9. Differential Item Functioning In Licensure Tests

Barbara S. Plake
University of Nebraska- Lincoln, bplake@unl.edu

DIFFERENTIAL ITEM FUNCTIONING IN LICENSURE TESTS

Barbara S. Plake

University of Nebraska-Lincoln

When test scores are used to make important decisions, as is typically the case with licensure tests, the validity of test score interpretations is extremely critical. The validity of the decision (e.g., pass or fail the licensure examination) relies heavily on the validity of the test score that is used in making the licensure decision. So, although validity is always a critical component in test score interpretation, it has increased importance when the score is used in high-stakes decision situations such as licensure testing.

Issues in validity for licensure tests have been addressed in Chapter 4 of this volume. The focus of this chapter is on techniques that have been developed for identifying one source of test interpretation invalidity: differential item functioning (DIF) by identifiable groups. The chapter begins with a discussion of what constitutes differential item functioning and under what circumstances differential item functioning poses a source of test interpretation invalidity. Next, various methods for identifying test items that function differentially are highlighted. This section focuses principally on multiple-choice test items although a separate subsection on applications of DIF methods with constructed-response type items is presented. The chapter ends with a conclusion section that makes recommendations for future developments in the area of identification of test items that function inappropriately for different subpopulations.

This chapter concentrates on the individual items that comprise the test, not on administrative or other aspects of testing that also might influence examinee test performance. Specifically, this chapter considers ways to identify items that function differentially for identifiable subpopulations. Other reasons for score performance differences (e.g., speeded conditions, administration medium, test anxiety/wiseness) are extremely important. However, these issues are beyond the scope of this chapter.

The focus of this chapter is on discussing different approaches that have promise for identifying items that function differentially in licensure tests. It is not the intent of this chapter to present step-by-step details on calculating these various methods. The reader should reference other books that present formulas for such calculations, particularly Berk (1982), Camilli and Shepard (1994), and Holland and Wainer (1993). Further, this chapter is not designed to be a comprehensive resource for DIF methods; instead, the chapter samples from these methods those techniques that are relevant or dominant in use for DIF analysis with licensure test applications.

WHAT IS DIFFERENTIAL ITEM FUNCTIONING?

It is expected that test items will show different performance across members of the examinee population. After all, if everyone performed exactly the same on the item, it would provide no useful information in differentiating those who qualify for licensure from those examinees who do not. Therefore, an item is not identified as functioning differentially based on overall differences in performance by examinees. When an item shows differences in performance for examinees in the population, however, the basis for that performance difference should be specifically that the examinees differ on the knowledge or achievement that is assessed by the item. When the item shows different performance as a function of differing levels on the trait the item is designed to assess, the item is functioning properly. However, when differences in performance are attributable to extraneous sources of variance, such as ethnic group membership, then the item is not functioning properly. If the item was scored as an operational item in the test, performance on the item could be a basis for invalid test interpretations.

Differential item functioning is often defined as differential item performance by subpopulations of examinees who are equal in the underlying trait measured by the test (Cole & Moss, 1989). To ascertain whether a test contains items that show DIF, many analytic methods are available to compare item performance by subgroups of examinees who have been matched on overall test performance.

Although any identifiable subgroup of examinees could be compared, typically DIF analyses have focused on detecting differential item performance for gender or ethnic groups. In most applications of the methods discussed in this chapter, two distinct groups of examinees are identified: the reference group and the focus group. In the study of DIF for ethnic groups, for example, the reference group is often white examinees and the focus group consists of members of a particular ethnic group, such as African-American examinees. For many to the methods discussed in this chapter, only two groups can be considered in the DIF analysis (e.g., males and females, white examinees and Hispanic examinees; low SES and high SES). In some instances, the methods can be generalized to more than two mutually exclusive groups; however, these extensions are beyond the scope of this presentation.

It is important to note that differential item performance, per se, is not *prima facie* evidence that the test item is biased. Bias is a judgment that may be made due to the presence of items in a test that show differential item performance by identifiable subgroups of examinees in the population. However, some sources of DIF by identifiable subgroups may be appropriate and contribute to valid test score interpretations. For example, on a broad-based licensure test for a discipline with subspecialties, differential item performance may be appropriate and expected by examinees with differential training in the subdisciplines. Therefore, differential item performance by some subpopulations of examinees does not necessarily warrant conclusions about item or test bias.

METHODS FOR DETECTING DIFFERENTIAL ITEM FUNCTIONING

Even when the best item writers are employed and the test development practices are excellent, there is the potential for inappropriate items to appear in the operational version of a licensure test. Most test developers desire to identify such items and eliminate them from the test score to improve the validity of test score interpretations. The purpose of this section is to identify several methods useful for identifying items that may be contributing to differential item performance. These methods are distinguished by when they are applied in the test process. The first set of methods is applied during the test development process. The second set of methods relies on test performance data by examinees. Illustrations of applications of these methods follow in the next section. Generalizations of these methods to tests that require examinee performances, as in constructed response tests or clinical sets, follow.

DIF METHODS DURING TEST DEVELOPMENT STAGES

Probably the best way to eliminate differentially functioning items from a licensure test is to use good test development practices. Through the table of specification (or test blueprint), all critical components that contribute to valid test score interpretations should be identified. These include, in addition to test content, appropriate levels of cognitive processing, and necessary levels of prerequisite skills. Therefore, test content areas that are deemed unnecessary should not be covered by the test questions. The items are written to command an appropriate level of cognitive processing and features such as readability level, test wiseness, and item flaws should have been considered in the item development process. A readability analysis could provide useful information about the level of reading skill needed to perform adequately on the test. Here is an example of a potential contributor to test score differences that may be appropriate: If reading at a specific level is relevant to the licensure decision, then examinees who differ on their reading skill should perform differently on the test questions. On the other hand, if minimal reading skill is needed, then a test with an elevated readability level would likely advantage good readers. Under those conditions, reading level would be a source of unwarranted differential item performance. Good test writing practices aid in eliminating unwarranted sources of test score variance, and therefore, in reducing the potential for differential item functioning by subpopulations in the examinee group.

A second approach used during the item development stage is to employ a panel of experts to review the test items for inappropriate characteristics. Often the panel consists of persons knowledgeable about the targeted subpopulations being considered in the differential item functioning analysis. These panel members are usually asked to review each item and identify items that have potential for being offensive or misleading to members of the targeted groups. Items so identified are typically revised or removed from the item pool.

DIF METHODS BASED ON ANALYSIS OF TEST PERFORMANCE

In order to use data-based DIF methods, a group of examinees must have taken the test under operational test conditions. Sometimes pilot data or pretest data are used to identify items that show differential item functioning. In order for these data to generalize to the operational administration, common administration features must be maintained.

These data-based methods seek to identify test items that show differences in test performance between members of identifiable subpopulations. It is important to remember, however, that it is not simply the difference in test scores between identifiable subgroups that signals a concern for differential item functioning. These identifiable groups may, in fact, differ in their knowledge or achievement the licensure test measures. If that is the case, this difference in test performance is a meaningful and warranted source of score interpretation. Instead, what indicates the presence of differential item functioning is differences in item performance between subgroups of examinees that have been matched on the knowledge or achievement measured by the test.

One important issue in the application of these analytic methods for identifying items that function differentially for matched subgroups of the examinee population is how to form the matched subgroups. Optimally, an external measure of the latent trait (or underlying construct or performance domain) would be used; however, that is almost never available (in fact, if such a valid and reliable alternative method existed, the licensure test probably would not be needed). Instead, most methods utilize the overall licensure test score as the matching criterion. Of course, this is potentially a source of invalidity because the matching variable consists of performances on the very items that are being investigated as suspect for contributing unwarranted score variance. Some of the methods address this problem through attempts to refine the matching criterion by eliminating those items that have been shown to have differential item performance (Clausen, Mazor, & Hambleton, 1993). Although this appears, logically, to be a needed step, reducing the number of items that contribute to the matching criterion weakens its reliability (Zwick, 1990). Therefore, this is not an accepted practice. Because the analytic methods are often used in tandem with methods used in the test development stages, the items that make up the total operational test often have already been subjected to one screening for sources of differential item functioning. It is hoped this serves to strengthen the use of the total test score as the matching criterion for these analytic methods.

Two general classes of analytic methods are presented: those that rely basically on classical test theory (CTT) and those that are founded in item response theory (IRT). The reader is referred to other chapters in this volume for fundamentals of these two theories.

CTT-Based Methods

Approaches that are based on classical test theory focus on item difficulty as a fundamental indicator of item performance. The subpopulations are matched on overall test score, or in test score ranges. Then the number of examinees in the identifiable subgroups correctly answering each item is compared. Three different variations of this approach are Scheuneman's Chi-Square, Log-linear analysis, and Mantel-Haenszel method.

Scheuneman's Chi-Square

This method, suggested by Scheuneman in 1975, begins with dividing the examinees into categories based on total test score (usually three to five categories are formed). For each item, Scheuneman's Index, $C2$, is computed as a function of the number of correct answers for members of each group, summed across the test score categories. As a test statistic, $C2$ asymptotically follows a chi-square distribution with degrees of freedom equal to the number of test score categories.

Several variations of this method have been proposed, including those by Camilli (1979) and Marascuilo and Slaughter (1981). The "full chi-square" method (Camilli, 1979) includes the number of incorrect as well as correct answers in the computation. These methods tend to produce very similar results; however, the sample size requirements for the full chi-square method are somewhat higher than those for Scheuneman's Chi-Square method.

Log-Linear Analysis

In applying log-linear approaches, nominal level data are all that is required. Three variables can be formed for a log-linear approach to identifying items showing differential item functioning: group membership (0 for reference group, 1 for focus group membership); total score category (typically three to five categories); and item response (0 for correct, 1 for correct). These variables form the bases of a three-way contingency table specified for each item in the test. Based on the specification of the models of interest, goodness-of-fit measures are then calculated (e.g., likelihood ratio chi-square, G^2). Significance test for differences in G^2 support conclusions regarding DIF. A model is specified containing terms (or components) reflecting possible sources of differential performance for examinee groups. This model, with each term adding sequentially to the others, forms a hierarchial model. The first term in the model focuses on the main effect of ability. The second term added to the model addresses the potential for a main effect difference between groups. The final, third term, then is sensitive to an interaction between group and ability. The process involves a sequential series of hypothesis tests, designed to assess the unique, additional contributions of individual components of a model to conclusions regarding differential item performance by examinee groups. If it is found through the sequential hypothesis testing procedure that the group and group by ability terms do not significantly improve the fit of the data to the model, it is generally concluded that no DIF exists. If the group term significantly improves the fit of the data to the model, then the conclusion is typically that uniform differences in item performance are present. It is only when the third, interaction term, provides a significant contribution to the fit of the data to the model that the interpretation of differential item performance is justified. More information on the log-linear approach to DIF can be found in Van der Flier, Mellenbergh, Ader, and Wijn (1984).

Mantel-Haenszel Method

The Mantel-Haenszel (MH) shows similarities to both the chi-square approaches and the log-linear methods presented above. Originally developed for use in medical applications, this method was introduced by Holland and Thayer (1986) as a technique for investigating differential item functioning.

The MH method is based on the odds ratio at each of the score points for the test. Two-by-two contingency tables are formed for each of the possible score values. Chi-square statistics are calculated at each of these score points, converted to odds ratios (similar to a proportion) in order to be on the same scale, and weighed by the product of the frequency of right and wrong responses divided by the frequency of responses. A significance test reveals those items for which it is more likely for a member of one group to get the item right than for a member of the other group.

Comparison of Scheuneman's Chi-Square, Log-linear, and Mantel-Haenszel Procedures

These three methods share a common characterization of the data as categorical. The two chi-square type methods, Scheuneman's Chi-Square and Mantel-Haenszel, differ primarily in the number of matched score categories. The Scheuneman method requires dividing the examinees into three to five categories based on total test score whereas the MH method creates distinct categories at every score point. Therefore, more data are needed for the MH method than for Scheuneman's Chi-Square. One important difference between the MH approach and the other two is that the MH method is not sensitive to inconsistency in differential item performance at differing score points in the distribution of test scores (e.g., interactions cannot be detected as in the log-linear method). Consider an item that revealed a complex pattern of performance difference such that low-scoring males were more likely to get the item right than their equally able low-scoring female counterparts, but for males with high overall test scores, they were less likely to get the item right than females with the same overall test score. The MH statistic is not sensitive to such inconsistent patterns of differential item functioning. If this kind of DIF was of interest, methods such as the log-linear approach would be more appropriate. Other methods, such as those based on item response theory (see below) are also sensitive to inconsistent patterns of DIF across the ability continuum and are attractive alternatives to the MH methods in those instances.

The chi-square based methods have been criticized for the use of gross categorization of test scores to form the ability groups. Obviously, the MH method, which employs as many ability groups as there are overall test score points, provides a more fine-grained analysis of item performance by ability for group members.

All three methods can be used with moderate numbers of examinees (e.g., 100 per identifiable subpopulation) and are relatively inexpensive to compute using standard statistical software packages. The log-linear method typically involves several analytical steps, which can result in higher cost than the other approaches based on classical methods.

Item Response Theory Based Methods

Item response theory provides a mathematical model that links performance on an item to specific features of the item (difficulty, discrimination, pseudo-guessing) with characteristics of the examinees (typically ability on the unidimensional trait being measured). This mathematical function may take on a variety of forms, depending on the specific item response theory model (1-, 2-, or 3 parameter models are frequently used in practice; for multiple-choice items, the 3-parameter model has been shown to have desirable features due to the inclusion of the pseudo-guessing parameter). Regardless of the specific item response model used, this mathematical relationship between item characteristic(s) and examinee ability can be described through an item characteristic curve (ICC). This curve represents the relationship between examinee ability and the probability the examinee will correctly answer the item. The key features from item response theory that show promise for detecting items that show differential item functioning are estimates of the item parameters (principally the difficulty parameter, b) and overall shape of the item characteristic curve.

IRT methods are very demanding in sample size and cost. Minimum sample size is generally given as 1,000 for the 3-parameter logistic model. Programs to perform the item calibrations and estimation of examinee ability can be difficult to implement and costly to run. Further, IRT models are based on the assumption of unidimensionality of the underlying latent trait being measured. Many licensure programs will find these requirements prohibitive for using item response theory approaches.

Wright, Mead, and Draba (1976) provide an index for quantifying the difference in b parameter values between two populations that is based on the Rasch model. In the Rasch model, the a parameter (discrimination) values are assumed to be invariant across the items in the test and no guessing is assumed. Therefore, the only reason for differences in item performance is the item's difficulty (i.e., the b parameter) and the examinee's ability (i.e., Θ). After calibrating the test items using data from the two groups and converting them to the same scale, Wright et al. suggest the calculation of an index that is approximately distributed as a t-statistic. They suggest using a critical value of plus or minus 2 to detect items that show differential item functioning.

Lord (1977, 1980) suggested an approach that involves a simultaneous test of the differences between the a and b parameters for two groups. This methods involves several calibrations: first with the two groups combined in order to get improved estimates of the *c* parameter. Then these *c* values are held constant and the *a* and *b* parameters are re-estimated for the two groups separately. These estimates would then need to be transformed to the same scale. An asymptotical chi-square test is available to test the simultaneous equality of the *a* and *b* parameters for the two populations of interest.

Linn and Harnish (1981) proposed a method that only requires one item calibration. Using the calibrations based on the total sample size, ability estimates (Θ) for members of the focal groups are determined. Then estimated test performance and actual test performance for focal group members are compared; DIF is assessed using a standardized difference score.

Rudner, Getson, and Knight (1980) proposed a method that is based on the item characteristic curves for the two groups. The items are calibrated separately for the two groups and then put onto a common scale. The area between these two ICCs is then determined. No statistical test is available to detect DIF using this approach. However, items showing large differences can be identified for further analysis or study.

Comparison of Item and Ability Estimation Approaches

Lord's method has not been used very much in empirical studies, in part due to the large demand for item calibrations (for total group and each of the comparison groups). Some research has shown that it does not agree well with other empirical methods for assessing DIF (Shepard, Camilli, & Averill, 1981). The Linn and Harnish method is promising as it only requires one calibration (for the total group). This is particularly important as many times there are insufficient numbers of members of the focus group to provide stable item parameter estimates. The Wright et al. method has been shown to confound other sources of model misfit with the DIF results, leading to inappropriate statements of DIF for certain items (Shepard, Camilli, & Williams, 1984). Rudner's approach is not used much in application due to the lack of appropriate statistical tests.

APPLICATIONS OF ANALYTIC METHODS TO TEST DEVELOPMENT

Test developers have used evidence about test items' performance to make decisions about test development, test scoring, and future test administration. The purpose of this section is to highlight some of these applications and to provide a critical analysis of their appropriateness for creating valid and reliable licensure examinations.

Golden Rule. One noteworthy application of item performance data for developing licensure examinations is what has come to be known as the "Golden Rule Method." This method resulted from an out-of-court settlement between the Golden Rule Insurance Company and Educational Testing Service. For more information about that case and the details of the settlement, see Phillips (1993).

Actually, this method does not incorporate differential item functioning data (that is why it was not identified as one of the methods for identifying items that perform differentially for subpopulations of examinees). Instead, this approach is based on overall performance differences by identifiable subgroups of examinees.

Based on pilot or pre-testing, the proportion of examinees correctly answering each item in each of the identifiable subgroups is determined (for example, Hispanic examinees and White examinees). When selecting test items for the operational test, items are selected first that show minimal between-group performance differences. Items that show large between-group performance differences are only considered for inclusion in the test if there are not other available items to satisfy the test specifications.

This method has received strong reactions from the measurement community. (See the 1987 issue of *Educational Measurement: Issues and Practice, 6,* for commentary by Faggen, Rooney, Linn & Drasgow, Bond, Jaeger, & Weiss.) Concerns focused on using empirical decisions, rather than table of specifications, for forming the test content. In 1987, then ETS President Gregg Anrig published a statement in which he details why ETS now feels the settlement was a mistake (Anrig, 1987).

Item Pool Maintenance. Many licensure test programs have item banks that are maintained over a period of years. Chapter 8 of this volume is specifically devoted to the development and maintenance of item banks for licensure test purposes. Typically, item information denoted in the bank consists of item classification, history of item administration and performance data, and occasionally information about DIF is detailed. Evaluations from panel members regarding appropriateness could also be maintained in the item bank data base. It is strongly recommended that DIF data be routinely gathered and reported in the item bank data base in order to monitor the status of the item with regard to differential item functioning. An item may have passed initial screening for DIF and subsequently be found to perform differentially for other, or even the same, identifiable subgroups. DIF analysis should be an ongoing part of the statistical analysis program.

Operational program applications. Even in the best of circumstances, when item development practices are exemplary and control/monitoring systems routinely in place, items occasionally will show differential performance on operational licensure examinations. The licensure administrator then has to decide on the best approach to deal with test scores that may not support valid and fair interpretations. First and foremost, any item that shows differential item functioning must be scrutinized for bias. If differential performance is supported by the construct being assessed, then the differential performance is valid, and the item should be maintained in the operational test score. However, if the differential item performance is an extraneous source of score variance, and not part of the construct being measured, serious problems exist when using the total test score for licensure decisions. One obvious solution would be to remove the item from the examination and rescore the test for all examinees. Although this has the advantage of removing the offending item from the test score, it has serious consequences. First, removing the item from the test changes the overall match of the test to the table of specifications. This is particularly worrisome for categories where limited numbers of items make up that component of the test. Further, changing the number, and character, of the items in the operational test will distort the cut score or standard previously established for determining those who pass the examination and those who do not.

This is another reason why differential item functioning is particularly crucial in licensure examinations. Not only are the decisions being made from performance on the examination high-stakes, and therefore, necessitate high standards for test validity, but decision reference points often are already in place and are subject to distortion when decisions to redesign the test occur after test administration. Test developers in licensure applications, therefore, must pay serious attention to those methods which are designed to diminish the presence of items that are potentially biased. Methods such of those described in this chapter are aimed at just that kind of effort.

APPLICATIONS OF DIF METHODS WITH PERFORMANCE-TYPE ASSESSMENT

The methods presented and discussed so far in this chapter are designed for use with multiple-choice items. Licensure programs have used performance-type assessments in their licensure tests for decades. These are frequently referred to as "clinical sets" in licensure testing applications. Unfortunately, there is very little known about the applicability or generalizability of these DIF methods to performance-type assessments.

The concern for differential item performance with performance-type assessments should be very high because there is additional potential for extraneous factors to influence test performance (Dorans & Schmitt, 1991; Miller, Spray, & Wilson, 1992; Oppler, Campbell, Pulakos, & Borman, 1992; Zwick, 1992). Often performance-type assessments are scored on a subjective basis. Many times, it is obvious to the scorer not only the quality of the performance, but the status of the examinee on many of the group identifiable traits used with objectively scored tests (ethnic group membership and gender, for example). Therefore, scorer subjectivity is a source of differential performance that was not present with multiple-choice tests.

In addition to scorer subjectivity, some forms of performance-type assessments may be more prone to tap construct-irrelevant factors. For example, in instances where the examinee brings prepared materials to the testing site (as in portfolios), there is the possibility that some candidates may have unequal access to support services or high quality materials. Although some advocates of the performance assessment movement speculate that the advent of performance-type assessments will reduce group differences and improve test fairness, some evidence suggests the opposite may in fact result (Dunbar, Koretz, & Hoover, 1991). Therefore, the need for strong methods for assessing potential differential performance on performance-type assessment tasks is extremely high.

When developing performance-type assessments, tasks rather than items are the units that are scored. If the performance-type tasks yield dichotomous performance outcomes (right/wrong, for example) then the methods described above will work. It is the polychotomous nature of the score scales the leads to problems in generalizing the current methods to performance-type assessments. Some of the issues that need to be addressed when generalizing DIF methods to polychotomously scored tasks are: (a) How should the matching variable be defined? and (b) What analysis should be used to ascertain the presence of differential task functioning?

With performance-type tasks, typically fewer tasks make up the assessment. Therefore, there are fewer data points to use when forming the matched groups. This reduces the reliability of group categorization decisions. Zwick, Donoghue, and Grima (1993) report on a simulation study testing the efficacy of several strategies for forming matched groups for the purposes of differential task functioning analysis. These authors also provide some suggestions for extensions of the MH method to polychotomously scored items. These methods show promise for applications with performance-type tasks used in licensure testing.

CONCLUSIONS

The purpose of this chapter was to discuss differential item functioning in licensure tests. The high-stakes nature of licensure testing creates an environment where validity of licensure test score interpretations (particularly as they relate to licensure decisions) is extremely crucial. Factors that improve the validity of licensure test scores should be enhanced and those factors that decrease the validity of interpretations from licensure tests should be removed or reduced as much as possible. Factors that are irrelevant to the construct being measured, and the licensure decision being made, are examples of factors that should be removed from the test scores.

One way of identifying such task-irrelevant factors is through differential item functioning analyses. The purpose of these methods is to draw attention to items that show unexpected differences in performance across equally able members of identifiable subgroups of the candidate population.

The methods discussed in this chapter show promise for aiding in the removal or reduction of factors irrelevant to the construct being assessed by the licensure test. However, these methods are typically only applicable to dichotomously scored assessments. Much attention is needed in the development of DIF methods useful with performance-type assessments, such as clinical sets and portfolio assessments.

In addition to concentrated efforts needed in the area of polychotomously scored assessments, better theoretical bases are needed for explaining extraneous sources of score variance. It is one thing to find items in a test that show differential item functioning between identifiable subgroups of the candidate population. It is quite another to be able to reason whether this shown difference is part of the construct being assessed or a source of test interpretation bias. Empirical methods are only useful in singling out items that show unexpected score differences; theory is needed to understand and improve interpretations based on these empirical results. Recent work by O'Neill and McPeek (1993) and Schmitt, Holland, and Dorans (1993) show promise in contributing to the theory of differential item functioning for identifiable subpopulations. With a theory to rely upon, test developers will have a foundation to use in developing test questions that, by design, reduce unwanted sources of test score differences between subgroups. Until we reach this level of sophistication, the empirical results will drive these decisions.

Only those methods that direct attention to performance differences between matched subgroups were discussed at length in this chapter. Many earlier methods that were based simply on differences in overall group performance between identifiable subgroups of the candidate populations (such as the transformed item difficulty method, the Golden Rule procedure) were not considered as true DIF methods. Two categories of empirical methods were presented, those based on classical test theory and those from item response theory.

Licensure testing programs with large examinee populations have the luxury of more choice when considering empirical DIF methods. The CTT approaches are amenable to both small and large testing programs and those with large and small testing support budgets. Only testing programs with large examinee populations to draw from, and relatively large human, computer, and fiscal support systems will be able to use the IRT-based methods. Recent research has shown that comparable results often occur between these two methods (Hambleton & Rogers, 1989). Another issue in deciding between CTT- and IRT-based methods is the degree to which the licensure decision is based on a unidimensional construct. IRT methods, as presented in this chapter, assume an underlying unidimensional construct. Many licensure areas consist of subcategories or subdisciplines that may not be strongly unidimensional as a set. These issues must be addressed before a decision about the methods is finalized.

Licensure testing, unlike other kinds of testing, typically ends with a final decision of pass or fail. The decision rule is often set in advance and is based on an analysis of the licensure test performance that is deemed sufficient for a pass decision. The cut score decision, therefore, is also inextricably tied to the validity of interpretations based on candidate performance on the licensure test. The validity of these decisions is linked to the validity of the interpretations that are made as a function of the candidates' test scores. Task-irrelevant influences on test scores, therefore, are doubly dangerous in licensure testing: They affect the validity of the test score and they affect the validity of the cut score. It is, therefore, extremely critical that licensure tests are scrutinized for unwarranted sources of test performance. Differential item functioning methods provide an approach for identifying potential sources of test invalidity. In the environment of high-stakes licensure testing the costs of errors are extremely high; DIF provides a means to purification of the test score to match more directly those knowledges, skills, and abilities that are salient to the licensure decision.

REFERENCES

Anrig, G. R. (1987, January). "Golden Rule": Second thoughts. *APA Monitor,* p. 3.

Berk, R. A. (Ed.). (1982). *Handbook of methods for detecting test bias.* Baltimore: Johns Hopkins University Press.

Bond, L. (1987). The Golden Rule settlement: A minority perspective. *Educational Measurement: Issues and Practice*, *6*, 18-20.

Camilli, G. (1979). *A critique of the chi square method for assessing item bias.* Unpublished manuscript, University of Colorado, Laboratory of Educational Research, Boulder.

Camilli, G., & Shepard, L. A. (1994). Methods for identifying biased test items. *Measurement methods for the social sciences* (vol. 4), Thousand Oaks, CA: Sage Publications.

Clausen, B., Mazor, K., & Hambleton, R. K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education*, *2*, 269-280.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R.,L. Linn (Ed.), *Educational measurement* (3rd ed.; pp. 201-219). New York: Macmillan Publishing Co.

Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach* (ETS Research Report 91-47). Princeton, NJ: Educational Testing Service.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, *4*, 289-303.

Faggen, J. (1987). Golden Rule revisited: Introduction. *Educational Measurement: Issues and Practice*, *6*, 5-8.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education*, *4*, 313-334.

Holland, P. W., & Thayer, D. T. (1986). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 12-145). Hillsdale, NJ: Lawrence Erlbaum and Associates, Inc.

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

Jaeger, R. M. (1987). NCME opposition to proposed Golden Rule legislation. *Educational Measurement: Issues and Practice*, *6*, 21-22.

Linn, R. A., & Drasgow, F. (1987). Implications of the Golden Rule settlement for test construction, *Educational Measurement: Issues and Practice*, *6*, 13-17.

Linn, R. A., & Harnish, D. (1981). Interactions between item content and group membership in achievement test items. *Journal of Educational Measurement*, *18*, 109-118.

Lord, F. M. (1977). A study of bias using item characteristic curve theory. In N. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19-29). Amsterdam: Switts & Vitlinger.

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

Marascuilo, L. A., & Slaughter, R. E. (1981). Statistical procedures for analyzing item bias based on chi square statistics. *Journal of Educational Measurement, 18,* 105-118.

Miller, T., Spray, J., & Wilson, A. (1992, July). *A comparison of three methods for identifying nonuniform DIF in polytomously scored test items.* Paper presented at the annual meeting of the Psychometric Society, Ohio.

O'Neill, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255-276). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc..

Oppler, S. H., Campbell, J. P., Pulakos, E. D., & Borman, W. C. (1992). Three approaches to the investigation of subgroup bias in performance measurement: Review, results and conclusions. *Journal of Applied Psychology, 77*, 201-217.

Phillips, S. E. (1993). The *Golden Rule* remedy for disparate impact of standardized testing: Progress of regress? *Education Law Reporter*, pp. 383-427.

Rooney, J. P. (1987). Golden Rule on "Golden Rule." *Educational Measurement: Issues and Practice, 6*, 9-12.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). Biased item detection techniques. *Journal of Educational Statistics, 5,* 213-233.

Scheuneman, J. S. (1975, April). *A new method of assessing bias in test items.* Paper presented at the meeting of the American Educational Research Association, Washington, DC.

Schmitt, A. P., Holland, P. W., & Dorans, N. J. (1993). Evaluation hypotheses about differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp.255-276). Hillsdale, NJ: Lawrence Erlbaum and Associates.

Shepard, L. A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6*, 317-375.

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

Van der Flier, H., Mellenbergh, G., Ader, H. J., & Wijn, M. (1984). An iterative item bias detection method. *Journal of Educational Measurement, 21*, 131-145.

Weiss, J. (1987). The Golden Rule bias reducation principle: A practical review. *Educational Measurement: Issues and Practice, 6*, 23-24.

Wright, B. D., Mead, R. J., & Draba, R. (1976). *Detecting and correcting test item bias with a logistic response model.*(Research Memorandum No. 22). Chicago: University of Chicago, Department of Education, Statistical Laboratory.

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. (1992, April). *Differential item functioning analysis for new modes of assessment.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning performance tasks. *Journal of Educational Measurement, 30*, 223-251.