ED 353 301 TM 019 365

AUTHOR Jones, Russell W.; Hambleton, Ronald K.
TITLE Recent Advances in Psychometric Methods.
INSTITUTION Massachusetts Univ., Amherst. Laboratory of Psychometric and Evaluative Research.
REPORT NO LPER-R-233
PUB DATE [Nov 92]
NOTE 22p.
PUB TYPE Information Analyses (070)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Cognitive Psychology; Computer Assisted Testing; Generalizability Theory; Item Bias; *Item Response Theory; *Measurement Techniques; *Psychometrics; *Technological Advancement; Test Construction; Test Format
IDENTIFIERS Empirical Research; Performance Based Evaluation

ABSTRACT

Increasing social awareness of the need for accurate and fair testing has combined with technological developments to produce dramatic advances in psychometric models, methods, and procedures. Three of the most important new directions within the measurement field concern the following areas: (1) item response theory; (2) item bias or differential item functioning (empirical methods and judgmental approaches); and (3) performance assessment and the associated development of new item formats. Recent developments in each of these three areas are described. Future applications of item response theory include extensions to current models to handle polytomous and multidimensional data. One of the most exciting areas within performance testing involves more valid assessments of higher-level cognitive skills. Important future directions for differential item functioning research include the development of techniques for the analysis of items that are scored polytomously and improvements in current judgmental methods. The following are other areas within the measurement field in which important advances may be expected: (1) expanded uses of generalizability theory in the analysis of educational data; (2) increased use of computer technology in test development and administration; and (3) the integration of cognitive psychology and psychometric methods to evolve new types of measurements that are especially useful to teachers. Included are 37 references. (Author/RLC)

Recent Advances in Psychometric Methods

by

Russell W. Jones and Ronald K. Hambleton
University of Massachusetts at Amherst

Abstract

Increasing social awareness of the need for accurate and fair testing has combined with technological developments to produce dramatic advances in psychometric models, methods, and procedures. Three of the most important new directions within the measurement field concern (1) item response theory, (2) item bias or differential item functioning, and (3) performance assessment and the associated development of new item formats. Recent developments in each of these three areas are considered in the paper.

Lab Report 233

Recent Advances in Psychometric Methods[1]

by

Russell W. Jones and Ronald K. Hambleton
University of Massachusetts at Amherst

Advances in psychometric models, methods, and procedures since the publication of Lord and Novick's Statistical Theories of Mental Test Scores in 1968, have been substantial. Also, technological breakthroughs such as the widespread availability of powerful personal computers have combined with increasing social pressure for fair and accurate assessment. Accurate and fair measurement has gone from being primarily a concern of educational systems to a concern of governments, credentialing organizations, industry, the military, litigators, and politicians. With the increased public and private interest in measurement issues has come a concomitant increase in funding for psychometric research. In turn, this has resulted in the development of more measurement techniques which has further increased public and private awareness of the value of accurate and fair assessment and the cycle continues.

The proof of this enormous growth is evident in the dramatic increase in the number of professional journals which publish articles concerning measurement, and in the quantity of papers presented at an increasing number of professional conferences and symposia at the national and international levels which deal with issues of a psychometric nature. Add to this the publication of a vast number of books addressing measurement concerns, and computer programs which have been developed to facilitate the application of

[1] Laboratory of Psychometric and Evaluative Research Report No. 233. Amherst, MA: University of Massachusetts, School of Education,

new measurement technology, and it readily becomes apparent that psychometric advances have grown dramatically in the last 20 years.

Examples of changes in testing practices are readily available. Consider, for example, computerized adaptive testing (Wainer, et al., 1990). This recent innovation has preserved the advantages, validity, and breadth of applications of the paper-and-pencil test, but, in addition, has dramatically changed the manner of item presentation and increased the efficiency of feedback. In addition, test lengths can often be shortened by 50%, without any loss in measurement precision. Similarly, the detection of potentially biased items, which, at one time, was carried out primarily by judgmental methods, has mushroomed into the application of a wide array of both empirical and judgmental methods. Some of the procedures are simple, many others are complex (see, for example, Scheuneman & Bleistein, 1989). In 1992, item bias detection analyses were nearly as common in the test development process as classical item analyses to detect potentially flawed test items. Other important changes include item-examinee sampling (i.e., matrix sampling) in program evaluation, the use of item response theory models in test design, test score equating, and diagnostic assessment, and the movement from group-centered test construction and interpretations (i.e., norm-referenced testing) to content-centered test construction and interpretations (i.e., criterion-referenced testing). Other prominent examples of modern measurement advances will be found in Linn's Educational Measurement (1989).

Of the many new directions in the measurement field, three of the most important will be considered in this paper: (1) item response theory, (2) item bias or differential item functioning, and (3) performance assessment and the development of new item formats. Developments in each of the three areas will be considered along with some predictions about the future.

Item Response Theory: A Theory for the 1990s

Although the psychometric field has been adequately served by classical test theory and associated models for a long time (see, for example, Gulliksen, 1950; Lord & Novick, 1968; de Gruijter & van der Kamp, 1984), there has been a shift in focus in recent years away from classical theory and towards item response theory (IRT). Over the last 20 years, item response theory (sometimes referred to in the literature as latent trait theory or item characteristic curve theory) has become a popular area for psychometric advances.

There are many well-documented shortcomings of classical test models and measurement procedures in the psychometric literature. One shortcoming is that the values of such classical item statistics as item difficulty and item discrimination depend on the particular examinee samples in which they are obtained. The mean and standard deviation of ability scores in an examinee group affect the values, often substantially, of the item statistics, and the reliability and validity statistics, too. One consequence of *sample dependent* item statistics is that these item statistics are only useful when constructing tests for examinee populations which are similar to the sample of examinees from which the item statistics were obtained. Unfortunately, it is all too often unreasonable to assume that a field test sample is representative of the population of examinees of interest.

A second shortcoming of classical test models is that comparisons of examinees on the test score scale are limited to situations where examinees are administered the same (or parallel) tests. The seriousness of this shortcoming becomes evident when it is recognized that examinees often take different forms of a test, or even different sections within a test. For example, one medical board in the United States requires examinees to take a "core section" and then three of six additional sections of the test. Examinees are compared using scores based on a test consisting of the core and three optional sections. Since the sections are not equally difficult, and there are 20 different combinations of three sections possible, comparisons among examinees become difficult. In fact, it is not fair to require the same passing score for examinees who have been administered tests which differ, perhaps substantially, in their difficulty. When several forms of a test which vary in difficulty are used, examinee scores across non-parallel forms are not comparable without resorting to complex equating procedures which may also contain inherent problems and difficulties.

Computer adaptive tests are being considered by a number of school districts, major testing agencies, and credentialing boards at the present time. Advantages include reduced testing time, increased test security, flexibility in test scheduling, quick score reporting, and increased measurement precision for many examinees. But, again, the non-equivalence of test forms makes comparisons among examinees or comparisons of test scores to passing scores difficult without the use of very complex and tedious to apply classical equating methods. Other shortcomings of classical test models have been described by Hambleton and Swaminathan (1985).

Item response theory purports to overcome the shortcomings of classical test models by providing an ability scale on which examinee abilities are independent of the particular choice of test items that they are administered. Item response theory postulates that (1) underlying examinee performance on a test is a single ability or trait, and (2) the relationship between examinee performance on each item and the ability measured by the test can be described by a monotonically increasing curve. This curve is called an item characteristic curve and it provides the probability of examinees at various ability levels answering the item correctly. Examinees with more ability have higher probabilities for giving correct answers to items than lower ability examinees. Item characteristic curves for dichotomously scored items are typically described by one, two, or three parameters.

Within an IRT measurement framework, ability estimates for an examinee obtained from tests which vary in difficulty will be the same, except for the usual measurement errors. This invariance feature in the ability estimates is obtained by incorporating information about the items (i.e., their statistics) into the ability estimation process. Furthermore, item statistics are defined on the same scale as examinee ability, and, in theory, item statistics are independent of the particular examinee sample used in obtaining the estimates. Item parameter invariance is accomplished by defining item parameters in a way that the examinee ability distribution does not influence the item parameters or their interpretations. Finally, by providing measurement errors associated with individual IRT ability estimates, rather than producing a single estimate of error (i.e., the standard error of measurement) and using it with all examinees, another of the criticisms of the classical test model can be overcome.

Item response theory models provide both invariant item statistics and ability estimates. Both features are of considerable value to test developers because they open up new directions for assessment, such as adaptively administered tests and item banking. Of course, the feature of invariance will not always be present. Item and ability parameter invariance will be obtained when there is (at least) a reasonable fit between the chosen IRT model and the test data. Not surprisingly, then, considerable importance is attached to determining the fit of an IRT model to the test data.

Presently, in some countries (for example, the United States and Canada), item response models, especially the one- (often called the "Rasch model" in the measurement literature) and three-parameter logistic models, are receiving increasing use from testing agencies, certification/licensure test agencies, government departments, state departments of education, and the Armed Services, in test item selection, in addressing item bias, and in equating and reporting test scores. Measurement specialists are also exploring the uses of item response theory in preparing computerized banks of test items and in computer-administered and computer-adaptive tests. Detailed descriptions of IRT procedures and applications are available in Hambleton, Swaminathan, and Rogers (1991) and Hambleton and Swaminathan (1985). At this time, it seems reasonable to predict that item response theory will continue to have a growing and substantial influence on the development and application of most, if not all, measurement applications.

The various IRT applications have been sufficiently successful that researchers in the IRT field have now shifted their attention from a comparative consideration of IRT model advantages and disadvantages in relation to classical test models, to consideration of the technical problems involved with IRT including goodness-of-fit investigations, model selection, parameter estimation, and the required steps for performing specific applications. Certainly some issues and technical problems remain to be solved in the IRT field, but it would seem that item response model technology is more than adequate at this time to serve a variety of uses.

The Detection of Differential Item Functioning

A widely accepted definition of differential item functioning (DIF) today is that an item is demonstrating DIF if examinees of equal ability, but from different subgroups (for example, gender or culture), do not have equal probability of correctly responding to that item (Hambleton & Rogers, 1989). Differential item functioning (DIF) is becoming an increasingly preferred term to item bias because this term focuses on the results of the analytical procedure rather than making inferences about the cause of the effect, as is the case when the term "bias" is used. The difference between the concepts of bias and DIF being made in the measurement literature today may be clarified by the following brief example. It is beyond dispute that there exist differences between males and females. Some of these differences result, for example, in the obvious disparity in height and weight, i.e., males are generally heavier and taller than females. But do we conclude that the weigh scale and ruler are biased? No, clearly it would be incorrect to conclude that the messenger (i.e., the measuring instruments) is acting in favor of one sex or the other. Similarly, it is incorrect to conclude that a test item exhibits bias because two subgroups of interest perform differently, even if they are matched on ability. It is more accurate to conclude that the item exhibits differential item functioning. This expression makes it quite clear that a difference does exist but does not involve an implied quantitative judgement in favor of either subgroup.

Due to numerous powerful ethical, political, and social pressures the detection of DIF has become an important concern of today's measurement industry and a priority for testing agencies, test publishers, and school districts. To this end, many methods of DIF detection have been developed each of which has inherent advantages and disadvantages and most of which are used to a greater or lesser extent in the measurement industry today. Methods of DIF detection may be categorized as either empirical or judgmental.

Empirical Methods

Empirical methods of DIF detection involve the application of statistical procedures to measure subgroup item performance. If differential performance between subgroups is detected then, where appropriate, statistical tests of significance can be applied to determine if the observed differences are attributable to reasons other than chance. An inherent assumption within empirical methods of DIF detection is that for those items where a significant statistical difference exists, the cause is in some rational way related to those characteristics which define the subgroups (Stahl, Lunz, & Snyder, 1990). Empirical procedures for the detection of DIF may be further divided into three categories; approaches using classical test theory, approaches using item response theory, and approaches using chi-squared methods.

A number of approaches to the detection of DIF have been developed from the principles of classical test theory. These approaches utilize examinee observed scores and usually involve a comparison of p-values (i.e., classical item difficulty values) between the subgroups of interest. Hence, classical methods are sample dependent. Also the assumption is made that the observed score is a reliable and valid measure of the ability that the test is intended to measure (Scheuneman & Bleistein, 1989). Methods which utilize classical test theory include ANOVA and correlational approaches, the transformed item difficulty (TID) or delta plot method (Angoff & Ford, 1973), item discrimination procedures (e.g., Green & Draper, 1972), partial correlation methods (Stricker, 1982), and contingency table approaches (Scheuneman, 1979). But little more needs to be said, as these methods have been found to be theoretically unsatisfactory since they fail to take real group differences in true ability into account (Hambleton & Swaminathan, 1985).

Approaches to DIF detection which operate within the framework of item response theory not only overcome the shortcomings of classical test theory, but also gain several desirable characteristics inherent within IRT. Moreover, a great deal of research (e.g., Ironson & Subkoviak, 1979; Merz & Grossen, 1979; Rudner, Getson, & Knight, 1980; Shepard, Camilli, & Averill, 1981; Subkoviak, Mack, Ironson, & Craig, 1984) has consistently indicated that methods of DIF detection which apply IRT are superior to those methods which apply classical measurement theory.. IRT provides a useful empirical framework for relating the probability of a correct response to examinee ability level (Cole, 1981). Essentially, these DIF detection procedures involve obtaining test item parameter estimates for the item characteristic curve (ICC). ICCs are obtained for each subgroup and compared. If DIF is not present, then the ICCs for each subgroup should be identical. Complexity of the approach and unstable DIF statistics are two of the main drawbacks.

A number of DIF detection procedures make use of a chi-square value as an index of DIF. These procedures are often collectively referred to as chi-square methods. Included within this group is perhaps the most popular method of DIF detection currently in use, the Mantel-Haenszel procedure. Originally developed by Mantel and Haenszel (1959) for the purposes of analyses within the biomedical research industry, this technique was adapted for measurement by Holland and Thayer (1988), and has been used in DIF studies by Bennett, Rock, and Novatkoski (1988), Zwick and Ercikan (1989), and many others. Essentially the MH procedure is a chi-square test with one degree of freedom which compares the item performance of the majority and minority groups across different score groups (Raju, Bode, & Larsen, 1989). Subgroups of interest are assumed to be comparable at each score group level, and if the item is nondifferentially functioning then there should be no relationship between performance on the item and group membership and hence the performance of both groups should be equal (Raju, Bode, & Larsen, 1989). If performance by the two groups is not equal then DIF is present. The popularity of this technique may be attributed to its intuitive appeal and the ease by which it can be applied. Other techniques which employ chi-square values include the modified contingency table method proposed by Veale (1977) and the logistic regression procedure (Swaminathan & Rogers, 1990). A main difference between DIF studies with IRT and the MH procedure is the independent variable: In IRT, it is the ability score; in the MH procedure, it is the test score. Also, a statistical test is readily available with the MH procedure.

Judgmental Approaches

Essentially, judgmental approaches to DIF detection require a number of judges to review a set of test items. The judges may act independently. However, in providing their ratings, frequently, groups of judges are brought together for useful and productive discussions. In addition to the issues of stereotyping and fair representation, the focus of the judges' attention is directed towards monitoring equal opportunity for the acquaintance of subgroups with item content and ensuring the overlap of items with instruction (Tittle, 1982). In this way, judgmental approaches to DIF detection perform the role of establishing construct validity (Tittle, 1982), content validity, and face validity, all valuable and important aspects of test development. Specifically, judgmental reviews must identify the following: (1) stereotyping (even though stereotyping does not usually cause DIF, it is still an undesirable property and as such should be removed), (2) testing of content not within the realm of experience of a particular subgroup, and (3) subgroups not having an equal opportunity to learn particular content.

Judgmental approaches to the detection of differentially functioning items have several advantages over other methods. First, the use of judgmental approaches to DIF detection can be considerably cheaper than statistical approaches which often require expensive and time consuming data collection and computer analyses. Second, the use of appropriate and suitable judges to review test items can provide the test developer with credible face validity evidence. In the social, ethical, and politically conscious arena in which test developers operate, the opportunity to have judges from focal groups of interest review and judge items as DIF or non DIF provides the test developer with a face validity that is both valuable and powerful. Third, judgmental reviews may be performed prior to any examinees being administered the test items. Inappropriate items can then be deleted or modified before the items are field tested. A judgmental review is particularly valuable for those test developers who are unable to field test their items. Fourth, if judges are selected who are familiar with curricular content, then judges can also be asked to check that the test as a whole exhibits content validity.

A number of disadvantages are inherent with the judgmental approach to DIF detection. Most notable of these is the frequent failure of statistical and judgmental approaches to agree on which items are flagged as exhibiting DIF. Other disadvantages include: (1) the expense incurred in bringing judges together, (2) the time and expense involved in training judges, and (3) the susceptibility of judges to fatigue, boredom and other conditions which may interfere with the validity and reliability of judgments.

Summary

The choice of the DIF detection procedure required for any particular test should probably consider the available computer facilities, available examinee samples and their sizes, the precision required as a function of the importance of the decision to be made, and the audience to whom the results are to be directed (Scheuneman & Bleistein, 1989). Also, multiple methods are usually advantageous because different methods, empirical and judgmental, can be helpful in identifying problematic items.

Performance Assessment and New Item Formats

One of the "hottest" topics to emerge in recent years in testing in the United States is "authentic testing." Although proponents treat authentic testing as if it were a totally new topic, in reality this concept has existed for some time under the more preferable term performance assessment (preferable because the term performance assessment avoids the implication that all other forms of measurement are not authentic). Performance-based assessment incorporates the following features (Horvarth, 1991): (1) by placing the emphasis on performance they assess not only what a student knows, but also what a student can do, (2) whenever possible, direct methods of assessment are used (e.g., speaking skills should be assessed through oral presentations, and writing should be assessed through writing samples), (3) they should incorporate a high degree of realism, and (4) in order to better mirror real-world situations, assessments may include activities for which there is no correct answer, assessment of groups rather than individuals, evaluation may continue over an extended period of time, and/or self evaluation is permitted.

The concept of performance assessment is clearly worthy of incorporation into current testing practices. However, although multiple-choice, true-false, essay, and short answer item formats have much to contribute to effective performance assessment, other item formats will need to be developed which preserve the necessary levels of validity and reliability while addressing the practical essence of performance assessment. To this end, a number of new item formats have already been implemented in some testing programs and the following are useful examples from the United States.

One of the most useful of these is the standardized patient format, which is suitable for performance assessment for those tasks which involve contact between the examinee and others (van der Vleuten & Swanson, 1990). For example, this new format may have potential in assessing the competencies of doctors, nurses, dentists, lawyers, etc. In the case of medical testing, an actor or actress is trained to display specific symptoms of a particular condition, i.e., to become a "standardized patient." The patient is then introduced to the candidate who must examine and question the patient to determine the nature of the problem and prescribe appropriate treatment (Swanson & Stillman, 1990). Certainly this test meets the desired requirement for a practical performance by the candidate. From a psychometric point of view, this format offers the additional challenge of scoring. For example, how should an examinee (i.e., candidate for a licence in medicine) be scored if the treatment prescribed is fatal? Clearly, the candidate may be expected to fail the item, but should consideration also be given to his/her failing the entire test?

Another format which has received some use in medical testing is called computer-based problem-solving. This format makes use of computers to present an examinee with information about a hypothetical problem and, again, the examinee is required to find a solution (Melnick, 1990). While computer-based problem-solving does not offer the same degree of reality as the standardized patient, it does offer the advantages of greater comparative economy and consistency when providing the examinee with information. Such is not always the case with trained actors who, in addition to being expensive and time-consuming to train, may also be prone to fatigue and to personal likes or dislikes of examinees which can interfere with the consistency of presentation during the course of the testing of consecutive examinees. Applications of this format to the assessment of examinees in many professional fields, such as air traffic controllers, meteorologists, engineers, etc., and in education, in the content fields of mathematics and science, appear promising. Cost of development will limit the number of applications. Scoring and validity assessment will be two of the technical problems that must be overcome.

These computer-based situations contain all the advantages inherent within any computer-based testing system. They include the potential for computerized adaptive testing, greater flexibility regarding when and where examinees are administered tests (i.e., examinees may no longer need to be brought together to a central location for the administration of a single paper-and-pencil test, but instead may be able to report to a regional center to sit a computer-generated version of a test), and the creation of computerized item banks.

Audio-visual context setting is a very realistic item format which creates the performance scenario through the use of audio and/or visual stimuli. Technological advances, such as the widespread availability of the video camera/recorder and VCR, have made it comparatively easy to film scenarios and present them to examinees in order to "set the stage" for a performance test. Indeed, the responsive ability of videodisc technology, which can provide the operator with a choice of options and immediate feedback, make it possible for examinees to be presented with an audio-visual representation of a scenario from which they are required to make a decision. Feedback regarding the outcome of this decision can be immediately provided to the examinee in the form of a modified scenario which continues to unfold until the examinee is required to make another decision, whereupon immediate feedback is again provided. A test can consist of one or several of these realistic scenarios through which the examinee is required to successfully work in order to provide a satisfactory performance.

Not all new formats place a hravy emphasis on the use of modern technology or acting. One such example is multiple-choice with justifications (Dodd & Leal, 1988; Murray, 1990). This item format retains many of the benefits of the multiple-choice format but gathers additional information from the examinee through the requirement that the examinee provide a brief written justification of his/her answer choice. This item format is particularly useful for formative assessments because valuable information is obtained regarding incorrect reasoning, misconceptions, and gaps in the knowledge base of examinees.

Another format which builds on common item formats rather than recent technology is the figural response format (Martinez, 1991). These items provide stimulus material in the form of graphical, diagrammatical, or pictorial illustration(s) and an examinee is typically required to answer a question, or series of questions, by recording their responses on the illustration. Because these responses require greater detail than those required by a multiple-choice item, greater insight into the cognitive reasoning behind an examinee's choice is obtained. This item format also has the capability of being machine scored.

The traditional "patient management problem" format (McGuire, Solomon, & Bashook, 1976) has been modified to evolve into the latent image item. This format is based on similar principles to that which will likely make the videodisc a valuable assessment tool in the near future, except that instead of a scenario being presented to the examinee via a series of related audio-visual presentations, the scenario is presented through a series of related written questions. Each question is answered by the examinee selecting a specified number of options and, through the use of a special pen (Singer, 1985), each option reveals additional information to the examinee. Armed with this additional knowledge, the examinee moves on to the next question which is itself determined by the response to the previous item. Items such as these are capable, to some extent, of mimicking real life situations where performance frequently relies on small packets of information provided via feedback, rather than the typical test situation where an examinee is provided with all the information he/she requires to answer an item at the onset.

Performance assessment is increasingly being incorporated into tests in the United States through the addition of a practical component to traditional multiple-choice tests. These components are called performance items and require examinees to perform a practical task. The task may be as straightforward as writing an essay or conducting a science experiment, or as complex as successfully preparing and delivering a lesson on solving linear equation as part of a teacher certification test.

Development and adoption of the item formats described in this section are not likely to replace the more typical formats. Instead, these new formats will likely be used to augment true-false, multiple true-false, matching, and multiple-choice formats to create assessments that are more effective in evaluating the skills and abilities in which the examiner is interested. In particular, the development of new item formats holds some promise that soon the assessment of higher order cognitive skills such as reasoning, problem-solving, and critical thinking will be possible. These are the outcomes in which many educators profess to be most interested.

Conclusion

Psychometric methods is a rapidly expanding discipline, and many areas besides the three which were reviewed here have contributed significant changes to measurement practice and will likely continue to do so in the future. Future applications of item response theory include extensions to current models to handle polytomous and multidimensional data. One of the most exciting areas within performance testing involves more valid assessments of higher-level cognitive skills. Important future directions for differential item functioning research include the development of techniques for the analysis of items which are scored polytomously, and improvements in current judgmental methods. Other areas within the measurement field in which important advances may be expected include (1) expanded uses of generalizability theory in the analysis of educational data, (2) more use of computer technology in test development and administration, and (3) the integration of cognitive psychology and psychometric methods to evolve new types of measurements which are especially useful to teachers.

References

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10, 95-106.

Bennett, R. E., Rock, D. A., & Novatkoski, I. (1988). Clusters as the unit of analysis in differential item functioning (Report No. RR-88-70). Princeton, NJ: Educational Testing Service.

Cole, N. S. (1981). Bias in testing. American Psychologist, 36, 1067-1077.

de Gruijter, D. N. M., & van der Kamp, L. J. Th. (1984). Statistical models in psychological and educational testing. Lisse, The Netherlands: Swets and Zeitlinger.

Dodd, D. K., & Leal, L. (1988). Answer justification: Removing the trick from multiple choice questions. Teaching of Psychology, 15, 37-38.

Green, D. R., & Draper, J. F. (1972, December). Exploratory studies of bias in achievement tests. Paper presented at the meeting of the American Psychological Association, Honolulu. (ERIC Document Reproduction Service No. ED 070 796.)

Gulliksen, H. (1950). Theory of mental tests. New York: John Wiley.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT and Mantel-Haenszel methods. Applied Measurement in Education, 2(4), 313-334.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer Academic Publishers.

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

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.

Horvarth, F. G. (1991, April). Assessment in Alberta: Dimensions of authenticity. Paper presented at the meetings of the NATD/NCME, Chicago, IL.

Ironson, G. H., & Subkoviak, M. J. (1979). A comparison of several methods of assessing item bias. Journal of Educational Measurement, 18, 209-225.

Linn, R. L. (Ed.). (1989). Educational measurement (3rd ed.). New York: Macmillan.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analyses of data from retrospective studies of the disease. Journal of the National Cancer Institute, 22, 719-748.

Martinez, M. E. (1991). A comparison of multiple-choice and constructed figural response items. Journal of Educational Measurement, 28, 131-145.

McGuire, C. H., Solomon, L. M., & Bashook, P. G. 976). Construction and use of written simulations. San Antonio, TX: The Psychological Corporation.

Melnick, D. E. (1990). Computer-based clinical simulation. Evaluation and the Health Professions, 13, 104-120.

Merz, W. R., & Grossen, M. (1979). An empirical investigation of six methods for examining test item bias (Final Rep. No. NIE-G-78-0067). Sacramento, CA: California State University.

Murray, J. P. (1990). Better testing for better learning. College Teaching, 38, 148-152.

Raju, N. S., Bode, R. K., & Larsen, S. L. (1989). An empirical assessment of the Mantel-Haenszel statistic for studying differential item performance. Applied Measurement in Education, 2, 1-13.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). A Monte Carlo comparison of seven biased item detection techniques. Journal of Educational Measurement, 17, 1-10.

Scheuneman, J. D. (1979). A method of assessing bias in test items. Journal of Educational Measurement, 16, 143-152.

Scheuneman, J. D., & Bleistein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. Applied Measurement in Education, 2, 255-275.

Shepard, L., 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.

Singer, D. (1985). Latent image processing can bolster the value of quizzes. Journal of College Science Teaching, 15, 114-116.

Stahl, J. A., Lunz, M. E., & Snyder, J. R. (1990, April). Validity of statistical detection of bias in test items. Paper presented at the meeting of the American Educational Research Association, Boston, MA.

Stricker, L. J. (1982). Identifying test items that perform differently in population subgroups: A partial correlation index. Applied Psychological Measurement, 6, 261-273.

Subkoviak, M. J., Mack, J. S., Ironson, G. H., & Craig, R. D. (1984). Empirical comparison of selected item bias detection procedures with bias manipulation. Journal of Educational Measurement, 21, 49-58.

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

Swanson, D. B., & Stillman, P. L. (1990). Use of standardized patients for teaching and assessing clinical skills. Evaluation and the Health Professions, 13, 79-103.

Tittle, C. K. (1982). Use of judgmental methods in item bias studies. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 31-63). Baltimore, MD: The Johns Hopkins University Press.

Veale, J. R. (1977). A note on the use of chi-square with "correct/incorrect" data to detect culturally biased items (Technical Report No. 4). Berkeley, CA: Statistical Research: Education and Behavioral Sciences.

van der Vleuten, C. P. M., & Swanson, D. B. (1990). Assessment of clinical skills with standardized patients: State of the art. Teaching and Learning in Medicine, 2, 58-76.

Wainer, H., et al. (Eds.). (1990). Computerized adaptive testing: A primer. Hillsdale, NJ: Erlbaum.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. Journal of Educational Measurement, 26, 55-66.