TABLE OF CONTENTS

1 Fixed theta estimation

Note that although this feature is not available in IRTPRO 2.1 or IRTPRO 3, it has been implemented in IRTPRO 4.

The EXTERNAL option of the INPUT command allows calibration of item parameters from data records with given test scores of the respondents. See the BILOGMG guide for more information describing this feature.

2 Posterior weights

The **PDISTRIB** keyword allows the user to save the points and weights of the posterior latent distribution at the end of the calibration phase. These quantities can be included as prior values following the `SCORE` command for later EAP estimation of ability from previously estimated item parameters.

3 Drift analysis

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 below). 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.

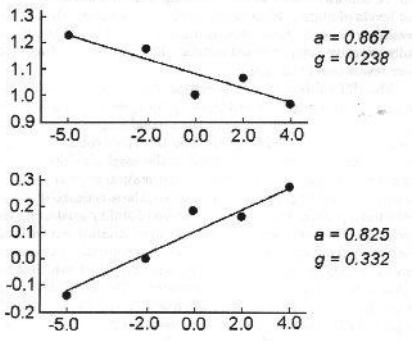

Figure: Drift of the location parameters of two items from a College Board Advanced Placement Examination in Physics

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, the relative difficulties of the items are allowed to differ from one group to another or one occasion to another.

When an item parameter drift (DRIFT) analysis is selected, the program provides estimates of the coefficients of the linear or polynomial function. Consult the BILOGMG guide for an illustration of a drift analysis.

5 Equivalent groups equating

See Example 4 in the BILOGMG guide for an illustration.

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.

The method of carrying out equivalent groups equating is somewhat different, 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.

6 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.

7 Vertical equating

See Example 5 in the BILOGMG guide for an illustration.

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*.

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, nonequivalent 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 input records 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 input records 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.

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".

8 Group-wise adaptive testing

See Example 8 in the BILOGMG guide for an illustration.

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 assessments see Bock & Zimowski (1989).

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.

9 Variant items

See Example 7 in the BilogMG guide for an illustration

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.

10 Parallel-form correlations

See Example 11 for the commands required.

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.*

Unlike the individual-level analysis, the aggregate-level permits a rigorous test of fit of the response pattern for the group.

The starting values computed in the input phase 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.

11 Estimating and scoring tests of increasing length

In Example 10 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`.