:              0.8940
VARIANCE:          0.7992

ROOT-MEAN-SQUARE POSTERIOR STANDARD DEVIATIONS

TEST:               MAINTEST
RMS:               0.4493
VARIANCE:          0.2019

EMPIRICAL RELIABILITY:     0.7984

MARGINAL LATENT DISTRIBUTION(S)
===============================

MARGINAL LATENT DISTRIBUTION FOR TEST MAINTEST
   MEAN     =     0.092
   S.D.     =     0.974
```

6.8 Group-wise adaptive testing

This example illustrates the use of BILOG-MG with multiple groups and multiple subtests. It is designed to illustrate some of the more complicated features of the program, including user-specified priors on the latent distributions and priors on the item parameters.

Based on previous test performance, examinees are assigned to two groups for adaptive testing. Out of a set of 45 items, group 1 is assigned items 1 through 25, and group 2 is assigned items 21 through 45. Thus, there are 5 items linking the test forms administered to the groups.

Twenty of the 25 items presented to group 1 belong to subtest 1 (items 1-15 and 21-25); twenty items also belong to subtest 2 (items 6-25). Of the 25 items presented to group 2, 20 belong to subtest 1 (items 21-40) and 20 to subtest 2 (items 21-25 and 31-45).

In all, there are 35 items from the set of 45 assigned to each subtest. (This extent of item overlap between subtests is not realistic, but it illustrates that more than one subtest can be scored adaptively provided they each contain link items between the test forms.)

This example also illustrates how user-supplied priors for the latent distributions are specified with IDIST=1 on the CALIB command. The points and weights for these distributions are supplied in the QUAD commands. Note that with IDIST=1, there are separate QUAD commands for each group for each subtest. Within each subtest the points are the same for each group. This is a requirement of the program. But as the example shows, the points for the groups may differ by subtest. If IDIST has been set to 2, sets of weights have to be supplied by group. The set of points then applies to all subtests.

The PRIOR command for each subtest is placed after the QUAD commands for that subtest. The presence of the PRIOR command is indicated using the READPRIOR option on the CALIB command. In this example, only the prior for the standard deviation of the thresholds is supplied on the PRIOR command. Default values are used for the other prior distributions. The means of the distributions are kept fixed at their specified values by using the NOFLOAT option on the CALIB command.

The score distribution in the respondent population is estimated in the form of a discrete distribution on NQPT=16 points by adding the EMPIRICAL option to the CALIB command. This discrete distribution will be used in the place of the prior in MML estimate of the item parameters. When NGROUP>1, separate score distributions will be estimated for the groups. The first group serves as the reference group (REFERENCE=1). If the REFERENCE keyword is omitted, the first group will by default be used as the reference group. When NGROUP>1, the FLOAT option is the default. By using NOFLOAT here, the means of the prior distributions on item parameters are kept fixed at the specified values during estimation.

In the scoring phase, the empirical prior from phase 2 is used as prior distribution for the scale scores (IDIST=3). Rescaling of scores to the scale and location in the sample of scale score estimates is requested by setting RSCTYPE to 3. The presence of the INFO keyword indicates that information output is required. In this case INFO=1 and test information curves will be printed to the phase 3 output file. In combination with the YCOMMON and POP options, the test information curves will be expressed in comparable units and an estimate of the classical reliability coefficient, amongst other information, will be calculated for each subtest.

```
EXAMPL08.BLM -
                GROUP-WISE ADAPTIVE TESTING WITH TWO SUBTESTS
>GLOBAL DFNAME='EXAMPL08.DAT', NPARM=2, NTEST=2, SAVE;
>SAVE   SCORE='EXAMPL08.SCO';
>LENGTH NITEMS=(35,35);
>INPUT  NTOT=45, SAMPLE=2000, NGROUP=2, KFNAME='EXAMPL08.DAT', NALT=5,
        NFORMS=2, NIDCH=5;
>ITEMS  INUM=(1(1)45),  INAME=(C01(1)C45);
>TEST1  TNAME=SUBTEST1, INAME=(C01(1)C15,C21(1)C40);
>TEST2  TNAME=SUBTEST2, INAME=(C06(1)C25,C31(1)C45);
>FORM1  LENGTH=25, INUM=(1(1)25);
>FORM2  LENGTH=25, INUM=(21(1)45);
>GROUP1 GNAME=POP1, LENGTH=25, INUM=(1(1)25);
>GROUP2 GNAME=POP2, LENGTH=25, INUM=(21(1)45);
(5A1,T1,I1,T1,I1,T7,25A1)
>CALIB  IDIST=1, READPRIOR, EMPIRICAL, NQPT=31, CYCLE=25, TPRIOR, NEWTON=5,
        CRITERION=0.01, REFERENCE=1, NOFLOAT;
>QUAD1  POINTS=(-0.4598E+01,-0.3560E+01,-0.2522E+01,-0.1484E+01,-0.4453E+00,
        0.5930E+00, 0.1631E+01, 0.2670E+01, 0.3708E+01, 0.4746E+01),
        WEIGHTS=(0.2464E-05, 0.4435E-03, 0.1724E-01, 0.1682E+00, 0.3229E+00,
        0.3679E+00, 0.1059E+00, 0.1685E-01, 0.6475E-03, 0.8673E-05);
>QUAD2  POINTS=(-0.4598E+01,-0.3560E+01,-0.2522E+01,-0.1484E+01,-0.4453E+00,
        0.5930E+00, 0.1631E+01, 0.2670E+01, 0.3708E+01, 0.4746E+01),
```

```
        WEIGHTS=(0.2996E-04, 0.1300E-02, 0.1474E-01, 0.1127E+00, 0.3251E+00,
        0.3417E+00, 0.1816E+00, 0.2149E-01, 0.1307E-02, 0.3154E-04);
>PRIOR  TSIGMA=(1.5(0)35);
>QUAD1  POINTS=(-0.4000E+01,-0.3111E+01,-0.2222E+01,-0.1333E+01,-0.4444E+00,
        0.4444E+00, 0.1333E+01, 0.2222E+01, 0.3111E+01, 0.4000E+01),
        WEIGHTS=(0.1190E-03, 0.2805E-02, 0.3002E-01, 0.1458E+00, 0.3213E+00,
        0.3213E+00, 0.1458E+00, 0.3002E-01, 0.2805E-02, 0.1190E-03);
>QUAD2  POINTS=(-0.4000E+01,-0.3111E+01,-0.2222E+01,-0.1333E+01,-0.4444E+00,
        0.4444E+00, 0.1333E+01, 0.2222E+01, 0.3111E+01, 0.4000E+01),
        WEIGHTS=(0.1190E-03, 0.2805E-02, 0.3002E-01, 0.1458E+00, 0.3213E+00,
        0.3213E+00, 0.1458E+00, 0.3002E-01, 0.2805E-02, 0.1190E-03);
>PRIOR  TSIGMA=(1.5(0)35);
>SCORE  IDIST=3, RSCTYPE=3, INFO=1, YCOMMON, POP, NOPRINT;
```

Phase 1 output

Phase 1 echoes the assignment of items to subtests, forms, and groups. Classical item statistics are computed for each subtest in each group. Output for subtest 1 and group 1 (POP1) is given below.

```
           SUBTEST   1    SUBTEST1
             GROUP   1    POP1         200 OBSERVATIONS
             GROUP   2    POP2         200 OBSERVATIONS

           SUBTEST   2    SUBTEST2
             GROUP   1    POP1         200 OBSERVATIONS
             GROUP   2    POP2         200 OBSERVATIONS

 SUBTEST    1    SUBTEST1
 ITEM STATISTICS FOR GROUP:  1    POP1
                                                      ITEM*TEST CORRELATION
ITEM   NAME      #TRIED    #RIGHT    PCT     LOGIT/1.7  PEARSON  BISERIAL
-------------------------------------------------------------------------
    1   C01         200.0     170.0   0.850    -1.02      0.408    0.625
    2   C02         200.0     164.0   0.820    -0.89      0.396    0.580
    3   C03         200.0     154.0   0.770    -0.71      0.451    0.626
    4   C04         200.0     143.0   0.715    -0.54      0.400    0.532
    5   C05         200.0     140.0   0.700    -0.50      0.586    0.772
    6   C06         200.0     135.0   0.675    -0.43      0.441    0.574
...
   19   C24         200.0      83.0   0.415     0.20      0.590    0.746
   20   C25         200.0      76.0   0.380     0.29      0.558    0.711
-------------------------------------------------------------------------
```

Phase 2 output

Phase 2 estimates empirical latent distributions for each group and item parameters for each subtest. The arbitrary mean and standard deviation of reference group 1 determine the origin and unit of the ability scales.

```
>CALIB   IDIST=1, READPRIOR, EMPIRICAL, NQPT=31, CYCLE=25, TPRIOR, NEWTON=5,
         CRITERION=0.01, REFERENCE=1, NOFLOAT;

ITEM      INTERCEPT    SLOPE    THRESHOLD   LOADING   ASYMPTOTE    CHISQ     DF
             S.E.       S.E.       S.E.       S.E.       S.E.      (PROB)
 ------------------------------------------------------------------------------
 C01       |   1.435  |   0.930  |  -1.542  |   0.681  |   0.000  |      8.5   6.0
           |   0.196* |   0.187* |   0.211* |   0.137* |   0.000* | (0.2037)
           |          |          |          |          |          |
 C02       |   1.196  |   0.823  |  -1.453  |   0.635  |   0.000  |      7.7   6.0
           |   0.162* |   0.163* |   0.215* |   0.126* |   0.000* | (0.2580)
           |          |          |          |          |          |
 C03       |   1.028  |   0.922  |  -1.115  |   0.678  |   0.000  |      5.8   6.0
           |   0.160* |   0.169* |   0.153* |   0.124* |   0.000* | (0.4441)
       ...
           |          |          |          |          |          |
 C38       |  -0.962  |   1.098  |   0.876  |   0.739  |   0.000  |      6.7   6.0
           |   0.164* |   0.182* |   0.115* |   0.122* |   0.000* | (0.3520)
           |          |          |          |          |          |
 C39       |  -1.144  |   0.879  |   1.302  |   0.660  |   0.000  |      1.7   5.0
           |   0.173* |   0.170* |   0.169* |   0.128* |   0.000* | (0.8927)
           |          |          |          |          |          |
 C40       |  -1.044  |   0.632  |   1.652  |   0.534  |   0.000  |      3.0   6.0
           |   0.133* |   0.123* |   0.268* |   0.104* |   0.000* | (0.8143)
 ------------------------------------------------------------------------------
                                                              * STANDARD ERROR
     LARGEST CHANGE =    0.008756                            171.9 233.0
                                                            (0.9990)

 PARAMETER       MEAN  STN DEV
 -----------------------------------
 SLOPE          0.862    0.154
 LOG(SLOPE)    -0.165    0.180
 THRESHOLD     -0.164    0.908
```

Phase 3 output

The only new feature in Phase 3 is the use of the YCOMMON option to place the information plots for the subtests on the same scale. This permits visual comparison of the relative precision of the subtests according to the heights of the information curves. To illustrate, the ICC for subtest 1, form is given below. The POP option also provides IRT estimated reliability for each subtest.

```
TEST INFORMATION CURVE FOR  TEST:   SUBTEST1        FORM:       1
STANDARD                                                              INFOR-
ERROR                                                                 MATION
    ------------------------------------------------------------------
4.41|                                                                |10.0923
    |                                                                |
4.19|                                                                | 9.5876
    |                                                                |
3.97|                              ++++                              | 9.0830
    |                             +    +                             |
3.75|                            +      +                            | 8.5784
    |                           +        +                           |
3.53|                          +          +                          | 8.0738
    |                                                                |
3.31|                         +            +                         | 7.5692
    |                                                                |
3.08|                        +              +                        | 7.0646
    |                                                                |
2.86|                       +                +                       | 6.5600
    |                                                                |
2.64|                      +                                         | 6.0554
    |                                         +                      |
2.42|                     +                                          | 5.5507
    |                                          +                     |
2.20|                    +                                          *| 5.0461
    |                                           +                  * |
1.98|*                  +                                            | 4.5415
    |                  +                         +                * |
1.76| *                                                          *   | 4.0369
    |  *              +                           +                  |
1.54|   *                                                       *    | 3.5323
    |                +                             +           *     |
1.32|    *                                                    *      | 3.0277
    |     **        +                               +        *       |
1.10|       *      +                                 +      *        | 2.5231
    |        *    +                                       **         |
0.88|         ** +                                    +  *           | 2.0185
    |           +*                                     +*            |
0.66|          +  **                                 ** +            | 1.5138
    |         +     ***                           ***    +           |
0.44|        +         ******                *****        ++         | 1.0092
    |      ++                ***************                ++       |
0.22|   +++                                                   +++    | 0.5046
    |+++                                                         ++++|
0.00|                                                                | 0.0000
    -+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
   -4.00   -3.00   -2.00   -1.00    0.00    1.00    2.00    3.00    4.00

MAXIMUM INFORMATION APPROXIMATELY  0.9046D+01  AT      0.0000
FOR A NORMAL POPULATION WITH MEAN   0.000 AND S.D.   1.000
AVERAGE INFORMATION=  0.7232D+01  AND RELIABILITY INDEX=   0.879
```

6.9 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). A complete description of these data are given in Section 2.14.1.

On the basis of item parameters they report, 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). The program command files as well as the data file (with N = 660) are included in the example folder of the BILOG-MG installation folder.

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 recali-
brating 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 6.1. 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.

Table 6.1: 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

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.

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 6.1. 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.

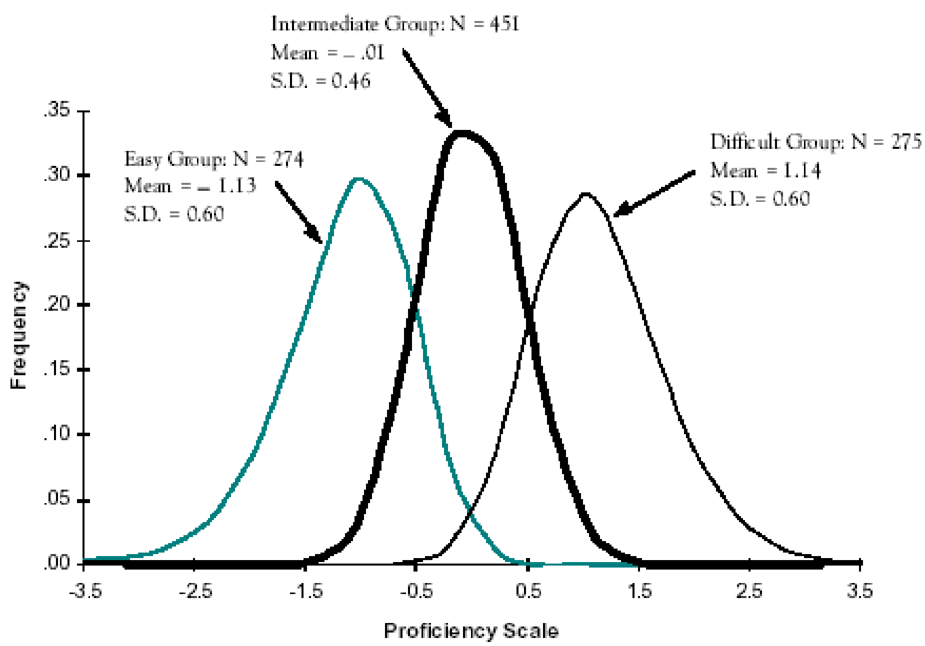


Figure 6.1. Prototype 1: estimated latent distributions from two-stage and one-stage spelling data

One may infer the measurement properties of the simulated two-stage spelling test from the information and efficiency calculations shown in Figure 6.2 and Figure 6.3, 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 6.2, this range of reliability is achieved in the two-stage results for spelling over much of the latent distribution.

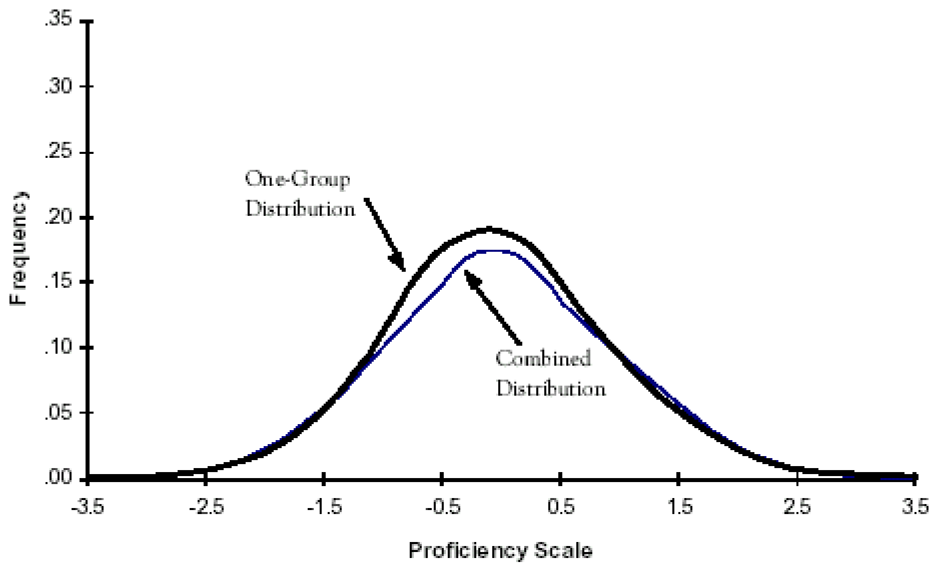


Figure 6.2. Prototype 1: two-stage spelling test

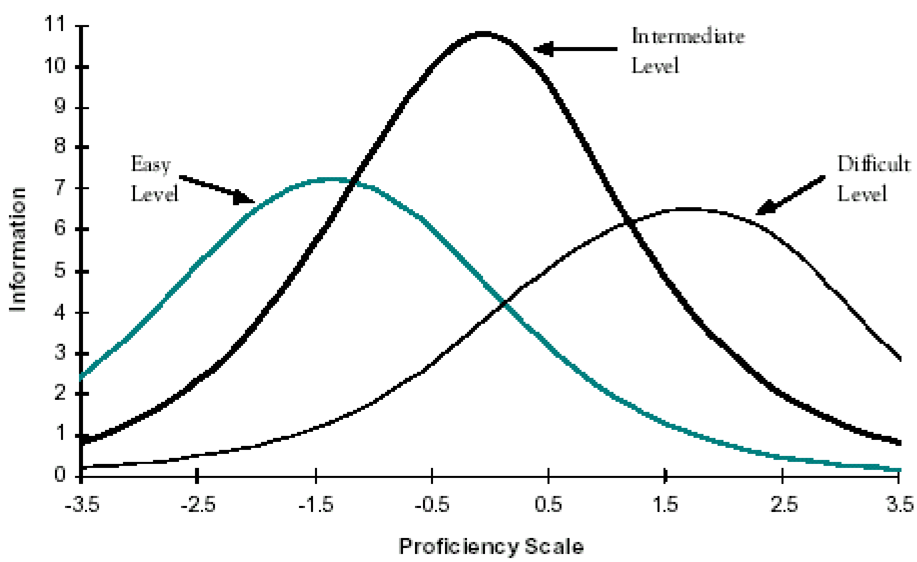


Figure 6.3. Prototype 1: efficiencies of the two-stage spelling tests

Finally, the efficiency curves in Figure 6.3 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.

6.10 Estimating and scoring tests of increasing length

In this example, commands for estimating item parameters and computing score means, standard deviations, variances, average standard errors, error variances, and inverse information reliabilities of maximum likelihood estimates of ability, are illustrated.

Note: to obtain the same results for EAP estimation, set METHOD=2 in the SCORE command; for MAP estimation, set METHOD=3.

```
EXAMPL10.BLM - MML estimation of item parameters
               ML estimation of case scores
>GLOBAL DFNAME='SIM01C0.SIM',NPARM=2,NTEST=6,SAVE;
>SAVE   SCORE='MLEVAL1.SCO';
>LENGTH NITEMS=(4,8,16,32,64,128);
>INPUT  NTOTAL=128,NIDCH=5,SAMPLE=3000;
>ITEMS  INUMBERS=(1(1)128),INAME=(ITEM001(1)ITEM128);
>TEST1  TNAME=LENGTH4,INUMBERS=(1(1)4);
>TEST2  TNAME=LENGTH8,INUMBERS=(1(1)8);
>TEST3  TNAME=LENGTH16,INUMBERS=(1(1)16);
>TEST4  TNAME=LENGTH32,INUMBERS=(1(1)32);
>TEST5  TNAME=LENGTH64,INUMBERS=(1(1)64);
>TEST5  TNAME=LENG128,INUMBERS=(1(1)128);
(11A1,1X,128A1)
>CALIB NQPT=40,CYCLE=25,TPRIOR,NEWTON=3,CRIT=0.001,NOSPRIOR,NOADJUST;
>SCORE  METHOD=1,INFO=1,YCOMMON,POP,NOPRINT;
```

Related topics

- SCORE command: METHOD keyword

6.11 Commands for parallel-form correlations

This example contains the syntax used for computing parallel form correlations and between test correlations for tests of different lengths. Set METHOD equal to 1, 2, or 3 in the SCORE command to obtain correlations for ML, EAP, and MAP estimated abilities respectively.

```
EXAMPL11.BLM - Correlation of independent ML estimates

>GLOBAL DFNAME='SIM01C0.SIM',NPARM=2,NTEST=12,SAVE;
>SAVE   SCORE='MAPCOR1.SCO';
>LENGTH NITEMS=(4,4,8,8,16,16,32,32,64,64,128,128);
```

```
>INPUT  NTOTAL=504,NIDCH=5,SAMPLE=3000;
>ITEMS  INUMBERS=(1(1)504),INAME=(ITEM001(1)ITEM504);
>TEST1  TNAME=LENGTH4a,INUMBERS=(1(1)4);
>TEST2  TNAME=LENGTH4b,INUMBERS=(5(1)8);
>TEST3  TNAME=LENGTH8a,INUMBERS=(9(1)16);
>TEST4  TNAME=LENGTH8b,INUMBERS=(17(1)24);
>TEST5  TNAME=LEN16a, INUMBERS=(25(1)40);
>TEST6  TNAME=LEN16b, INUMBERS=(41(1)56);
>TEST7  TNAME=LEN32a, INUMBERS=(57(1)88);
>TEST8  TNAME=LEN32b, INUMBERS=(93(1)120);
>TEST9  TNAME=LEN64a, INUMBERS=(121(1)184);
>TEST10 TNAME=LEN64b, INUMBERS=(185(1)248);
>TEST11 TNAME=LEN128a ,INUMBERS=(249(1)376);
>TEST12 TNAME=LEN128b, INUMBERS=(377(1)504);
(11A1,1X,504A1)
>CALIB NQPT=40,CYCLE=25,NEWTON=3,CRIT=0.001,NOSPRIOR,NOADJUST;
>SCORE METHOD=1,INFO=1,YCOMMON,POP,NOPRINT;
```

Related topics

- SCORE command: METHOD keyword

6.12 EAP scoring of the NAEP forms and state main and variant tests

The syntax in this example was used to score NAEP forms and state main and variant tests. It is included here as an example of a more complicated analysis and contains numerous TEST and FORMS commands.

The use of the INUMBERS keyword on the FORM commands to assign items to the various forms is of interest, as is the naming convention used with the INAMES keyword on the ITEMS command. Finally, note that none of the tests are calibrated (SELECT=0 for all tests on the CALIB command). Scoring is done according to a previously generated item parameter file **gr4fin.par** read with the IFNAME keyword on the GLOBAL command.

```
EXAMPLE12.BLM - Scoring of main and variant tests
                Grade 4 Reading
>COMMENTS
******************************************************************************
                This example is for illustration purposes only.
                The actual data to run the command file is not available.
******************************************************************************


The syntax in this example was used to score NAEP forms and state main and variant
tests. It is included here as an example of a more complicated analysis and contains
numerous TEST and FORMS commands.

The use of the INUMBERS keyword on the FORM commands to assign items to the various
forms is of interest, as is the naming convention used with the INAMES keyword on the
ITEMS command. Finally, note that none of the tests are calibrated (SELECT=0 for all
tests on the CALIB command). Scoring is done according to a previously generated item
parameter file (gr4fin.par) read with the IFNAME keyword on the GLOBAL command.

The variant items in a test are not intended to be scored as a test. They are included
in the analysis to obtain preliminary information on there item characteristics with
respect to the latent variable measured by the main test.
```

```
>GLOBAL NPARM=3, NTEST=6, NVTEST=6, DFNAME='GR4FIN.DAT',
          IFNAME='GR4FIN.PAR', SAVE;
>SAVE   SCORE='GR4FIN.SCO';
>LENGTH NITEMS=(82,20,56,82,25,47), NVARIANT=(82,20,56,82,25,47);
>INPUT  NTOTAL=230, SAMPLE=3000, NIDCH=10, NFORM=16,
        KFNAME='GR4FIN.DAT', OFNAME='GR4FIN.DAT', NFNAME='GR4FIN.DAT';
>ITEMS  INUMBERS=(1(1)230), INAME=
        (MC01,MC02,MC03, OEtw04,OETW05,OEtw06, OEth07,OEth08, OEfo09,
         MC10(1)MC14, OEtw15(1)OEtw20, OEfo21,
         MC22(1)MC28, OEtw29(1)OEtw31, OEfo32,
         MC33(1)MC37, OEtw38(1)OEtw41, OEfo42,
         MC43(1)MC46, OEtw47(1)OEtw51, OEfo52,
         MC53(1)MC55, OEth56(1)OEth60, OEfo61,
         MC62(1)MC64, OEtw65, OEth66(1)OEth69, OEfo70,
         MC71(1)MC76, OEtw77(1)OEtw81, OEfo82,
         READ07,READ08,READ16,READ17,READ24,READ25,
         LIST04,LIST05,LIST06,LIST08,LIST09,LIST12,LIST16,
         LIST17,LIST18,LIST19,LIST20,
         WSAM01,WSAM02,
         DRP01(1)DRP56,
         READ01,READ02,READ03,READ04,READ05,READ06,READ09,READ10,
         READ11(1)READ15,READ18(1)READ23,
         LIST01,LIST02,LIST03,LIST07,LIST10,LIST11,LIST13(1)LIST15,
         WRIT01(1)WRIT45);
>TEST01  TNAME=NAEP1,  INUMB=(1(1)82);
>TEST02  TNAME=LISTV,  INAME=(LIST01(1)LIST20);
>TEST03  TNAME=LISTM,  INAME=(LIST01(1)LIST20);
>TEST04  TNAME=NAEP2,  INUMB=(1(1)82);
>TEST05  TNAME=DRPV,   INAME=(DRP01(1)DRP56);
>TEST06  TNAME=DRPM,   INAME=(DRP01(1)DRP56);
>TEST07  TNAME=NAEP3,  INUMB=(1(1)82);
>TEST08  TNAME=READV,  INAME=(READ01(1)READ25);
>TEST09  TNAME=READM,  INAME=(READ01(1)READ25);
>TEST10  TNAME=NAEP4,  INUMB=(1(1)82);
>TEST11  TNAME=WRITEV, INAMB=(WRIT01(1)WRIT45,WSAM01,WSAM02);
>TEST12  TNAME=WRITEM, INAME=(WRIT01(1)WRIT45,WSAM01,WSAM02);
>FORM01  LEN=169, INUM=(83(1)230,10(1)14, 1(1)3,15(1)20, 4(1)6, 7, 8, 21,9);
>FORM02  LEN=168, INUM=(83(1)230, 1(1)3,22(1)28, 4(1)6,29(1)31, 7, 8, 9,32);
>FORM03  LEN=168, INUM=(83(1)230,22(1)28,62(1)64,29(1)31,65, 66(1)69,32,70);
>FORM04  LEN=169, INUM=(83(1)230,62(1)64,10(1)14,65, 15(1)20,66(1)69,70,21);
>FORM05  LEN=171, INUM=(83(1)230,10(1)14,22(1)28, 15(1)20,29(1)31,21,32);
>FORM06  LEN=166, INUM=(83(1)230, 1(1)3,62(1)64, 4(1)6,65, 7, 8,66(1)69,9,70);
>FORM07  LEN=170, INUM=(83(1)230,33(1)37,71(1)76,38(1)41,77(1)81, 42,82);
>FORM08  LEN=170, INUM=(83(1)230,71(1)76,43(1)46,77(1)81,47(1)51, 82,52);
>FORM09  LEN=167, INUM=(83(1)230,43(1)46,53(1)55,47(1)51,56(1)60, 52,61);
>FORM10  LEN=167, INUM=(83(1)230,53(1)55,33(1)37,38(1)41,56(1)60, 61,42);
>FORM11  LEN=168, INUM=(83(1)230,33(1)37,43(1)46,38(1)41,47(1)51, 42,52);
>FORM12  LEN=169, INUM=(83(1)230,71(1)76,53(1)55,77(1)81,56(1)60, 82,61);
>FORM13  LEN=170, INUM=(83(1)230,43(1)46,10(1)14,47(1)51,15(1)20, 52,21);
>FORM14  LEN=166, INUM=(83(1)230,53(1)55, 1(1)3, 4(1)6, 56(1)60, 7, 8,61, 9);
>FORM15  LEN=169, INUM=(83(1)230,22(1)28,33(1)37,29(1)31,38(1)41,32,42);
>FORM16  LEN=169, INUM=(83(1)230,62(1)64,71(1)76,65,77(1)81,66(1)69,70,82);
(10A1,2X,I2/15X,19A1,1X,129A1/15X,23A1)
>CALIB SELECT=(0(0)6);
>SCORE METHOD=2, NOPRINT, NQPT=(25(0)6);
```

Related topics

- CALIB command: SELECT keyword
- GLOBAL command: IFNAME keyword
- FORM command

- FORM command: INUMBERS keyword
- ITEMS command: INUMBERS keyword
- TEST command

6.13 Domain scores

This is an attempt to reconstruct the domain scores demonstration application reported in “The Domain Score Concept and IRT: Implications for Standards Setting” by Bock, Thissen & Zimowski (2001). We use the dataset **spell.dat**. All 100 items of the 100-word spelling test seem to be there, but there are only 660 records (instead of the 1,000 that Bock *et. al.* report). In a first run (**spell1.blm**), we CALIBRATEall 100 items and save the parameters in an external file. The syntax is shown below.

```
SPELL1.BLM - CALIBRATION OF THE 100 WORD SPELLING TEST
             TWO-PARAMETER MODEL
>COMMENTS
We are trying first to reproduce the table with slope and location parameters for the
100 words as Bock et al. report in "The Domain Score Concept and IRT:
          Implications for Standards Setting."

The SCORE command is included to obtain the percent correct for each examinee (= the
true domain scores).

>GLOBAL DFNAME='SPELL.DAT', NPARM=2, SAVE;
>SAVE   PARM='SPELL1.PAR';
>LENGTH NITEMS=(100);
>INPUT  NTOTAL=100, NIDCHAR=10, KFNAME='SPELL.DAT';
>ITEMS  INAMES=(S001(1)S100);
>TEST1  TNAME='SPELLING', INUMBERS=(1(1)100);
(10A1,1X,25A1,1X,25A1,1X,25A1,1X,25A1)
>CALIB  NQPT=31, CYCLES=100, CRIT=0.001, NOFLOAT;
>SCORE;
```

The item parameters of the first 5 items, as reported in the item parameter file **step1blm.par,** are shown in Table 6.2.

Table 6.2: Selected item parameters from step1blm.par

Item	Slope	S.E.	Threshold	S.E.
S001	0.79494	0.07978	-0.34466	0.06899
S002	0.38723	0.07299	-3.53823	0.61667
S003	0.24041	0.04784	-3.04033	0.61821
S004	0.72020	0.07353	-0.54159	0.08115
S005	0.69253	0.07367	-1.41137	0.13523

The parameter values are in close agreement with Table 1 from Bock. *et al*. (results for the first 5 items shown in Table 6.3 below), showing also that we have a correct dataset, with the items in the right order (of the table), albeit not all records.

Table 6.3: Selected item parameters from Bock *et. al.*

Item	Slope	Threshold
S001	0.843	-0.339
S002	0.351	-3.623
S003	0.239	-3.073
S004	0.785	0.727
S005	0.269	2.273

In a second run (**spell2.blm**), we let the program compute the expected domain scores for all 660 examinees from the saved parameter file. The DOMAIN and FILE keywords on the SCORE command are used. We skip the calibration phase with the SELECT keyword on the CALIB command. The scores are saved to file by using the SCORE keyword on the SAVE command.

The contents of **spell2.blm** are shown below. All the command files and data discussed here are available to the user in the **domscore** subfolder of the BILOG-MG installation folder.

```
SPELL2.BLM - CALIBRATION OF THE 100 WORD SPELLING TEST
             TWO-PARAMETER MODEL
>COMMENTS
In a second step, we test the "DOMAIN" keyword on the score command. The item parame-
ter file from the SPELL1.BLM analysis has been edited and saved as SPELL2.PAR in ac-
cordance with the FILE keyword format requirements. We save the score file.

>GLOBAL DFNAME='SPELL.DAT', NPARM=2, SAVE;
>SAVE   SCORE='SPELL2.SCO';
>LENGTH NITEMS=(100);
>INPUT  NTOTAL=100, NIDCHAR=10, KFNAME='SPELL.DAT';
>ITEMS  INAMES=(S001(1)S100);
>TEST1  TNAME='SPELLING', INUMBERS=(1(1)100);
(10A1,1X,25A1,1X,25A1,1X,25A1,1X,25A1)
>CALIB  SELECT=(0);
>SCORE  DOMAIN=100, FILE='SPELL2.PAR', METHOD=2;
```

The parameter file that we read in through the FILE keyword on the SCORE command had to be created from the saved parameter file (**spell1blm.par**) in the **spell1.blm** run. First we deleted everything before the first line with parameter estimates. Then we deleted all the columns that were not slope, threshold, or guessing parameters, leaving just those three columns and in that order. Then, we added a column with weights as the first column, in the same format. We used 1.0000, because we want all items weighed equally. We then added the variable format statement (4F10.5) as the first line in the file and renamed it to **spell1.par**.

The estimated domain scores from **spell2.blm** are fairly well recovered as **spell2.ph3** shows. Here are the results for the first five examinees:

```
GROUP   SUBJECT IDENTIFICATION            DOMAIN SCORE    S.E.   MARGINAL
WEIGHT   TEST      TRIED  RIGHT  PERCENT     ABILITY      S.E.     PROB
-------------------------------------------------------------------------
 1  01021119001                             64.89      4.92   |
 1.00 * SPELLING    100     65    65.00 |   -0.1501    0.4187 |  0.000000
 1  01041122001                             57.14      5.43   |
 1.00 * SPELLING    100     56    56.00 |   -0.7839    0.4321 |  0.000000
 1  01051219001                             54.25      5.40   |
 1.00 * SPELLING    100     57    57.00 |   -1.0132    0.4269 |  0.000000
 1  01061219001                             71.52      1.80   |
 1.00 * SPELLING    100     69    69.00 |    0.4499    0.1768 |  0.000000
 1  01071219001                             80.77      2.68   |
 1.00 * SPELLING    100     81    81.00 |    1.5475    0.4000 |  0.000000
-------------------------------------------------------------------------
```

If the estimated expected domain scores are not close, something is probably wrong, so this is a good test.

In a third and final step (**step3.blm**), we take a random sample of 20 items, adapt the parameter file (**spell3.par** as described previously) and produce a new score file (**spell3.sco**).

The contents of **spell3.blm** are as follows:

```
SPELL3.BLM - CALIBRATION OF THE 100 WORD SPELLING TEST
             TWO-PARAMETER MODEL
>COMMENTS
In this third step we use a random sample of 20 items from the 100-word spelling test
to score the examinees with the item parameters from the first step. The score file is
saved.

>GLOBAL DFNAME='SPELL.DAT', NPARM=2, SAVE;
>SAVE   SCORE='SPELL3.SCO';
>LENGTH NITEMS=20;
>INPUT  NTOTAL=100, NIDCHAR=10, KFNAME='SPELL.DAT';
>ITEMS  INAMES=(S001(1)S100);
>TEST1  TNAME='SPELLING', INUMBERS=(4, 9, 10, 13, 22, 26, 36, 51, 55, 65,
        69, 73, 74, 82, 83, 88, 89, 91, 94, 97);
(10A1,1X,25A1,1X,25A1,1X,25A1,1X,25A1)
>CALIB  SELECT=(0);
>SCORE DOMAIN=20, FILE='SPELL3.PAR', METHOD=2;
```

These are the results for the first five examinees:

```
GROUP   SUBJECT IDENTIFICATION            DOMAIN SCORE    S.E.   MARGINAL
WEIGHT   TEST      TRIED  RIGHT  PERCENT     ABILITY      S.E.     PROB
-------------------------------------------------------------------------
 1  01021119001                             72.51      6.95   |
 1.00 * SPELLING     20     14    70.00 |    0.4109    0.7214 |  0.000000
 1  01041122001                             62.76      7.94   |
 1.00 * SPELLING     20     12    60.00 |   -0.4985    0.6850 |  0.000000
 1  01051219001                             63.42      7.90   |
 1.00 * SPELLING     20     12    60.00 |   -0.4414    0.6870 |  0.000000
 1  01061219001                             74.25      6.61   |
 1.00 * SPELLING     20     14    70.00 |    0.5971    0.7300 |  0.000000
```

```
 1  01071219001                             79.14      5.32   |
 1.00 * SPELLING     20     17    85.00 |    1.2047    0.7604 |  0.000000
--------------------------------------------------------------------------
```

As can be seen, a decent recovery of the “population domain scores” with the random sample of only 20 items.

Related topics

- CALIB command: SELECT keyword
- SAVE command: SCORE keyword
- SCORE command: DOMAIN keyword
- SCORE command: FILE keyword

References

Aitchison, J., & Silvey, S. D. (1960). Maximum-likelihood estimation procedures and associated tests of significance. *Journal of the Royal Statistical Society, Series B*, **22**, 154-171.

Andersen, E. B. (1973). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society, Series B*, **32**, 283-301.

Andersen, E. B., & Madsen, M. (1977). Estimating the parameters of a latent population distribution. *Psychometrika*, **42**, 357-374.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, **43**, 561-573.

Baker, F. B. (1992). Item response theory: parameter estimation techniques. Reading, NY: Marcel Dekker.

Bartholomew, D. J. (1980). Factor analysis for categorical data. *Journal of the Royal Statistical Society, Series B*, **42**, 293-321.

Bergan, J. R., & Stone, C. A. (1985). Latent class models for knowledge domains. *Psychological Bulletin*, **98**, 166-184.

Binet, A., & Simon, T. (1905). Methods nouvelles pour le diagnostic du niveau intellectuel des anormaux. *Année Psychologique*, **11**, 191-244.

Birnbaum, A. (1957). Efficient design and use of tests of a mental ability for various decision making problems. *Series Report No. 58-16*. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas.

Birnbaum, A. (1958a). On the estimation of mental ability. *Series Report No. 15*. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas.

Birnbaum, A. (1958b). Further considerations of efficiency in tests of a mental ability. *Technical Report No. 17*. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas.

Birnbaum, A. (1967). Statistical theory for logistic mental test models with a prior distribution of ability. *Research Bulletin*, No. 67-12. Princeton, NJ: Educational Testing Service.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & R. M. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Bliss, C. I. (1935). The calculation of the dosage mortality curve (Appendix by R. A. Fisher). *Annals of Applied Biology*, **22**, 134-167.

Bock, R. D. (1966). Estimating multinomial response relations. *Research Memorandum*, No. 5. Chicago: University of Chicago Educational Statistics Laboratory.

Bock, R. D. (1970). Estimating multinomial response relations. In R. C. Bose, *et al*. (Eds.), *Contributions to statistics and probability*. Chapel Hill, NC: University of North Carolina Press, 111-132.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, **37**, 29-51.

Bock, R. D. (1975). *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill, 2nd edition, 1985, Scientific Software International, Chicago.

Bock, R. D. (1976). Basic issues in the measurement of change. In D. N. M. de Gruijter & L. J. T. van der Kamp (Eds.), *Advances in Psychological and Educational Masurement*. London: Wiley & Sons, 75-76.

Bock, R. D. (1983a). Within-subject experimentation in psychiatric research. In R. D. Gibbons & Dysken (Eds.), *Statistical and methodological advances in psychiatric research*. New York: SP Medical & Scientific Books, 59-90.

Bock, R. D. (1983b). The mental growth curve re-examined. In D. Weiss (Ed.), *New horizons in testing*. New York: Academic Press, 205-219.

Bock, R. D. (1983c). The discrete Bayesian. In H. Wainer & S. Messick (Eds.), *Principles of psychometrics*. Hillsdale, NJ: Erlbaum, 103-115.

Bock, R. D. (1989). Measurement of human variation: a two-stage model. In R. D. Bock (Ed.), *Multilevel analysis of educational data*. New York: Academic Press, 319-342.

Bock, R. D. (1993). Different DIFs. In: P. W. Holland & H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Erlbaum, 115-122.

Bock, R. D. (1997). The nominal categories model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*. New York: Springer Verlag, 33-65.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*, **46**, 443-445.

Bock, R. D., Gibbons, R. D., & Muraki, E. (1988). Full information item factor analysis. *Applied Psychological Measurement*, **12**, 261-280.

Bock, R. D., & Jones, L. V. (1968). *The measurement and prediction of judgment and choice*. San Francisco: Holden-Day.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika*, **35**, 179-197.

Bock, R. D., & Mislevy, R. J. (1981). An item response model for matrix-sampling data: the California Grade Three Assessment. In D. Carlson (Ed.), *Testing in the states: beyond accountability*. San Francisco: Jossey-Bass, 65-90.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, **6**, 431-444.

Bock, R. D., Muraki, E., & Pfiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, **25**, 275-285.

Bock, R. D., Thissen, D., & Zimowski, M. F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement* **34**, 197-211.

Bock, R. D., Wolfe, R., & Fisher, T. H. (1996). *A Review and analysis of the Tennessee Value-Added Assessment System*. Nashville, TN: Office of Education Accountability, State of Tennessee, Comptroller of the Treasury.

Bock, R. D., & Zimowski, M. F. (1989). *Duplex Design giving students a stake in educational assessment.* Chicago: Methodology Research Center NORC.

Bock, R. D., & Zimowski, M. F. (1995). Multiple group IRT. In W. van der Linden & R. Hambleton (Eds.), *Handbook of item response theory*. New York: Springer-Verlag.

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*. New York: Springer Verlag, 433-448.

Bock, R. D., & Zimowski, M. F. (1998). *Feasibility Studies of Two-Stage Testing in Large-Scale Educational Assessment: Implications for NAEP*, 34-41. Commissioned by the NAEP Validity Studies (NVS) Panel. May 1998.

Bock, R. D., & Zimowski, M. F., (1999). *Application of disattenuation analysis to correlations between matrix-sample assessment results and achievement test scores*. Addendum to D. H. McLaughlin, R . D. Bock, E. A. Arenson & M. F. Zimowski. Palo Alto, CA: American Institutes for Research.

Bowers, J. (1972). A note on comparing *r*-biserial and *r*-point biserial. *Educational and Psychological Measurement*, **32**, 771-775.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs. I. Method of paired comparisons. *Biometrika*, **39**, 324-345.

Browne, M. W., & du Toit, S. H. C. (1992). Automated fitting of nonstandard models. *Multivariate Behavioral Research*, **27**, 269-300.

Burt, C. (1921). *Mental and scholastic tests*. London: P. S. King & Son.

Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika*, **10**, 1-19.

Clogg, C. C. (1979). Some latent structure models for the analysis of Likert-type data. *Social Science Research*, **8**, 287-301.

Clogg, C. C., & Goodman, L. A. (1984). Latent structure analysis of a set of multi-dimensional contingency tables. *Journal of the American Statistical Association*, **79**, 762-661.

Cronbach, L. J., Glaser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Davis, J. A. (1975). Codebook for the Spring 1976 General Social Survey. Chicago: NORC.

De Leeuw, J., & Verhelst, N. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational Statistics*, **11**, 193-196.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1-38.

Digvi, D. R. (1979). Calculation of the tetrachoric correlation coefficient. *Psychometrika*, **44**, 169-172.

Fechner, G. T. (1860). *Elemente der Psychophysik*. Leipzig: Breitkopf & Hartel.

Finney, D. J. (1952). Probit analysis: A statistical treatment of the sigmoid response curve, 2nd ed. London: Cambridge University Press.

Fisher, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, **22**, 699-725.

Fisher, R. A., & Yates, F. (1938). Statistical tables for biological, agricultural and medical research. New York: Hafner.

Follman, D. (1988). Consistent estimation in the Rasch model based on nonparametric margins. *Psychometrika*, **53**, 553-562.

French, J. L., & Hale, R. L. (1990). A history of the development of psychological and educational testing. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children*. New York: Guilford Press, 3-28.

Gibbons, R. D., & Hedeker, D. R. (1992). Full information item bi-factor analysis. *Psychometrika*, **57**, 423-436.

Glas, C. A. W. (1996). Detection of differential item functioning using Lagrange Multiplier Tests. *Research Report*, No. 96-02. Enchede: University of Twente, Faculty of Educational Science and Technology.

Glass, G. V., & Stanley, J. C. (1970). *Statistical Methods in Education and Psychology*. Englewood Cliffs, NJ: Prentice-Hall.

Goldstein, H. (1983). Measuring changes in educational attainment over time. *Journal of Educational Measurement*, **20**, 369-377.

Green, B. F. (1951). A general solution for the latent class model of latent structure analysis. *Psychometrika*, **16**, 151-166.

Green, B. F. (1952). Latent structure analysis and its relation to factor analysis. *Journal of the American Statistical Association*, **47**, 71-76.

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement*, **37**, 827-838.

Guilford, J. P. (1954). *Psychometric Methods*. (2nd Ed.) New York: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Dekker.

Haberman, S. J. (1977). Log-linear models and frequency tables with small expected cell counts. *Annals of Statistics*, **5**, 1148-1169.

Haberman, S. J. (1979). Analysis of qualitative data, Vol. 2. New developments. New York: Academic Press.

Hambleton, R. K., & Jurgensen, C. (1990). Criterion referenced assessment of school achievement. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children*. New York: Guilford Press, 456-477.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory*. *Principles and applications*. Boston: Kluwer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Harman, H. H. (1976). *Modern Factor Analysis*. Chicago: The University of Chicago Press.

Harvey, W. R. (1970). Estimation of variance and covariance components in the mixed model. *Biometrics*, **26**, 485-504.

Harwell, M. R., Baker, F. B., & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and an EM algorithm: A didactic. *Journal of Educational Statistics*, **13**, 243-271.

Hendrickson, E. A., & White, P. O. (1964). Promax: A quick method for rotation to oblique simple structure. *British Journal of Mathematical and Statistical Psychology*, **17**, 65.

Henryssen, S. (1971). Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), *Educational Measurement, 2nd ed.,* Washington, DC: American Council on Education.

Hively, W. (1974). *Domain-referenced testing*. Englewood Cliffs, NJ: Educational Technology Publications.

Holland, P. W., & Rubin, D. B. (Eds.) (1982). *Test equating*. Hillsdale, NJ: Erlbaum.

Holland, P. W., & Wainer, H. (1993). *Differential Item Functioning*. Hillsdale, NJ: Erlbaum.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, **2**, 41-54.

Horst, P. (1933). The difficulty of a multiple-choice test item. *Journal of Educational Psychology*, **24**, 229-232.

Irving, L. M. (1987). Mirror images: Effects of the standard of beauty on women's self and body esteem. Unpublished Masters Thesis, University of Kansas.

Jenkins, C.D., Rosenman, R.H. & Zyzanski, S.J. (1972). *The Jenkins Activity Survey of Health Prediction*. New York: The Psychological Corporation.

Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika*, **36**, 149-176.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's Reference Guide*. Chicago: Scientific Software International, Inc.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, **23**, 187-200.

Kendall, M., & Stuart, A. (1961), *Inference and Relationship*, Vol. 2 of The Advanced Theory of Statistics, first ed. London: Charles Griffin & Company.

Kelley, T. L. (1947). *Fundamentals of statistics*. Cambridge, MA: Harvard University Press.

Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, **27**, 887.

Klassen, D., & O'Connor, W. A. (1989). Assessing the risk of violence in released mental patients: A cross-validation study. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, **1**, 75-81.

Kolakowski, D., & Bock, R. D. (1981). A multivariate generalization of probit analysis. *Biometrics*, **37**, 541-551.

Lawley, D. N. (1943). On problems connected with time selection and test construction. *Proceedings of the Royal Society of Edinburgh*, **61A**, 273-287.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star & J. A. Clausen, *Measurement and prediction*. Princeton, NJ: Princeton University Press, 362-412.

Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh*, **61**, 273-287.

Likert, R. (1932). A technique for the measurement of attitude. *Archives of Psychology*, 140.

Linacre, J. M., & Wright, W. D. (1993). *FACETS: Many-facet Rasch analysis with FACFORM Data Formatter*. Chicago: MESA Press.

Linn, R. L., & Hambleton, R. K. (1991). Customized sets and customized norms. *Applied Measurement in Education*, **4**, 185-207.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Longford, N. T. (1989). Fisher scoring algorithm for variance component analysis of data with multilevel structure. In R. D. Bock (Ed.), *Multilevel analysis of educational data*. San Diego: Academic Press, 297-310.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph*, No. 7.

Lord, F. M. (1953). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika*, **18**, 57-76.

Lord, F. M. (1971). A theoretical study of two-stage testing. *Psychometrika*, **36**, 227-242.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores* (with contributions by A. Birnbaum). Reading, MA: Addison-Wesley.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **44**, 226-233.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

Mantel, N. (1966). Models for complex contingency tables and polytomous dosage response curves. *Biometrics*, **22**, 83-95.

Marshall, J. C., & Hales, L. W. (1972). *Essentials of Testing*. Reading, MA: Addison-Wesley.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, **47**, 149-174.

Masters, G. N. (1985). A comparison of latent trait and latent class analyses of Likert-type data. *Psychometrika*, **50**, 69-82.

Meng, X. L., & Schilling, S. (1996). Fitting full information factor models and an empirical investigation of bridge sampling. *Journal of the American Statistical Association*, **91**, 1254-1267.

Mislevy, R. J. (1983). Item response models for grouped data. *Journal of Educational Statistics*, **8**, 271-288.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, **49**, 359-381.

Mislevy, R. J. (1986). Bayesian modal estimate in item response models. *Psychometrika*, **51**, 177-195.

Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement*, **11**, 81-91.

Mislevy, R. J., & Bock, R. D. (1982). Biweight estimates of latent ability. *Journal of Educational and Psychological Measurement*, **42**, 725-737.

Mislevy, R. J., & Bock, R. D. (1983). BILOG: Analysis and scoring of binary items and one-, two-, and three-parameter logistic models. Chicago: Scientific Software International, Inc.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item Analysis and Test Scoring with Binary Logistic Models*. Chicago: Scientific Software International, Inc.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, **17**, 131-154.

Naylor, J. C., & Smith, A. F. M. (1982). Applications of a method for the efficient computation of posterior distributions. *Applied Statistics*, **31**, 214-225.

Neyman, J. A., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrika*, **16**, 1-22.

Olsson, U., Drasgow, F., & Dorans, N. J. (1982). The polyserial correlation coefficient. *Psychometrika*, **47**(**3**), 337-347.

Owen, R. J. (1969). A Bayesian approach to tailored testing. *Research Bulletin* No. 69-92. Princeton, NJ: Educational Testing Service.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational Measurement (3rd edition)*. New York: American Council on Education-Macmillan, 221-262.

Ramsey, J. O. (1975). Solving implicit equations in psychometric data analysis. *Psychometrika*, **40**, 337-360.

Rasch, G. (1960; reprinted 1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, **4**, 321-324.

Richardson, M. W. (1936). The relationship between difficulty and the differential validity of a test. *Psychometrika*, **1**, 33-49.

Roche, A. F., Wainer, H., & Thissen, D. (1975). *Skeletal Maturity: The knee joint as a biological indicator*. New York: Plenum.

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, No. 17.

Samejima, F. (1972). A general model for free-response data. *Psychometrika Monograph Supplement*, No. 18.

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika*, **39**, 111-121.

Samejima, F. (1979). A new family of models for the multiple-choice item. *Research Report*, No. 79-4, Department of Psychology, University of Tennessee.

Schilling, S. (1993). *Advances in Full Information Item Factor Analysis using the Gibbs Sampler*. Unpublished doctoral dissertation, University of Chicago.

Schilling, S. G., & Bock, R. D. (1999). High-dimensional maximum marginal likelihood item factor analysis. (In press.)

Schultz, M. E., & Nicewander, W. A. (1997). Grade equivalent and IRT representations of growth. *Journal of Educational Measurement*, **34**(**4**), 315-332..

Smith, M. C., & Thelen, M. H. (1984). Development and validation of a test for bulimia. *Journal of Consulting and Clinical Psychology,* **52**, 863-872.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, **15**, 201-293.

Stouffer, S. A., & Toby, J. (1951). Role conflict and personality. *American Journal of Sociology*, **56**, 395-406.

Stroud, A. H., & Secrest, D. (1966). *Gaussian Quadrature Formulas*. Englewood Cliffs, NJ: Prentice-Hall.

Symonds, P. M. (1929). Choice of items for a test on the basis of difficulty. *Journal of Educational Psychology*, **20**, 481-493.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika*, **47**, 175-186.

Thissen, D., & Steinberg, L. (1984). A response model for multiple-choice items. *Psychometrika*, **49**, 501-519.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, **51**, 566-577.

Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin*, **104**, 385-395.

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement*, **26**, 161-176.

Thissen, D., Steinberg, L., & Money, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, **26**, 247-260.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of DIF using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Erlbaum, 67-113.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, **47**, 397-412.

Thorndike, E. L. (1927). *The measurement of intelligence*. New York: Teacher's College, Columbia University.

Thorndike, R. L. (1982). *Applied Psychometrics*. Boston: Houghton-Mifflin.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, **16**, 433-451.

Thurstone, L. L. (1930). The learning function. *Journal of General Psychology*, **3**, 469-493.

Thurstone, L. L. (1947). *Multiple-factor analysis*. Chicago: University of Chicago Press.

Tsutakawa, R. K. (1992). Prior distribution for item response curves. *British Journal of Mathematical and Statistical Psychology*, **45**, 51-71.

Tsutakawa, R. K., & Lin, H. Y. (1986). Bayesian estimation of item response curves. *Psychometrika*, **51**, 251-267.

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika*, **11**, 1-13.

Urban, F. M. (1908). The application of statistical methods to the problems of psychophysics. Philadelphia: Psychological Clinic Press.

Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Verhulst, P-F. (1844). Recherches mathematiques sur la loi d'accrossement de population. *Memoires de l'Academe royal de Belgique*, Volume 18.

Wainer, H. (Ed.) (1990). *Computerized adaptive testing: a primer*. Hillsdale, NJ: Erlbaum.

Wainer, H. (1995). Precision and differential item functioning on a testlet based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education*, **8**, 157-187.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, **24**, 185-201.

Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, **54**, 427-450.

Wilmut, J. (1975). Objective test analysis: some criteria for item selection. *Research in Education*, **13**, 27-56.

Wood, R. (1977). Inhibiting blind guessing: the effect of instructions. *Journal of Educational Measurement*, **13**, 297-307.

Zimowski, M. F. (1985). *Attributes of spatial test items that influence cognitive processing*. Unpublished doctoral dissertation, University of Chicago.

Zimowski, M. F., & Bock, R. D. (1987). Full information item factor analysis of test forms from the ASVAB CAT pool. Chicago: NORC.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items*. Chicago: Scientific Software International, Inc.

Zwick, R. (1987). Assessing the dimensionality of NAEP reading data. *Journal of Educational Measurement*, **24**, 293-308.