An NCME Instructional Module on

Using Statistical Procedures to Identify Differentially Functioning Test Items

Brian E. Clauser, *National Board of Medical Examiners*
Kathleen M. Mazor, *University of Massachusetts School of Medicine*

This module is intended to prepare the reader to use statistical procedures to detect differentially functioning test items. To provide background, differential item functioning (DIF) is distinguished from item and test bias, and the importance of DIF screening within the overall test development process is discussed. The Mantel-Haenszel statistic, logistic regression, SIBTEST, the Standardization procedure, and various IRT-based approaches are presented. For each of these procedures, the theoretical framework is presented, the relative strengths and weaknesses of the method are highlighted, and guidance is provided for interpretation of the resulting statistical indices. Numerous technical decisions are required in order for the practitioner to appropriately implement these procedures. These decisions are discussed in some detail, as are the policy decisions necessary to implement an operational DIF detection program. The module also includes an annotated bibliography and a self-test.

Test results are routinely used as the basis for decisions regarding placement, advancement, and licensure. These decisions have important personal, social, and political ramifications. It is crucial that the tests used for these decisions allow for valid interpretations. One potential threat to validity is *item bias.*[1] When a test item unfairly favors one group over another, it can be said to be biased. Such items exhibit *differential item functioning (DIF),* a necessary but not a sufficient condition for item bias.

Differential item functioning is present when examinees from different groups have differing probabilities or likelihoods of success on an item, after they have been matched on the ability of interest. The last clause of this definition, requiring that differences exist after matching on the ability of interest, is essential. It implies that differences in performance, in and of themselves, are not evidence of bias. In some circumstances, examinees from different groups may in fact differ in ability, in which case differences in performance are to be expected. This result is referred to as *item impact,* rather than item bias.

Unfortunately, matching examinees on the ability of interest is not a trivial task. Test items are often developed to measure complex skills delineated in content specifications. It may be difficult to identify a matching criterion that represents the specific skill (or set of skills) that an item has been developed to measure. If differences are found after conditioning on a particular ability, then performance on that item depends on some ability other than that which has been taken into account. The question then becomes whether that second ability is relevant to the purpose of testing. That is, is the additional ability of interest, or does its presence represent a nuisance?

As an example, consider a mathematics word problem for which the correct solution depends on both the ability to perform calculations and reading comprehension. If examinees are matched only on the ability to perform calculations, and one group is less proficient than the other in reading comprehension, between-group differences in performance, evidenced as DIF, are likely. However, the item may not be biased. Whether or not such an item is judged to be biased will depend on whether reading comprehension is considered a relevant ability with respect to the purpose of the test. Thus, DIF is a necessary, but not sufficient, condition for item bias.

It is important to differentiate both item bias and DIF from inappropriate (potentially offensive) item content. DIF analyses do not ensure that item content is appropriate. Testing organizations often use panels of experts to review items with the goal of eliminating material that may involve gender or ethnically based stereotyping or be otherwise offensive to

Brian E. Clauser is a Senior Psychometrician at the National Board of Medical Examiners, 3750 Market St., Philadelphia, PA 19104. His specializations are applied measurement and psychometric methods.

Kathleen M. Mazor is a Co-Director of Research and Evaluation, Office of Medical Education at the University of Massachusetts, School of Medicine, 55 Lake Ave. N., Worcester, MA 01655. Her specializations are applied measurement and psychometric methods.

Series Information

ITEMS is a series of units designed to facilitate instruction in educational measurement. These units are published by the National Council on Measurement in Education. Professor George Engelhard, Jr., Emory University, has served as editor for this module. Information regarding the development of new ITEMS modules should be addressed to: Dr. Michael Zieky, Educational Testing Service, Mail Stop 16-C, Rosedale Rd., Princeton, NJ 08541.

minority examinees (Ramsey, 1993). Sensitivity reviews are separate and distinct from DIF analyses—both are important, and neither can substitute for the other.

DIF analyses are also separate from other validation studies. DIF procedures are designed to identify individual items that function differentially, relative to some identified criterion. If all items advantage one group over the other, DIF procedures, using total test score as the criterion, will be ineffective (Camilli, 1993). Meaningful interpretation of DIF statistics, therefore, presupposes appropriate construct and predictive validity evidence.

This article focuses on statistical techniques for identifying differentially functioning items. DIF analyses are only one step in the overall test development process and must be interpreted within that context. DIF analyses do not lend themselves to a cookbook approach. Most of the steps require judgment, and most require consideration of other aspects of the test development process. The decisions made at each step will be driven by practical considerations, knowledge of the test content, purpose, examinee population, empirical results, technical knowledge of the strengths and limitations of various statistical DIF indices, and political pressures. The test developer will need to identify which groups of examinees will be compared. An appropriate matching criterion (or criteria) must be identified. One (or more) DIF statistic must be selected for the analysis. Regardless of which statistical procedure is used, numerous technical issues will need to be considered before the analysis is implemented. Finally, the results of the analysis must be interpreted, and decisions must be made regarding the final test content, or scoring. Again, interpretations of results must be placed within the context of the overall test development process. The results of DIF analysis are one form of validity evidence. As with other validation efforts, the focus is not on the test per se but on the application being made and on the associated interpretation of the results (Cronbach, 1988; Messick, 1988).

The remainder of this module is presented in seven sections. The first describes the most commonly used DIF statistics, noting strengths and limitations. The second section describes implementation decisions. The third section provides guidance for interpreting the output from DIF detection software. The fourth section discusses the policy decisions that will be necessary both to carry out the analysis and interpret the results. The fifth section is a glossary of terms. The sixth section is a self-assessment. The final section provides an annotated bibliography including sources of software and recommended readings.

Statistical Procedures

Literally dozens of DIF screening procedures have been described in the literature. Based on theoretical strengths and the results of numerous empirical comparisons, a relatively small number of these methods have emerged as preferred. All of these approaches provide for comparison of performance on a studied item after matching examinees on the ability of interest. A brief description of each of these follows.

IRT Methods

Methods based on item response theory (IRT) provide a useful theoretical framework for DIF because between-group differences in the item parameters for the specific model can be used to model DIF. There is no single IRT method—numerous approaches are based on a range of IRT models. They share the use of a matching variable which is an estimate of latent ability rather than the observed score. All the various IRT methods conceptualize DIF in terms of differences in the model parameters for the comparison groups. The general framework involves estimating item parameters separately for the *reference* and *focal groups*. After placing them on the same scale, differences between the item parameters for the two groups can then be compared. When the parameters are identical for the two groups, the item does not display DIF. In the absence of DIF, the focal and reference group item characteristic curves (ICCs), showing the probability of a correct response as a function of examinee ability, will be coincident (see Figure 1). In the simplest case of DIF, items may differ

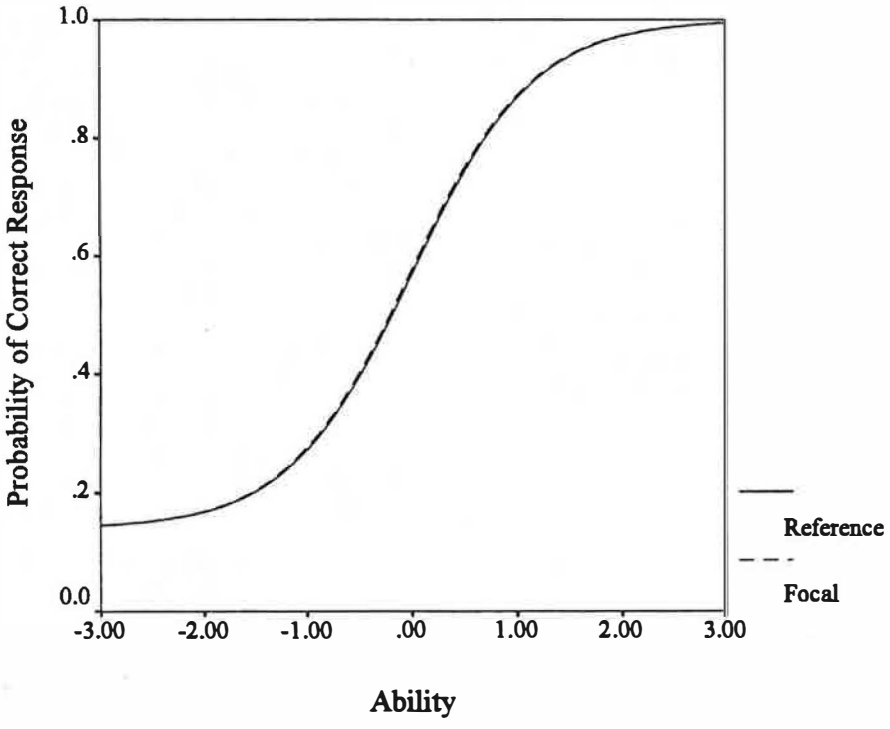

FIGURE 1. *ICCs for the focal and reference groups for an item that displays no DIF*

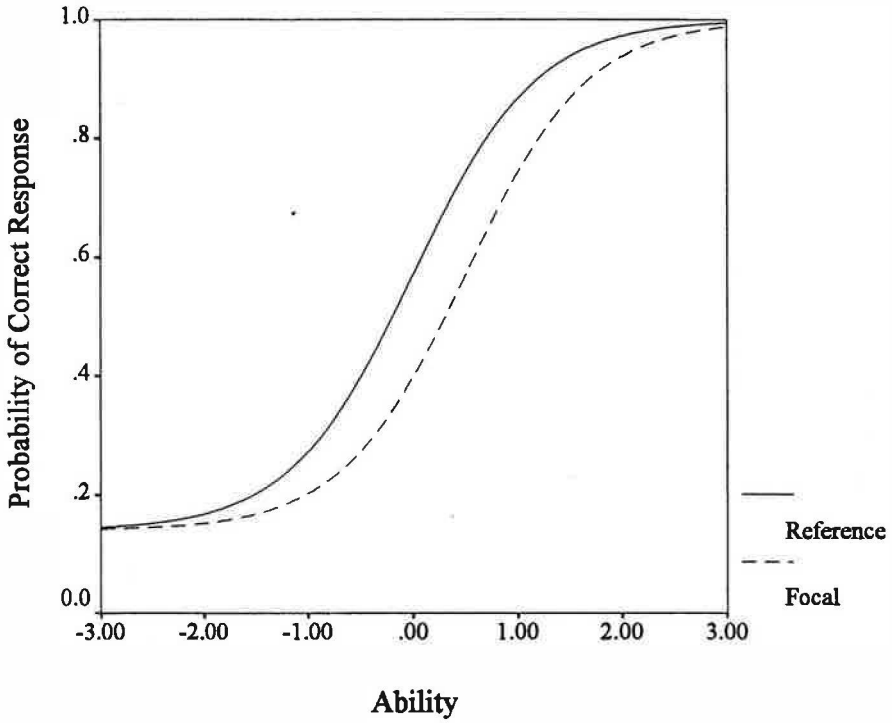

FIGURE 2. *ICCs for the focal and reference groups for an item that displays DIF, resulting from a difference in item difficulty parameters of 0.5*

across groups solely in terms of difficulty (see Figures 2 and 3). Alternatively, items may differ across groups in terms of any or all of the a, b, and c parameters (discrimination, difficulty, and pseudo-guessing). Figure 4 shows an example in which there are group differences in the discrimination parameter. Figure 5 provides an example of DIF resulting from differences across multiple parameters.

Numerous approaches are available for identifying the presence of between-group differences in item parameters. Estimates of effect size and/or statistical significance can be made based on comparison of item parameters across groups (Linacre & Wright, 1986; Lord, 1980); differences between both the difficulty and discrimination parameters can be quantified by estimating the area between the ICCs for the

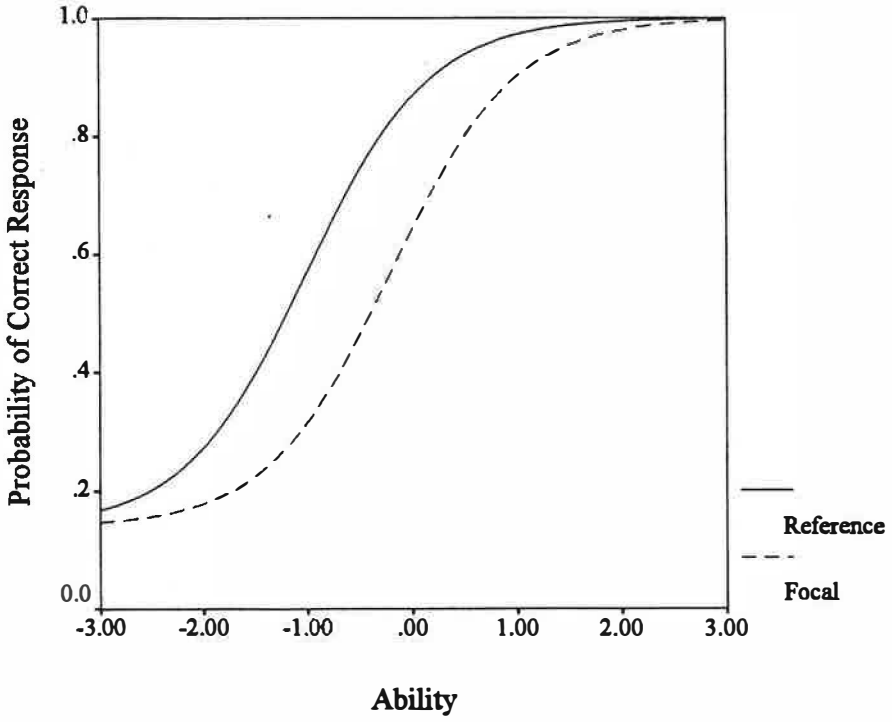

FIGURE 3. *ICCs for the focal and reference groups for an item that displays DIF, resulting from a difference in item difficulty parameters of 0.8*

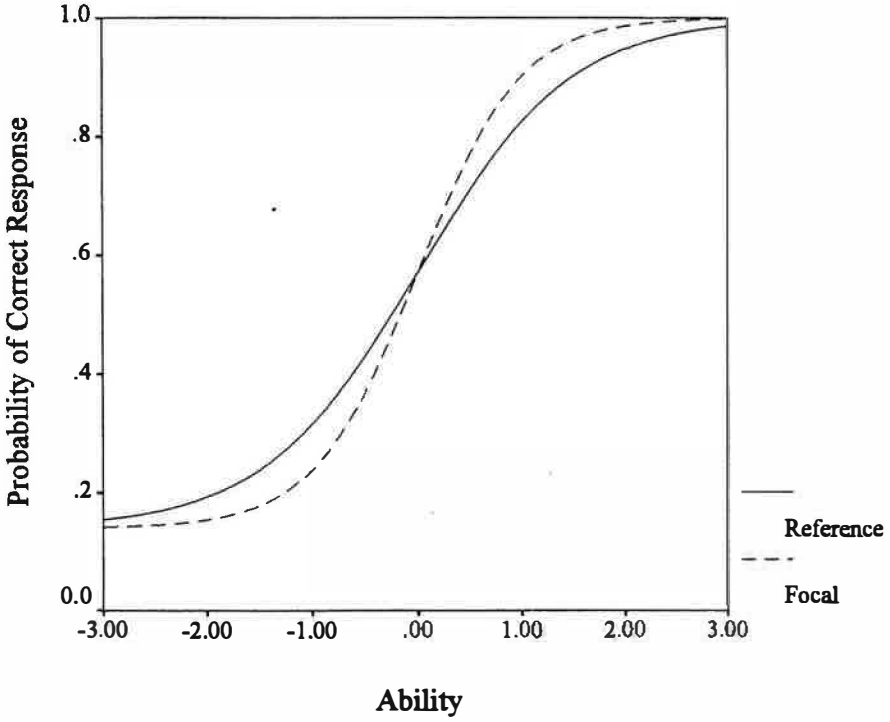

FIGURE 4. *ICCs for the focal and reference groups for an item that displays DIF, resulting from a difference in item discrimination parameters of 0.4*

two groups (Raju, 1988); or improvement in fit for the model can be tested, comparing fit with and without separate group parameter estimates (Thissen, Steinberg, & Wainer, 1993). The limitation of IRT methods is that the data must meet the strong (unidimensionality) assumption of the models. These methods also require large examinee samples for accurate parameter estimation if the two- or three-parameter model is used. Obviously, researchers choosing this approach will also need a working understanding of the required IRT model(s) and associated parameter estimation procedures.

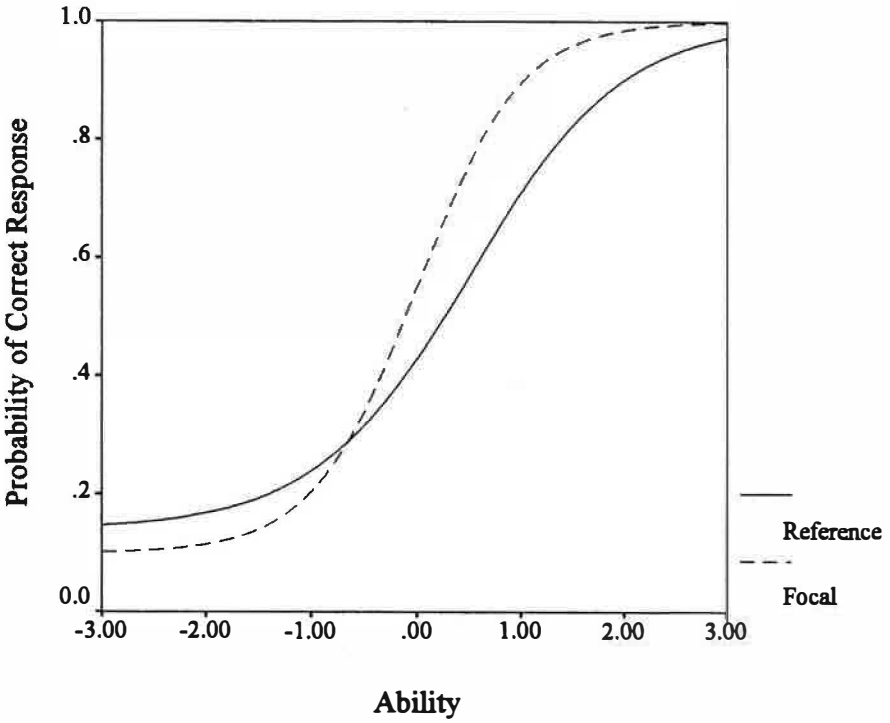

FIGURE 5. *ICCs for the focal and reference groups for an item that displays DIF, resulting from differences in item difficulty, discrimination, and pseudo-guessing parameters*

Mantel-Haenszel Statistic

As an alternative to the IRT methods, several approaches have been suggested based on analysis of contingency tables. These methods differ from IRT approaches in that examinees are typically matched on an observed variable (such as total test score), and then counts of examinees in the focal and reference groups getting the studied item correct or incorrect are compared. The Mantel-Haenszel statistic may be the most widely used of the contingency table procedures and has been the object of considerable evaluation since it was first recommended by Holland and Thayer (1988). With this method, the relevant comparison is implemented in terms of the likelihood of success on the item for members of the two groups (matched on ability). The resulting index is based on the ratio of these likelihoods. In addition to this measure of effect size, the statistic has an associated test of significance, distributed as a chi-square. The method has been shown to be effective with reasonably small examinee samples (e.g., 200 examinees per group). It is highly efficient both in terms of statistical power and computational requirements.

The procedure is implemented by first dividing examinees into levels based on ability. Typically, the total test score is used for matching examinees. A 2 × 2 table, with the following configuration, is formed for each level of the matching criterion:

	Score on Studied Item		
Group	1	0	Total
Reference	A_j	B_j	N_{rj}
Focal	C_j	D_j	N_{fj}
Total	M_{1j}	M_{0j}	T_j

The Mantel-Haenszel statistic then tests the H_0 against the alternative

$$H_1: \frac{P_{rj}}{Q_{rj}} = \alpha \frac{P_{fj}}{Q_{fj}} \qquad j = 1, 2, \ldots, k$$

where $\alpha \neq 1$ and k is the number of levels of the matching criterion. The following formula estimates α,

$$\hat{\alpha}_{MH} = \frac{\sum_j A_j D_j / T_j}{\sum_j B_j C_j / T_j}.$$

The MHχ^2 takes the form

$$MH\chi^2 = \frac{\left(\left|\sum_j A_j - \sum_j E(A_j)\right| - \frac{1}{2}\right)^2}{\sum_j \text{var}(A_j)}$$

$$\text{var}(A_j) = \frac{N_{rj} N_{fj} M_{1j} M_{0j}}{T_j^2 (T_j - 1)}.$$

Software is available as part of commercially available statistical packages. Specialized software for DIF analysis is also readily available. The major documented limitation of this procedure is that it may be unable to detect *nonuniform DIF.*

Standardization

The standardization procedure, proposed by Dorans and Kullick (1986), is included because of its recurrent appearance in the literature and the intuitive appeal of the resulting index. Numerous studies have been based in part or entirely on this statistic. The value of interest is the standardized difference in the proportion correct, given by

$$D_{std} = \sum_s W_s (P_{fs} - P_{rs}).$$

P_{fs} represents the proportion correct on the studied item for focal group members within Score Group S. P_{rs} is the respective value for reference group members. W_s is the relative frequency of standardization group members (usually the focal group) within Score Group S.

In addition to intuitive appeal and a relatively large body of research and literature, this method has the advantage of simplicity. Its major limitation is the lack of an associated test of significance.

SIBTEST

SIBTEST (Shealy & Stout, 1993) is a relatively recent addition to the list of DIF statistics. Conceptually, it is similar to the standardization procedure. However, it includes significant innovations. Most notable among these is a test of significance, based on the ratio of the weighted difference in proportion correct (for reference and focal group member) to its standard error. It also includes several conceptual innovations. The first of these is that the matching criterion is a latent, rather than observed, score. Estimation of this matching score includes a regression-based correction that has been shown to be useful in controlling Type I error (Roussos & Stout, 1996; Shealy & Stout, 1993). Additionally, SIBTEST allows for evaluation of *DIF amplification* or *cancellation* across items within a *testlet* or *bundle*. Finally, this software is designed to perform the evaluation iteratively. Initially, all items are used in the matching criterion. Items displaying DIF are then removed from the matching criterion, and the analysis is repeated until a *valid subtest* of items that are "DIF-free" is identified for use as the final matching criterion. (An option for the user to specify the subtest to be used for matching is also available.)

The available software for SIBTEST is user friendly. In spite of its recent appearance, SIBTEST has been the object of substantial study. It has been shown to perform similarly to the Mantel-Haenszel statistic in identifying uniform DIF. It produces *Type I* errors at approximately the nominal level, has reasonable statistical power, and performs well with relatively small examinee samples (Narayanan & Swaminathan, 1994; Roussos & Stout, 1996).

Logistic Regression

Logistic regression (Swaminathan & Rogers, 1990) may be conceptualized as a link between the contingency table methods (Mantel-Haenszel, standardization, SIBTEST) and the IRT methods. The contingency table methods form groups based on discrete score categories. By contrast, logistic regression treats total score as a continuous variable and predicts performance on the studied item based on score and group membership. The basic model,

$$P(U = 1) = \frac{e^z}{1 + e^z}$$

allows for considerable flexibility in specifying the hypothesis to be tested. When $Z = T_0 + T_1\theta + T_2 G$, T_2 provides a measure of uniform DIF (θ is the matching ability, and G is the group coded 0, 1). Adding a term representing interaction between ability and group, $Z = T_0 + T_1\theta + T_2 G + T_3(\theta G)$, allows for testing for uniform and nonuniform DIF. The presence of uniform and/or nonuniform DIF can be tested simultaneously by comparing the fit of the augmented model (including T_0, $T_1\theta$, $T_2 G$, and $T_3[\theta G]$) to that of the compact model (including only T_0 and $T_1\theta$). This flexibility in specifying the model also allows for simultaneous conditioning on multiple abilities (i.e., multidimensional matching).

Studies with both simulated and real data have shown that this procedure produces results similar to the Mantel-Haenszel statistic when testing for uniform DIF. It is superior to the Mantel-Haenszel statistic for identifying nonuniform DIF (Clauser, Nungester, Mazor, & Ripkey, 1996; Rogers & Swaminathan, 1993a).

DIF Statistics for Polytomous Items

With increasing interest in performance assessment and other formats that require *polytomous scoring,* a clear need has arisen for DIF statistics applicable to polytomous models. Logistic regression (Rogers & Swaminathan, 1993b), SIBTEST (Chang, Mazzeo, & Roussos, 1996), the Mantel-Haenszel statistic (Zwick, Donoghue, & Grima, 1993), and IRT methods (Wainer, Sireci, & Thissen, 1991) can all be extended for use in this context. Additionally, Miller and Spray (1993) have suggested a variant on logistic regression: logistic discriminant function analysis. With this approach, group membership, rather than success status on the studied item, is used as the dependent measure in the regression equation.

Substantially less research is available on the performance of these methods than for the applications with *dichotomous* items. Several authors have presented results comparing procedures (Chang, Mazzeo, & Roussos, 1996; Welch & Hoover, 1993; Welch & Miller, 1995). Initial results suggest that the procedures perform reasonably well. However, these studies have tended to highlight the fact that the implementation decisions are considerably more complex than those decisions required for implementation.

Summary

In general, research comparing the various DIF statistics described above suggests that they produce similar results. With the possible exception of IRT methods, all the procedures described for use with dichotomous data have Type I error rates at or near nominal levels (see Cohen, Kim, & Wollack, 1996, for a discussion of the Type I error with IRT DIF analysis). They also have reasonable power with small samples. Variations on each of the procedures are also available to increase the sensitivity to nonuniform DIF. Again, with the possible exception of the IRT methods, each of these procedures can be implemented by individuals with limited sophistication in terms of statistics and computer skills. For routine DIF screening, the choice of a procedure is likely to be based on familiarity and personal preference. Specific demands of the intended analysis will have obvious implications in less routine settings. (For example, if multivariate matching is important, logistic regression may be preferred. If DIF amplification within a testlet is of interest, SIBTEST is an obvious choice.) These considerations are discussed in greater detail below.

Implementation Decisions

The relatively positive review of all the procedures given above does not imply that there are not requirements and limitations for their appropriate use. The following briefly describes several considerations essential for appropriate implementation of most, if not all, of these procedures.

Internal Versus External Matching Criteria

The validity of any DIF screening procedure is based on appropriate matching of examinees from reference and focal groups. The choice of a criterion is central in this regard. The obvious limitation of *internal criteria* is that such criteria provide no basis for identifying pervasive DIF. When total test score is used for matching, across all items, the total measure of effect size for DIF items favoring the reference group must approximately offset that for items favoring the focal group. Additionally, at the point that DIF screening is occurring, the validity of the test score must be in some doubt. A demonstrably valid external measure of the ability measured by the studied test would be optimal. Unfortunately, it is unlikely that such a measure will typically be available. There are few studies reporting on analysis with *external criteria.* Hambleton, Bollwark, and Rogers (1990) presented a comparison of internal versus external criteria. In that study, although the two criteria were only moderately correlated, results of DIF screening (with the Mantel-Haenszel statistic) were similar.

Purification of the Matching Criterion

Although the use of internal criteria may be by default rather than choice, it is a common practice. Holland and Thayer (1988) recommended an iterative application of the Mantel-Haenszel procedure in order to ameliorate the effect of DIF items within the matching criterion when an internal criterion is used. The total test score is used as the matching criterion for the initial step. Items identified as DIF are omitted, and the score is recalculated. This score is then used as the matching criterion for a second Mantel-Haenszel analysis. Again all items are assessed. This type of purification process has obvious appeal and could be applied to any of the procedures described above, including the IRT methods. In addition to intuitive appeal, the procedure has empirical support. Across simulated conditions, use of this purified criterion with the Mantel-Haenszel procedure produced results equal or superior to those for the nonpurified criterion, without inflation of the Type I error rate (Clauser, Mazor, & Hambleton, 1993). A similar iterative approach is operationalized as an available option in the SIBTEST software to produce a valid subtest for matching.

Including Studied Item in the Matching Criterion

Holland and Thayer (1988) recommended that when the purification process is used the studied item be included in the matching criterion even if it was identified as displaying DIF on the initial screening and excluded from the criterion for all other items. It has been repeatedly demonstrated that failure to adhere to this recommendation may result in inflated Type I error rates (Lewis, 1993; Zwick, 1990). Simulation results presented by Zwick et al. (1993) indicate that this recommendation also holds for implementation of the Mantel-Haenszel statistic for assessment of DIF in polytomously scored items. Simulation results from evaluation of the standardization procedure indicate that this recommendation is appropriate for use with that procedure as well (Donoghue, Holland, & Thayer, 1993).

Choice of a Matching Criterion When the Total Test Is a Multidimensional Composite

The choice of total test score as the matching criterion is based on the assumption that this score is the most reliable available measure of the ability of interest. If the test does not approximate unidimensionality, this score may not be an appropriate basis for matching examinees. In this circumstance, it will be necessary to attempt to account for the dimensional structure of the test. Several options are available. When subtests can be identified that reasonably approximate unidimensionality, items within the subtests can be analyzed separately using the subtest score as the matching criterion. This approach can be used with any of the procedures described above. Subtests can be formed based on content or factor analytic evidence. When factor analysis is used, content review should be implemented to verify the validity of the resulting subtest score. When the subtests are based on content, statistical analysis will be required to confirm that the resulting subtest does more closely approximate unidimensionality.

Several more complex approaches are available. Ackerman (1992) has suggested that items could be identified that fall

within a *validity sector* (i.e., a valid subtest). The score based on these items may approximate the dimension that best represents the test. The SIBTEST software attempts to operationalize this approach by using an iterative purification process to identify items that make up the valid subtest. It assumes that items outside the validity sector represent nuisance dimensions. Alternatively, the multidimensional structure of the test may be intentional. In this circumstance, it may be more appropriately represented with multivariate matching. Subtest scores, again formed based on factor analysis or content review, can be identified. The logistic regression approach can efficiently incorporate multiple criteria in a single model (Clauser, Nungester, Mazor, & Ripkey, 1996; Mazor, Hambleton, & Clauser, in press). Logistic regression also allows for modeling the dimensionally complex structure of a test using a combination of internal and external criteria (Mazor, Kanjee, & Clauser, 1995; Clauser, Nungester, & Swaminathan, 1996). A version of the SIBTEST software that allows for multidimensional matching (using two scores) has also been developed (Stout, Li, Nandakumar, & Bolt, 1997).

Reliability of the Matching Criterion

In order to test the hypothesis that an item displays DIF, it is necessary to match the focal and reference group examinees on ability. This obviously requires a reasonably reliable measure of that ability. When between-group differences in ability exist, use of an unreliable matching criterion will lead to identifying the most discriminating items as displaying DIF. This should not typically be a problem when the total test score is used for matching. When subsets of items are analyzed using subtest scores for matching, this requirement may be more limiting. How many items are required will depend on characteristics of the items. In a simulation study, Donaghue, Holland, and Thayer (1993) obtained satisfactory results when the matching criterion contained at least 10 items.

Thick Versus Thin Matching

Valid comparison of performance across groups requires that examinees be matched as accurately as possible. When observed score (on a test made up of dichotomous items) is used as the matching criterion, the total number of score categories will be one more than the number of items. In general, simulation research supports the practice of matching examinees using the narrowest score categories possible (*thin matching*). There are, however, circumstances under which it may be appropriate to use broader score categories. In order for data from a given score category to be included in the calculation of the Mantel-Haenszel statistic, they must include both correct and incorrect responses from the reference and focal groups. When the number of items (and score categories) is large and the number of examinees is small, this requirement may not be met. The examinees whose responses are represented in incomplete categories are lost from the calculations. This results in a loss of power for the statistic. When reference and focal groups differ substantially in ability, such incomplete cells are particularly likely to be present at unusually high and low scores. Using wider score categories (*thick matching*), particularly at the extremes, may be useful in this circumstance (Donoghue & Allen, 1993).

Sample Size

As with any statistical procedure, the power of a DIF procedure is related directly to sample size. With very small samples of reference and/or focal group members, even items displaying substantial DIF will go undetected. Numerous simulation studies have examined the power issue. Results show that more is better. Samples of 200 to 250 per group have been consistently shown to be suitable for use with the Mantel-Haenszel statistic, logistic regression, and SIBTEST (Narayanan & Swaminathan, 1994; Rogers & Swaminathan, 1993a). It is typically suggested that larger samples are required for use with IRT methods, when two or three parameter models are used. With these methods, the choice of model will probably influence the sample requirements.

Nonuniform DIF

Statistical DIF detection procedures represent an effort to model performance of test items that may not perform in an equitable manner across groups. Not all models are sensitive to every manifestation of DIF. Two distinctly different forms of DIF have come to be known as *uniform* and *nonuniform* DIF.

In the framework of IRT models, uniform DIF exists when the item characteristic curves across groups differ only in terms of the difficulty parameter. The relative advantage for the focal (or reference) group is uniform across the score scale. Nonuniform DIF occurs when ICCs for two groups differ in their discrimination parameters and/or pseudo-guessing parameters. Figures 2 and 3 represent examples of uniform DIF. Figures 4 and 5 represent nonuniform DIF.

All the procedures described above are sensitive to uniform DIF. By contrast, nonuniform DIF may go undetected unless the procedure used is specifically designed to be sensitive to this type of DIF. DIF detection procedures based on two- (and three-) parameter IRT models are by definition intended to be sensitive to this type of DIF. Each of the other procedures is sensitive to nonuniform DIF with appropriate modification. Among the advantages of logistic regression noted by Swaminathan and Rogers (1990) is the fact that the model can be modified to include a term representing the interaction between group membership and ability. This allows for sensitivity to nonuniform DIF. SIBTEST (Li & Stout, 1996), the standardization procedure (Dorans & Kullick, 1986), and the Mantel-Haenszel statistic (Mazor, Clauser, & Hambleton, 1994) have all been modified to allow for identification of nonuniform DIF. In practice, initial studies to examine the prevalence of nonuniform DIF within the test of interest may be warranted before an operational form of the statistic is chosen.

Item Characteristics

Several simulation studies have examined the sensitivity of DIF statistics to items based on the difficulty and discrimination. Not surprisingly, items with lower discrimination were less likely to be identified. Items that were very difficult or very easy, relative to the examinee ability distribution, were also likely to go undetected. In most circumstances, these limitations are likely to be of little concern. When the purpose of the test is to select a very small percentage of the most (or least) competent examinees, this latter limitation could be a problem. In this circumstance, it may be appropriate to limit the analysis to the range of scores representing only high (or low) performers. Obviously, this could limit the power of the analysis making it necessary to collect a relatively large sample of examinee responses from the range of interest.

Defining the Focal and Reference Groups

The previous discussion has assumed that the researcher has identified the comparison groups of interest. Policy aspects of this decision will be discussed in a subsequent section. Regardless of the groups identified, the importance of careful sampling should not be ignored. For example, when the groups are defined in terms of ethnicity, ethnicity based on the self-reports of examinees may not be equivalent to actual ethnicity because of nonresponses in self-reported examinee data. Similarly, it is important to ensure that, when DIF analyses are based on pretest results, the pretest sample represents the actual group of interest.

Testlets Versus Single Items

DIF is the assessment of item functioning. However, there are cases in which the item may not be the appropriate unit of analysis. When items are grouped in clusters or testlets—such as, a group of items based on a single reading passage—there may be local item dependence. This violates the assumptions of IRT-based DIF analysis.

When the test is composed of clusters of items, testlet rather than item analysis may be preferred. This approach simultaneously deals with the problem of local dependence and allows for increased sensitivity to DIF within the testlet. With testlet level analysis, low levels of DIF associated with individual items can accumulate within a testlet to show significant DIF for the cluster.

There are several approaches to examining testlet level DIF. SIBTEST is specifically designed to allow for evaluation of whether item DIF accumulates or cancels across items within a testlet (Douglas, Roussos, & Stout, 1996). Alternatively, treating the testlet as a single polytomous item also allows for appropriate evaluation. Wainer (1995) provides an example of this approach with IRT-based DIF analysis. Any of the polytomous models have potential in this application.

Interpreting the Results

Having selected one or more DIF detection procedures and implemented the analysis for a set of items, it remains to interpret the results. This section describes the indices produced by the various procedures and provides information relevant to using those indices to identify potentially biased items. To provide an illustrative example, responses were generated to simulate a 40-item test administered to 500 examinees from both the reference and focal groups. Results are reported for three items, two of which were simulated to display DIF and one which was not. Items were generated using the three-parameter logistic model. The resulting test had a reliability of .87. Focal and reference groups were simulated to have the same ability distributions; this allows observed *p*-value differences between groups to be used as a frame of reference for interpreting the various DIF statistics. Descriptive information on the items and DIF indices produced by the various methods are shown in Table 1.

Item Response Theory Approaches

As noted previously, there is no single IRT method—there are numerous methods using this framework. Each of the methods share the same conceptual null hypothesis; the item parameters are the same for the focal and reference groups. In the simplest case (i.e., the Rasch model), DIF is manifest as a difference in the difficulty parameter for the studied item. With the two- and three-parameter models, DIF is quantified in terms of the area between the item characteristic curves for the two groups. Holland and Thayer (1988) have shown that the α_{MH} can be conceptualized as $e^{b_F - b_R}$, where b_F and b_R are the difficulty parameters for the focal and reference groups. Again, conceptually, the log of this value is equivalent to T_2 from the logistic regression test for uniform DIF (i.e., $\ln(e^{b_F - b_R}) = T_2$). As with the previous comments on nonuniform DIF in the context of logistic regression, quantifying DIF with the two- and three-parameter models takes on additional complexity. The combined effect of differences in a- and b parameters can be quantified as the area between the item characteristic curves for the focal and reference groups (Raju, 1988). Raju (1990) has also provided asymptotic sampling distributions and tests of significance for use with the area method. Alternatively, the distance between the two item characteristic curves can be calculated at each ability level. For examinees performing at the given ability level, this distance is directly interpretable as the difference in probability of success on an item for examinees from the two groups.

Several approaches to testing the significance of parameter differences are available. Lord (1980) suggested comparing the difference in the b parameters (or both a- and b parameters) to the standard error of the difference. He noted that this approach may be limited because it requires large samples and assumes that the parameters are estimated but the examinee abilities are known. As a check on the appropriateness of the statistic, he suggested examining its performance under the null condition in which examinees are randomly assigned to focal and reference groups. Subsequently, Shepard, Camilli, and Williams (1984) suggested using random assignment of examinees as a means of establishing an empirical estimate of significance for estimates of differences between item characteristic curves. Alternatively, simulated data can be used to establish this type of baseline (Rogers & Hambleton, 1989).

More recently, Thissen, Steinberg, and Wainer (1993) have suggested using likelihood ratios to test improvement in the fit of the IRT model associated with adding additional parameters to account for differences in the studied items' performance for focal and reference group members. Their article describes several specific models and provides code for implementing the procedures with readily available commercial software packages. Conceptually, their approach is similar to that suggested by Lord (1980). However, their methodology takes advantage of innovations in model fitting and hypothesis testing that have been developed since Lord presented his procedure.

The item characteristic curves for the simulated sample items were shown in Figures 1, 2, and 3. For the first item, the parameters used for generating the responses were the same for both groups (a = 1.00, b = 0.0, c = .14). Consequently, the curves for this item are coincidental. Under these circumstances, the expected values for differences between item parameters will be zero. (The observed values will vary around this expectation due to estimation error.) Because group membership provides no information regarding examinee response for this item, there will also be no expected dif-

Table 1
Descriptive Information and DIF Indices for Sample Items

	Item parameters				*p*-value				Standardized	
Item	a	b_{ref}	b_{foc}	c	difference	Area	Δ_{MH}	β	*p* difference	T_2
1	1.0	0.0	0.0	.14	.02	.00	0.11	.01	.01	–.08
2	1.0	0.0	0.5	.14	–.05	.42	–1.09	–.07	–.06	.37
3	1.0	–1.0	–0.2	.14	–.11	.67	–3.45	–.13	–.13	1.34

ference in the fit of the model with and without separately estimated parameters for the studied item.

For the second sample item, the discrimination is identical for the two groups (a = 1.00) as is the pseudo-guessing parameter (c = .14). The difficulty of the item differs across groups by 0.5 (reference group b = 0.0; focal group b = 0.5). For this item, the area between the curves is .42. For the simulated sample, this represented an observed p-value difference of .05.

For sample Item 3, the a and c parameters are again unchanged. The b parameter for the reference group was –1.00; for the focal group, it was –0.20. This results in a difference in b parameters across groups of .80 and an area between the curves of .67. For the simulated sample, this represented an observed p-value difference of .11.

Mantel-Haenszel Statistic

The Mantel-Haenszel statistic provides three values that are of interest, α, Δ_{MH}, and $MH\chi^2$. The α is the odds ratio—that is, the ratio of the odds that a reference group examinee will get an item correct to those for a matched focal group examinee. For items favoring the reference group, α takes values between one and infinity; for items favoring the focal group, α takes values from zero to one. The asymmetry of this scale makes interpretation difficult. To produce a more readily interpretable scale, Holland and Thayer (1988) suggested a logistic transformation. Taking the log of α transforms the scale so that it is symmetric around zero. Multiplying the resulting value by –2.35 produces the Δ_{MH}. This places the value on Educational Testing Service's delta scale, with items favoring the reference group having values from minus infinity to zero and items favoring the focal group having values from zero to infinity. The α and Δ_{MH} are both measures of effect size. The $MH\chi^2$ provides a significance test distributed as an approximate chi-square with one degree of freedom. As with other tests of significance, the statistic is dependent on the sample size.

To avoid identifying items that display practically trivial but statistically significant DIF, item identification with the Mantel-Haenszel statistic is often based on a combination of statistical significance and effect size. A well-known example of this approach is the three-level classification system used by Educational Testing Service (Zieky, 1993; Zwick & Ercikan, 1989). Items classified in the first level, A, have a Δ_{MH} with an absolute value of less than 1.0 and/or have a value that is not significantly different than zero ($p < .05$). Items in the third level, C, have a Δ_{MH} with absolute value greater than 1.5 and are significantly greater than 1.0 (i.e., 1.0 is outside the confidence interval, around the estimated value). Items in the second level, B, are those that do not meet either of the other criteria. Items classified as A are considered to display little or no DIF and are considered appropriate for use in test construction. Items classified as B are used only if no A item is available to fill the content requirement of the test. Items classified as C are to be used only if the content experts consider them essential to meet the test specifications.

Turning to the Δ_{MH} and $MH\chi^2$ for the three sample items presented in the previous section, Item 1, simulated to have no DIF, had a Δ_{MH} close to zero (.11), and the $MH\chi^2$ value was nonsignificant. Item 2 had a Δ_{MH} with an absolute value greater than 1.0, and the $MH\chi^2$ value was significant at $p < .01$. Based on this result, Item 2 would be classified as a B item, using the classification system described in the previous paragraph. Item 3 had a Δ_{MH} with an absolute value substantially greater than 1.5 (–3.45), and the $MH\chi^2$ value was significant at $p < .01$. A real test item that performed like this example would be classified as a C item, and it would warrant very careful attention.

Standardization

The standardization procedure produces a single measure of effect size, the standardized p difference. The value represents the average p difference in response to the studied items for members of the focal and reference groups. This value is similar to the β produced by SIBTEST except that the matching criterion is the observed score rather than an estimated true score and the standardization is typically produced by weighting the p differences by the proportion of focal group examinees at each level. Unlike the other methods described previously, there is no associated test of significance. Dorans (1989) suggests a standardized p difference of 0.10 as a reasonable level for flagging items for review. The three simulated items produced values of .01, –.06, and –.13, using this method. These values correspond reasonably closely to the observed p-value differences of .02, –.05, and –.11 for the same items. Using the 0.10 criterion, only Item 3 would be identified for review.

SIBTEST

The primary output from SIBTEST is an estimate of β and a z statistic representing β divided by its standard error. The estimate of β represents the difference in the probability of a correct response to the studied item for examinees from the focal and reference groups (matched on the ability of interest). When SIBTEST is used to examine testlets or item bundles, β is estimated separately for each item in the bundle and then aggregated. When the individual items in the bundle primarily show DIF in the same direction and the bundle DIF is not the result of one or two items with large β values, bundle DIF may be represented by the mean β value within the bundle. For large samples, β follows a standard normal distribution. This allows for testing the significance of β; the significance test is included in the SIBTEST output. A positive value for β indicates that the item favors the reference group; a negative value indicates that it favors the focal group. As with the Mantel-Haenszel statistic, it is possible to use SIBTEST results to classify items based on both effect size and statistical significance. The three simulated sample items had β values of .01, –.07, and –.13. For Items 2 and 3, these values were significant at $p < .01$. As with the standardization values, the β values approximate the observed p-value differences. Reasonable interpretation (or categorization) of these items would be similar to that for the Mantel-Haenszel results, with Item 1 meeting the expectation for an item that does not display DIF, Item 3 clearly requiring careful review, and Item 2 showing evidence of more moderate DIF.

Logistic Regression

As with SIBTEST and the Mantel-Haenszel statistic, logistic regression provides both a significance test and a measure of effect size. The significance test is in the form of a chi-square test of improvement of fit for the model associated with adding a dichotomous variable representing group membership. Significance reflects improvement in fit above and beyond that obtained with a model predicting item performance from examinee ability alone. The test for uniform DIF involves comparing the fit for the augmented model (including ability and group membership variables) to the fit for the compact model (predicting performance on the studied item from ability only). It is distributed as a chi-square with one degree of freedom. The simultaneous test of uniform and nonuniform DIF requires an augmented model including a variable for group membership and a second variable for the interaction between group membership and ability; the associated chi-square test has 2 degrees of freedom.

The measure of effect size for the test of uniform DIF is the estimate of T_2 associated with the group membership variable. This represents the logit difference or log odds associ-

ated with a unit change in the variable, after accounting for ability. If group membership is coded 0/1, T_2 will equal the logit difference between groups for success on the studied item, for examinees of equal ability. If a marginal coding approach is used—so that, for example, the reference group is coded −1 and the focal group is coded 1—the logit difference between groups will be $2T_2$. This interpretation of the measure of effect size holds equally when multivariate matching is used to account for examinee ability.

For the test of uniform DIF (with group membership coded 0/1), the T_2 for simulated Item 1 was −.08. The test of improvement in the model associated with including the group variable was nonsignificant ($p < .65$). The test for improvement in the model associated with including both group and the interaction term was also nonsignificant. For Item 2, the T_2 for the group variable was .37. Including this variable was associated with an improvement in fit that was significant at $p < .02$. The less powerful test of improvement in fit associated with including both the group variable and interaction term was nonsignificant at $p < .05$. The T_2 for the group variable in Item 3 was 1.34. The test of improvement in fit of the model was significant at $p < .01$, both with and without the interaction term included. Again, as with the Mantel-Haenszel and SIBTEST results, logistic regression analysis would indicate that the first item displayed no evidence of DIF; the third item required close scrutiny, and the second item was marginal (i.e., it would most likely be identified for review, depending on the specific criterion selected).

When identification of nonuniform DIF is of interest, interpretation of the effect size (i.e., T_2) is less straightforward. If the interaction term is significant, the main effect for group membership cannot be interpreted because the difference between groups will vary as a function of ability. When it is of interest to apply a classification system like the one described previously, one or more benchmark test score values can be identified, and the effect size can be estimated at these scores.

Policy Decisions

The main focus of this module has been the use of statistical procedures for identifying DIF. The results of these procedures should be informative, but they will not provide answers per se. The results of DIF screening must be interpreted within the context of the intended use of the test. The use of the information must be shaped by policy. Specifics of the DIF detection methodology may be guided by technical considerations, but the very decision to implement the analysis is a policy decision. Because there are numerous comparison groups that could be defined, the choice of which ones to use is a policy decision. The level of DIF required for an item to be identified is a matter of policy. A policy decision is required to determine if items should be identified for review only if they favor the reference group or regardless of which group they favor. Finally, it is a matter of policy as to what should be done when an item is identified as displaying DIF. The issue is whether identified items are considered biased until proven valid or valid until proven biased. Arguments have been made on each side. The former approach requires rules to be established restricting use of DIF items until they are revised or additional evidence can be collected to support the validity of the item(s). Because DIF is a necessary but not sufficient condition for bias, this approach will likely lead to deletion of valid items. By contrast, items identified as displaying DIF could be targeted for review by content experts. Items could be deleted or maintained based on their judgments. The problem with this approach is that there is little evidence to support the validity of these judgments. Biased items would be likely to remain in the test. Clearly, either approach leads to errors and determination of the relative cost of these (while it may be guided by empirical evidence) will be a policy decision.

Although relatively little has been written to provide guidance in these areas, Zieky (1993) and Linn (1993) have commented on these issues. It seems that the short answer to each of these questions is "it all depends." Practicality and the intended use of the test are dominant considerations. Groups are generally defined in terms of ethnicity or gender. Linn notes the potential value of additional analyses but suggests that practicality limits the number of analyses that can reasonably be executed. Both Zieky and Linn agree that the three-level classification system used by Educational Testing Service is appropriate. However, this begs the question of what is to be done with an item after it is classified. The Educational Testing Service DIF classification scheme (Zieky, 1993; Zwick & Ercikan, 1989) seems to imply that DIF items will be treated similarly regardless of which group is potentially disadvantaged. The scheme is also associated with a decision rule for use and does not seem to call for judgment. Nonetheless, Zieky (1993) and Linn (1993) both seem to consider judgment essential throughout the process. Linn also suggests that the judgment process may produce a result that is less symmetrical than the classification categories imply.

The practitioner is left to make the decisions. As with other aspects of test validation, DIF analysis is a process of collecting evidence. Weighing and interpreting that evidence will require careful judgment. There is no single correct answer.

Glossary

Dichotomous scoring: Dichotomously scored items are scored either correct or incorrect, 1/0.

Differential item functioning (DIF): An item displays DIF if examinees from different groups have differing probabilities or likelihoods of success on the item after conditioning or matching on the ability the item is intended to measure.

DIF amplification/DIF cancellation: DIF amplification is present when the items within a studied subset do not individually exhibit a meaningful level of DIF until they are accumulated across the subset when the total level of DIF becomes meaningful. DIF cancellation occurs when the amount of DIF in a particular set of items is meaningful when the items are considered separately, but, when the items are grouped, and the amount and direction of the differential are considered across all items, neither group has a meaningful advantage.

Internal/External matching criterion: Internal matching criteria are measures based on the test from which the studied item is taken. These criteria may be either total test scores or subtest scores. External matching criteria are measures not based on the test under study. Examples include scores from other tests and course grades.

Item bias: An item is considered biased against examinees of a particular group if members of that group are less likely to answer that item correctly than examinees of another group because of some aspect of the test item or the testing situation which is not relevant to the purpose of testing.

Item impact: Item impact is present when examinees from different groups have differing probabilities of responding correctly to an item because of true differences in the underlying ability that the item measures.

Polytomous scoring: Polytomously scored items have more than two score categories. For instance, the scoring instructions for an essay item might allow scores of 0, 1, 2, or 3.

Reference group/Focal group: The terms *reference* and *focal group* refer to the examinee classifications used to define the hypothesis examined in DIF analysis. Typically, groupings are made on the basis of gender, ethnicity, or other demographic information. Holland and Thayer (1988) defined the

focal group as the group "of primary interest" (p. 130). The reference group is the standard against which the focal group is compared. The decision as to which group is considered the reference group and which the focal group is somewhat arbitrary. Often, researchers treat the majority group, or the group more likely to be advantaged, as the reference group, and the minority group as the focal group.

Testlets and item bundles: The terms *testlet* and *item bundle* refer to groups of items that are treated as a single unit. Testlet typically refers to items that are developed as a unit, such as multiple items that refer to a single passage of text. Item bundles are more likely to be formed after test construction, based on item characteristics such as content or format.

Thick/Thin matching: When thin matching is used, examinees are matched to the finest extent possible. For a test with n items, examinees would be grouped into $n + 1$ categories (assuming there was at least one examinee at each score level). By contrast, in thick matching, examinees are grouped in score categories that include a range of score values. For instance, for a 100-item test, examinees might be grouped into 20 score groups, each of width five.

Type I error: In DIF analyses, a Type I error refers to the case in which an item is identified as displaying DIF when there is no between-group performance difference in the population.

Uniform DIF/Nonuniform DIF: Uniform DIF exists when one group is advantaged across the entire ability scale, and that advantage is roughly the same (uniform) across all points on the scale. Conversely, nonuniform DIF refers to the case when one group has an advantage at one end of the ability scale, but that advantage increases, decreases, or reverses at other ability levels. For example, reference group members may have an advantage on a given item at the lower ability levels, but at higher ability levels that advantage could shift, so that the item then favored focal group examinees. In IRT terms, uniform DIF refers to the case when there is a difference only in item difficulty, and nonuniform DIF refers to the case when there is a difference in item discrimination. Differences between groups in the lower asymptote (represented by the c or pseudo-guessing parameter) will also result in nonuniform DIF. Figures 2 and 3 (presented previously) provide examples of uniform DIF. Figure 4 shows an example of nonuniform DIF resulting from differences in the item discrimination across groups. Figure 5 provides an example of an item that displays DIF resulting from differences in both difficulty and discrimination. This is sometimes referred to as mixed DIF.

Valid subtest: A valid subtest is a set of items that have been determined to be valid, unbiased measures of the ability of interest.

Validity sector: A validity sector is a geometric construct used to define a narrow sector within a two-dimensional framework that includes items making up a valid subtest. This subtest is intended to represent the composite that the test is intended to measure. Items falling outside of this sector are considered to be too heavily influenced by abilities or factors considered irrelevant to that composite.

Self-Test

One of your newly assigned responsibilities as a member of a test development team is to plan and implement appropriate DIF analyses. This team is responsible for a test designed to measure reading and math achievement in seventh graders. You are brought on just as they finish pretesting an "almost final" version of the test. You are given the following information:

The test consists of 50 math items and 50 reading items. Pretest data are available for 3,100 students. Student responses to the ethnicity item break down as follows: 2,000 are self-identified as White, 500 as Black/African American, 250 as Hispanic, 75 as Native American, 180 as Other, and 95 did not respond. On the gender item, 1,490 are self-identified as males, 1,560 as females, and 50 did not respond.

1. The project leader comes into your office and says: "I just had a call from an angry teacher. She said she counted 33 uses of boys' names in the math word problem section and only 21 references to girls. She says that this is obviously unfair and that, if girls don't do as well as boys on the math test, it's because of inequities such as this. Do me a favor; run one of your analyses on this." How do you respond?
   (a) Run one of "your analyses"—such as the Mantel-Haenszel or logistic regression—and see if the items which contain references to boys are easier for boys compared to girls, after conditioning on total score.
   (b) Count the number of boy/girl references in the entire test, and see if it balances out overall.
   (c) Tell the project leader that this is not really a DIF issue, and DIF analyses would not necessarily address the teacher's concern.

Answer: c and a.

This type of concern is usually addressed in a judgmental or sensitivity review of the item, which would typically take place before pretesting. It is unlikely that boys will actually do better than girls on items which include a boy's name. Once a DIF analysis is completed, it would be possible to examine whether the items naming boys are in fact easier for boys than girls (after conditioning on ability). However, should some of the boy-referenced items be flagged as DIF, it does not necessarily follow that the use of a boy's name is the problematic aspect of the item.

2. You have experience and expertise with logistic regression, so you use that procedure to conduct the DIF analyses. One of the members of the team tells you that he recently read that the Mantel-Haenszel is considered the industry standard. The project leader asks if you think that you should re-run all of the analyses using the Mantel-Haenszel "just to be sure." How do you respond?
   (a) Tell him that, in fact, IRT methods are considered the most theoretically defensible and that, if you are going to re-run anything, you should use an IRT procedure.
   (b) Tell him that there is a substantial body of literature that says that the Mantel-Haenszel and logistic regression yield virtually the same results and that there is no literature which demonstrates that the Mantel-Haenszel statistic is better than logistic regression.
   (c) Tell him that SIBTEST provides results that are generally comparable to the Mantel-Haenszel but also allows for identification of nonuniform DIF, so that, if he wants you to check the logistic regression results, you prefer to use SIBTEST.

Answer: b.

There is no reason to believe that you would get better, more accurate results using a different method. Logistic regression has a substantial body of literature that has established it as an appropriate and accurate procedure for detecting DIF. This is not to say you could not have chosen

the Mantel-Haenszel method, one of the IRT methods, or SIBTEST and made the same argument.

3. Another colleague tells you that she thinks that some of the test items are biased against Native Americans. There is a team meeting tomorrow afternoon, and she asks you to run a White/Native American DIF analysis before then and to bring the results to that meeting. What do you do?
   (a) Tell her there are not enough Native Americans to make the analysis meaningful.
   (b) Run the analysis and bring the results to the meeting as requested.
   (c) Tell her that less than one workday is not sufficient time for you to run the analysis she wants.

Answer: a.

There are not enough Native Americans in the sample to make the analysis meaningful. It is unlikely that you would be able to identify any differentially functioning items with only 75 examinees in the focal group, unless the conditional between-group difference is unusually large. Additionally, with so few Native Americans, idiosyncratic response patterns will be unduly influential. Finally, conducting a DIF analysis with an inadequate sample may be worse than a waste of time—it may be harmful in that the absence of any flagged items might be interpreted to mean the absence of DIF when, in fact, there may be DIF items that were not flagged because of lack of power. (Time is not likely to be a concern, as running an analysis for a single reference/focal group comparison can be done in a matter of minutes. It is the planning of the analyses and the interpretation of the results that are time consuming.)

4. Your colleague who has been reading about the Mantel-Haenszel statistic enters your office, smiling broadly. He ran a Mantel-Haenszel analysis using males and females as the focal and reference groups—the same analysis that you ran with logistic regression. He admits he was hoping to prove you wrong and to flag different items, but he did not. Instead, he found something that he thinks is even better. He added up all the Mantel-Haenszel Δ_{MH} values and found that they sum to zero. Thus the 8 items that you flagged as DIF cancel one another out. For the test as a whole, DIF is not a problem. How do you respond?
   (a) Re-run the Mantel-Haenszel analysis yourself to determine whether he made an error.
   (b) Point out that this is exactly what one would expect from a Mantel-Haenszel analysis when total score is used as the matching criterion. This result should not be interpreted to mean that "the bias cancels out."
   (c) Check to see whether the logistic regression beta values also sum to zero.

Answer: b.

5. Your colleague returns to your office, smiling again. He says that there is pervasive bias against girls throughout the math test, which means that total score is an inappropriate matching criterion and, further, that the results of the male/female analyses are invalid. He shows you a graph of the male and female math score distributions, which highlights that the means of the two distributions are separated by approximately one half a standard deviation, or approximately 5 points. How do you respond?
   (a) Tell him that it is not unusual for reference and focal group score distributions to differ and that such differences are not evidence of pervasive bias.
   (b) Tell him that, while there is pervasive bias, such bias does not invalidate your results. By conditioning on total score, you took that bias into account, and so your conclusions stand.
   (c) Re-run your analysis of the math items, this time using reading score as the matching criterion, because there are no male/female differences in mean reading test scores.

Answer: a.

Differences in total test score distributions are not evidence of pervasive bias. Similarly, the lack of such differences is not evidence of the lack of such bias. Pervasive bias is best prevented by sound test construction practices. If construct and predictive validity can be established, pervasive bias is unlikely to be a concern. (Conditioning on total test score does not remove pervasive bias. Conditioning on some other external criterion is useful only if that criterion is a valid measure of the ability the test is intended to measure. Thus, the reading score is not an appropriate substitute for the math score in this instance.)

Notes

[1] Terms defined in the glossary are printed in italics when they first appear in the text.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67–91.

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 397–418). Hillsdale, NJ: Erlbaum.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage Publications.

Chang, H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33,* 333–353.

Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (1993). The effects of purification of the matching criterion on identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education, 6,* 269–279.

Clauser, B. E., Nungester, R. J., Mazor, K., & Ripkey, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. *Journal of Educational Measurement, 33,* 202–214.

Clauser, B. E., Nungester, R. J., & Swaminathan, H. (1996). Improving the matching for DIF analysis by conditioning on both test score and an educational background variable. *Journal of Educational Measurement, 33,* 453–464.

Cohen, A. S., Kim, S., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement, 20,* 15–26.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

Donoghue, J. R., & Allen, N. L. (1993). Thin versus thick matching in the Mantel-Haenszel procedure for detecting DIF. *Journal of Educational Statistics, 18,* 131–154.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 137–166). Hillsdale, NJ: Erlbaum.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education, 2,* 217–233.

Dorans, N. J., & Kullick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23,* 355–368.

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement, 33,* 465–484.

Hambleton, R. K., Bollwark, J., & Rogers, H. J. (1990). *Factors affecting the stability of the Mantel-Haenszel item bias statistic* (Laboratory of Psychometric and Evaluative Research, Research Report 203). Amherst, MA: University of Massachusetts, School of Education.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Lewis, C. (1993). A note on the value of including the studied item in the test score when analyzing test items for DIF. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 317–320). Hillsdale, NJ: Erlbaum.

Li, H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika, 61,* 647–677.

Linacre, J. M., & Wright, B. D. (1986). *Item bias: Mantel-Haenszel and the Rasch model* (MESA Psychometric Laboratory, Memorandum No. 39). Chicago: University of Chicago, Department of Education.

Linn, R. L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 349–364). Hillsdale, NJ: Erlbaum.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1994). Identification of non-uniform differential item functioning using a variation of the Mantel-Haenszel procedure. *Educational and Psychological Measurement, 54,* 284–291.

Mazor, K., Hambleton, R. K., & Clauser, B. E. (in press). The effects of conditioning on two internally derived ability estimates in multidimensional DIF analysis. *Applied Psychological Measurement.*

Mazor, K., Kanjee, A., & Clauser, B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement, 32,* 131–144.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale, NJ: Erlbaum.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30,* 107–122.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18,* 315–328.

Norusis, M. J. (1993). *SPSS for windows advanced statistics.* Chicago: SPSS.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 54,* 495–502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14,* 197–207.

Ramsey, P. A. (1993). Sensitivity review: The ETS experience as a case study. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 367–388). Hillsdale, NJ: Erlbaum.

Rogers, H. J., & Hambleton, R. K. (1989). Evaluation of computer simulated baseline statistics for use in item bias studies. *Educational and Psychological Measurement, 49,* 355–369.

Rogers, H. J., & Hambleton, R. K. (1994). MH: A FORTRAN 77 program to compute the Mantel-Haenszel statistic for detecting differential item functioning. *Educational and Psychological Measurement, 54,* 101–104.

Rogers, H. J., & Swaminathan, H. (1993a). A comparison of logistic regression and the Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17,* 105–116.

Rogers, H. J., & Swaminathan, H. (1993b, April). *Logistic regression procedures for detecting DIF in non-dichotomous item responses.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement, 33,* 215–230.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58,* 159–194.

Shepard, L. A., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics, 9,* 93–128.

Stout, W., Li, H., Nandakumar, R., & Bolt, D. (1997). MULTISIB: A procedure to investigate DIF when a test is intentionally two-dimensional. *Applied Psychological Measurement, 21,* 195–213.

Stout, W., & Roussos, L. (1995). *SIBTEST manual* (2nd ed.). Unpublished manuscript, University of Illinois at Urbana–Champaign.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361–370.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–114). 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,* 157–186.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28,* 197–219.

Welch, C., & Hoover, H. D. (1993). Procedures for extending item bias detection techniques to polytomously scored items. *Applied Measurement in Education, 6,* 1–19.

Welch, C. J., & Miller, T. R. (1995). Assessing differential item functioning in direct writing assessments: Problems and an example. *Journal of Educational Measurement, 32,* 163–178.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–348). Hillsdale, NJ: Erlbaum.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15,* 185–197.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30,* 233–251.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement, 26,* 44–66.

Annotated Bibliography

The following is a selective list of sources for software and information on DIF detection. The selected sources are intended to provide the information necessary for implementation of the procedures described in the previous sections. For more detailed information on specific topics, the reader is directed to the references listed throughout the article. Complete citations for the books and articles listed here are included in the reference section.

Books

Differential Item Functioning, P. W. Holland & H. Wainer (Eds.), 1993

This book is certainly the most important single source for information on DIF research. It provides descriptive articles

on the Mantel-Haenszel statistic, the standardization procedure, the theoretical basis for SIBTEST, and a particularly useful presentation on IRT-based DIF detection. It also includes several important articles providing historical and theoretical perspectives on DIF detection. The only limitation of the book follows unavoidably from the fact that most of the articles were written for a conference held in 1989. Subsequent developments in the area of DIF detection are largely absent.

Methods for Identifying Biased Test Items, G. Camilli & L. A. Shepard, 1994

The book provides an introduction to DIF detection. Perhaps most useful is the section that provides practical help in implementing the statistical procedures. Specifically, it provides examples of SPSS output, and so forth, and directs the reader's attention to the relevant lines. The authors also provide a useful discussion of some of the policy issues, particularly focusing on review of flagged items.

Articles and Chapters

Differential Item Performance and the Mantel-Haenszel Procedure, P. W. Holland & D. T. Thayer, 1988

This is an essential article for anyone interested in DIF detection procedures. It provides a theoretical framework and historical perspective on DIF detection procedures. It then shows how the Mantel-Haenszel statistic fits this framework and how the results relate to IRT definitions of DIF.

Detecting Differential Item Functioning Using Logistic Regression Procedures, H. Swaminathan & H. J. Rogers, 1990

This article presents the argument for using logistic regression to identify DIF. It provides a clear explanation of the model as well as the associated estimation and distribution theory. It describes the conceptual link between logistic regression and the Mantel-Haenszel statistic. The authors then provide an example of one of the advantages of the flexibility of the logistic regression model; they demonstrate that by adding an interaction term it is possible to identify nonuniform DIF which would be undetected with the Mantel-Haenszel statistic.

A Model-Based Standardization Approach That Separates True Bias/DIF From Group Ability Differences and Detects Test Bias/DTF As Well As Item Bias/DIF, R. Shealy & W. Stout, 1993

This article presents the theoretical basis for the SIBTEST procedure, a detailed explanation of the mathematics involved, and the results of a simulation study comparing SIBTEST to the Mantel-Haenszel statistic. The authors also provide a discussion of the difference between DIF and bias.

Fundamentals of Item Response Theory, R. K. Hambleton, H. Swaminathan, & H. J. Rogers, 1991

This book includes a chapter on IRT approaches to DIF detection which provides an excellent and reasonably nontechnical summary of IRT-based DIF detection techniques, including comparison of item characteristic curves, comparison of item parameters, and comparison of fit.

Detection of Differential Item Functioning Using the Parameters of Item Response Models, D. Thissen, L. Steinberg, & H. Wainer, 1993

This article sets out the theoretical basis for identifying DIF using IRT models. It provides several examples and includes an appendix with sample control cards for implementing DIF analysis using several different computer programs.

A Didactic Explanation of Item Bias, Item Impact, and Item Validity From a Multidimensional IRT Perspective, T. Ackerman, 1992

This article provides a clear statement of the relationship between DIF and dimensionality.

Computer Software

MH: A FORTRAN 77 Program to Compute the Mantel-Haenszel Statistic for Detecting Differential Item Functioning, H. J. Rogers & R. K. Hambleton, 1994

This software calculates the Mantel-Haenszel statistic and produces the $MH\chi^2$ and Δ_{MH} for each item. It produces results based on both the total test score and a "purified" total test score. The user can also select the option of applying an external criterion. The program additionally provides cell counts for reference and focal groups for all flagged items, as well as complete descriptive statistics. The software also implements the standardization procedure. The program is able to handle up to 100 items without limit on the number of examinees. Contact: Dr. Ronald K. Hambleton, University of Massachusetts, 152 Hills South, Amherst, MA 01003-4140.

SIBTEST Manual, W. Stout & L. Roussos, 1995

The program produces β, the z statistic, and the associated level of significance for each item. The user may specify which items are to be included in the matching criterion, apply an external criterion, or allow the software to identify a purified criterion by selecting items from the full test. The software runs on a PC and can handle up to 150 items and 7,000 examinees per group. The software also allows for identification of nonuniform DIF and DIF in polytomous items. Contact: Dr. William Stout, Dept. of Statistics, University of Illinois, 101 Illini Hall, 725 S. Wright St., Champaign, IL 61820.

SPSS for Windows Advanced Statistics, M. J. Norusis, 1993

The SPSS Advanced Statistics package contains a logistic regression procedure. The program is user friendly and flexible. It allows for simultaneous or consecutive testing of the hypothesis of no uniform DIF and no nonuniform DIF. The program also makes it possible to enter multiple continuous and/or categorical matching variables. Logistic regression can also be run using SAS or BMDP. Contact: Marketing Department, SPSS, 444 North Michigan Ave., Chicago, IL 60611.

List of Reviewers for ITEMS Modules

The following individuals served as peer reviewers for the Instructional Materials for Educational Measurement (ITEMS) modules under the editorship of Professor George Engelhard, Jr., 1994–1997:

Linda Crocker
Mary Garner
Thomas R. Guskey
Ronald K. Hambleton
Richard M. Jaeger
Brenda Loyd
George Marcoulides
Barbara S. Plake
Nam Raju
Lori Rothenberg
Michael Zieky