CONFUCIUS' TEACHING ON ODES, RITES AND MUSIC

BY

CHANG, CHING-CHUAN*

The teaching of Confucius on character was compiled mainly in the three books entitled Odes, Rites and Music. Odes was written to enrich man's language, Rites was written to regulate man's behavior, and Music was written to beautify man's mind. In content, they are sequential trinity; in spirit, they are an inter-connected whole.

First, the purpose of the paper is to explain the meaning of Odes, Rites and Music, Then, secondly, the effectiveness of character cultivation, and make Confucius as the word modle on learning and character.

* Tearching Assistant, Department of Chinese, National Changhua University of Education

Methods of Detecting Test Item Bias

by

Hui-Fen Lin*

ABSTRACT

The purpose of this study was to investigate, through literature review, the differences between various methods of detecting test item bias. They are :the transformed item difficulty method, the item characteristic curves method, the Chi-square technique, the analysis of variance, the point biserial correlation approach and the judgemental approach. Reviewing the comparative studies about debiasing methods, several findings were found. They are: (1) although the judgemental review is criticized by many researchers, most test publishers still integrate it with statistical methods for identifying biased items; (2) among the various psychometric item bias detecting methods, the item characteristic curves approach is the most sensitive to bias in item discrimination, but it is restricted by large sample size, cost and complicated computer analysis; (3)for practical consideration, the Chi-square method and the transformed difficulty approach are recommended for most users. The future research is also recommended in the study.

* Associate Professor, Department of Special, National Changhua University of Education.

Methods of Detecting Test Item Bias

Introduction

Testing is a foundation upon which eqality is built (Wooten, 1982). The use of tests, therfore, has been widespread in schools, clinics, industry and government. For example, tests are used for placement, as when a student is placed in a remedial reading class. They are also used for selection as when a company sėlects new employees from a group of job application. Through the use of testing, an adequate educational placement for a student could be made, and a potential employee is supposed to be selected objectively.

However, the issues of bias in testing and related issues such as bias in selection and the fair use of tests have become major concerns for the educational and professional measurement over the past decade. The issues of bias in achievement tests will be mainly deal with in this paper. Through the review of literature, the definitions of bias in testing, the methods used for proposed as well as the approaches utilized by test publishers to detect bias will be discussed and presented.

Definitions of Bias

A number of definitions of bias in test have been defined by the researchers. Scheuneman (1979) defined an item as unbiased if, for persons with the same ability in the area being measured, the probability of a correct response is the same regardless of ethnic group memberships. Osterlind (1983) and Ironson (1982) had similar statements. They considered an item biased if equally able members of different groups have unequal chances of success on the item. Angoff & Ford (1973), and Plake (1979) suggested an item is biased if, compared to other items on the test, it is rėlativėly more difficult for one group than for another.

In addition, Cleary (1968) asserted that a test is biased if scores for subgroups are consistently predicted too high or too low. Green (1975) defined a biased test as one that produces results systematically unfair to some group. Unfairness, Green (1975) maintained, is not found in the nature of the test but in the way in wich scores are used.

Furthermore, Cole (1978) suggested that an item is thought to be biased if it dose not measure the same things in different groups. In a recent study along this line, Shepard (1980) claimed that bias would be suspected if children in one group consistently receive lower scores than would be expected from observed classroom performance.

There are still other definitions of item "bias". Most of them are different from those offered here, depending on the particular methodology used to identify the bias. However, common to each definition is that "bias" in test is referred to not truly reflect a examinee's capability.

Psychometric Item-Bias Detection Methods

Logical and empirical analyses are two models utilized for detecting bias. In logical analyses, the editorial reviews and reviews by minority representative panels are conducted during the test development process. They are also referred to as judgmental methods. Many researchers questioned their subjective validity (Osterlind, 1983; Jense, 1979). After conductig several studies, Reynold (1982) also found dismal evidence for the effectiveness of such reviews in discovering biased test items. However, many test publishers still implement such counsels as their first steps in the test development process. The reasons, as suggested by Shepard (1982), are their sensitivity to ambiguous items and offensive test formats as well as any pervasive reduction in performance caused by perceived unfairness in the questions.

In empirical analyses, numerous statistical techniques have been proposed for operationalizing item bias. They include Transformed Item Difficulty (Delta Approach), Item Characteristic Curves, Chi-Square Approach, Analysis of Variance, and Point Biserial Correlation Approach.

Transformed Item Difficulty

The Transformed Item Difficulty, proposed by Angoff, considers bias to be a characteristic inherent in all test items. It defines an item as biased if the item is comparatively more difficult for one group to answer than it is for the other (Osterlind, 1983).

With TID, the p-value (the proportion of subjects answering items correctly) is obtained first for two different groups on a set of items. Each p-value is then transformed into delta scale, with a mean of 13 and standard deviation of 4. The pairs of deltas are plotted on a bivariate graph. Ideally, the plot would be a perfect 45 degree of line indicating the absence of bias. The items filling at some distance from the 45 degree of line are regared as the result of item-by-group interaction. On the other hand, these are items that are especially more difficult for one group than the other, and bias is suggested.

In identifying the truly biased item, several different acceptable tolerance limits have been proposed. Strassberg-Rossenberg and Donlon (1975) suggested that any item varying from 45 degree of line by greater than 1.5 times the standardized deviation, the residual is defined as biased item (cited in Merz & Runder, 1978). Groome and Groome (1979) identified a biased item as any item which lies outside of the 95% confidence interval, while Runder (1978) suggested a fixed item-regression line distance of .75 z-score units.

Item Characteristic Curves

The item Characteristic Curves is a graphic representation describing the relationship between ability and probability of answering the questions correctly (Allen & Yen, 1979). For each item, the ICC is estimated by a graph with total test scores on the horizontal axis and the proportion of examination passing the item on the vertical axis. Derived from Item Response Theory, the item characteristic curves approach is based on the assumptions of : (1) dimensionality of the latent source that means an examinee's performance on a test can be attributed to a single trail or ability, and (2) local independence of items that means an examinee's performance on any item is unaffected by the performance on any other item in the test (Osterlind,1983). Ironson (1982) stated that the ICCs are the best theoretical approach to the study of item bias just because of their property of parameter invariance.

The curves are described by a cumulative logistic or normal ogive function. The difficult parameter, discrimination parameter, and guessing parameter are three parameters used to described such a curve. The Rasch model, also called a one-parameter logistic parameter, uses the difficulty parameter to describe an item. Small sample size can be used in this approach. Because it focuses on one parameter, many researchers have implemented it to study either a fit analysis or a difficulty shift analysis (Ironson, 1982).

The three-parameter model is another type of ICC. Compared with the Rasch model, the three-parameter approach is more realistic and fits the data better because it includes discrimination value and guessing parameter (Ironson, 1982). Computer programs such as Logist and Ancilles are available for estimating parameter. Osterlind (1982) stated that the ICC-3 approach is the most favorable research technique for conducting bias item detecting work because it comes close to describing psychometrically multiple-choice tests when they are presently constructed and used. The disadvantages of the ICC-3 approach, suggested by Ironson (1982), are complex and expensive analysis required as well as a large sample-size.

Chi-Square Technique

The Chi-Square approach to the identification of test item bias is to examine the probability of test-takers from different groups with the same ability level to correctly answer an item (Osterlind, 1983). The essential strategy of this technique is to remove biased item identification from the dependency on the groups by item interaction. The major assumption of chi-square approach is that total test scores can be used as an estimate of validity. In performing chi-square procedure, the total score scale is divided into three to five validity intervals or score levels. The observed andexpected frequencies for each interval then are used to calculate a chi-square.

Based on the chi-square technique, a variety of statistical methods have been proposed. Scheuneman's approach (1975, 1979) focuses on correct responses to a given item. Camilli (1979) has introduced a modification that makes use of both the correct and incorrect responses (cited on Berk, 1982).

Small sample size, intuitively easy to understand and good for rapid screening are the advantages of the chi-square approach (Marascuilo & Slaugh ter, 1981; Scheuneman, 1977). The disadvantages, suggested by Ironson (1982) includ (1) the total test score is an imperfect measure of ability, and (2) procedures are sensitive to the distribution of total test scores and to the differential sample size of subgroups.

Analysis of Variance

The usual analysis of variance analyzes variance as a number of additive components that together equal the total score variance. That is, the total variance can be attributed to the difficulties of items, the difficulties of groups, and the interaction of groups with items. However, the analysis of variance in testing item bias does not allow the researchers to partition out the variance. Instead, it specifies only the interaction of groups with items. In other words, an item is biased if a significant interaction between groups and items is revealed.

Procedually, the p values for one group are plotted along the abscissa, and those for the other groups are plotted on the ordinate. A 45-degree line is then drawn. Items with identical p values for both groups will fall exactly on the line; items easier or more difficult for one group than for the other will fall away from the line.

When a significant groups-by-item interaction is exposed, a post hoc procedure needs to be conducted in order to identify specific items are biased. Plake and Hoover (1979) recommended a Bonfereoni-type comparison, while Osterlind (1982) proposed the transformed item difficulties as the most logical post hoc method. Neither Tukey's nor Scheffe's procedure is recommended, because each is too conservative. In addition, Tukey's test requires equal sample sizes, which are not easily achieved in real situation (Osterlind,1983).

Point Biserial Correlation Approach

This approach examines the correlation between item performance and total score. First, item difficulty levels are obtained using one group as a reference against which the other groups are calculated. Items are arranged in order of difficulty for the reference group from most difficult to least difficult. Then, point biserial correlations are computed for each group on each item. Correlations are compared to identify items which for a particular group does not contribute to total scores, that is, items with a low item total score correlation for a specific group are identified as biased (Merz & Runder, 1978).

Review of Comparative Studies About "Debiasing" Methods

The various techniques of detecting item bias differ in their conceptualization and computation as well as in their interaction. In order to determine which of the proposed methods of identifying item bias is most useful in accomplishing that end, a number of studies have been conducted. Recently, several researchers have also sought to evaluate the methods they propose in terms of item pools, samples required, and item- or cost-effectiveness.

Comparative Study Using Simulated Data

Several researchers have attempted to compare the proposed procedures for detecting item bias by applying the simulated data. Such data are generated using a Monte Carlo procedure to establish a priori in both the amount and the nature of the bias in each test. Thus, methods may be evaluated not only for their relative similarity, but also for the extent of "spurious" bias they may identify.

In the Runder, Geston, and Keight (1980) study, seven techniques were applied:

transformed item difficulty-Major Axis (TID-MA); transformed item difficulty-45 degrees of line with 1 z-score unite (TID-45); item characteristic curves with three parameters(ICC-3); item charactèristic curves with one parameter-Fix statistics (ICC-Fix, also called Rasch model), item characteristic curves with one parameter- difference in item Easiness (ICC-1E); Chi-square with five intervals (CHI-5) and Chi-square with multiple interval (CHI-M). Data were generated for 112 different combinations of test conditions : 7 test lengths, 4 different amounts of bias in discrimination and 4 different amounts of bias in difficulty. Birnbaum's (1968) three-parameter logistic model was chosen to relate item and examine characteristics to item responses. Two groups of 1200 examinees were drawn with one standard deviation difference in the mental level of performance.

In studying the correlations of the detected and generated amount of bias, the ICC-2 and ICC-3 techniques were found to be the most accurate with correlation of. .80 and .73 respectively. The transformed item difficulties were found to be sensitive to bias in item difficulty, but relatively insensitive to bias in item discrimination. The item characteristic curves theory approaches with one parameter consistently correlated poorly with the amount of generated bias. Runder and his colleagues recommended the item characteristic curve theory with three parameters or the chi-square techniques with five intervals for detecting biased test items, regardless of test length, amount of bias, or nature of bias. They further suggest that when only item p-values are available, the transformed item difficulties-45 degrees of line can be used.

Groome and Groome (1974) conducted a study to compare two transformed item difficulties approaches-45 degree of line with 1.5 z-score unites and 95% confidence interval with three-parameter item characteristic curves approach. Four 6-item tests, based on item simulated data generated using that ICC methodology, were developed and each group contained 1,000 examinees.

The results found that TID approach adequately identified the known biased item when either of the two criteria was implemented. However, the correlation analysis between known-to-be-biased and identified-as-biased items exposed a high degree of congruence between the two criteria. Groome and Groome interprated this phenomena as a result of the lack of a good criterion for identifying biased items. Consequently, they suggested that the transformed item difficulties approaches can not serve as a reliable criterion when applied to a standardize test in which bias has not been previously controlled.

Six procedures were compared in Merz's and Grossen's study (1974): transfomed item difficulties, factor analysis, point biserial correlation, chi-square technique, and one and three-parameter item characteristic curves procedures. The three-paramter model was used to generate simuated data, but only the difficulty parameter was manipulated in this study.

The findings indicated that all of the approaches, except the point biaserial, correctly identified the number of biased items. The transformed item difficulty approach appeared to function best, which correlated .95 and .97 with generated bias across all conditions, while the point biserial method consistently under-identified the bias items.

Comparative Studies Using Empirical Data

Rather than using simulated data, several researchers have utilized "real" data to compare the proposed techniques for detection item bias.

Subkoviak, Mack, Ironson and Craig (1984) had compared three item-bias detection procedures: the transformed item difficulty approach, the chi-square approaches (including Scheueman's and Camilli's model), and the three parameter item characteristic curve. A 50-item vocabulary test was constructed, in which 10 items intentionally favored the black students.

The results of this study indicated that the ICC-3 procedure was most effective in detection the a priori bias, with correlations of .872 and .875, respectively, for signed and unsigned measures. The chi-square approaches had high correlation with ICC-3 in the analysis of inter-correlation.

Subkoviak et al. (1984) suggested ICC-3 to be a preferred method in detection the a priori bias, provided the researcher had adequate numbers of items (40 or more) subjects, as well as computer facilities. When sample size or other consideration precluded the use of three-parameter ICC or chi-square methods, the transformed item difficulty approach could be used as a practical alternative.

Ironson and Subkoviak (1979) applied the transformed item difficulty, chi-square, point biserial and three-parameter item characteristic curves methods to conduct their study. Test data from the 1972 Longitudinal Study was used. A percentage of agreement index between pairs of methods was computed, using twenty-four items identified to be the most biased by each method. The chi-square and ICC approaches exhibited the highest agreement (54.2%) in their choice of the most biased item, compared with 37.5 percent for chi-square with transformed item difficulty, and 33.3 percent for transformed item difficulty with the ICC approach. No significant correlation was found between the point biserial approach and any of the other techniques.

Ironson and Subkoviak (1979) stated that point biserial approach may be inadequate for use in bias studies. Due to the restriction by the requirement of a large sample size and computational complexity, the item characteristic curves approach, they suggested, may be feasible only for the test publishers. For most users, the chi-square as well as the transformed item difficulty techniques may be more practical.

Using two district-development tests, Monaco (1985) compared the chi-square, transformed item difficulty, and Lin-Harnish three-parameter item response approaches for detection of sex and ethnic bias.

In this study, the chi-square and Lin-Harnish approaches were found to have the greatest degree of agreement in detection the most and least biased items, while the least amount of agreement was found to exist between the chi-square and the transformed item difficulty method. The transformed item difficulty approach could identify the greatest number of index as being biased, while the chi-square approach detected the least number of indices as being biased.

From her study, Monaco (1985) suggested that chi-square and Lin-Harnish approaches detected the biased items in a similar manner. When the sample sizes are relatively small, the Lin-Harnish approach can be used in lieu of other restricted item response approaches.

Using data from the Standford Achievement Test, Runder and Convey (1978) compared four approaches, including transformed item difficulties (45 degrees of line), Scheuneman's chi-square technique, item characteristic curves, and factor scores. The highest correlation existed between ICCs and chi-square (r=.67) in detection biased items. The factor-score and chi-square approaches showed the least degree of similarity with the other approaches. They suggested that the ICC theory and transformed item difficulty approaches appeared to be the most attractive among the methods studies.

Shepard, Camilli and Averill (1980) compared four approaches, including chi-square analysis (both Scheuneman's and Camilli's procedures), Angoff's delta-plot approach, point biserial discrimination, and item characteristic curves (both ICC-3 and ICC-1). Test data was from the Lorge-Thorndike Test, which is a group-administered mental ability test for three diverse cultural groups (Black, white and Chicanos) (cited in Berk, 1983).

A correlation of .99 was found between Angoff's procedure and the one parameter ICC approach in the identification of extreme items. When using internal and external criteria of ability to examine how well each procedure correlated with itself, the ICC-1 was found to have the highest correlation (.99), while the chi-square as well as the ICC-3 had moderate correlation.

Shepard et al. (1980) suggested that the three-parameter methods are the most theoretically sound. Comparing the transformed item difficulty technique with the item characteristic curves parameter procedure for bias detection, they recommended the use of TID because ICCs involved complicated computation. They also suggested that when samples are not large enough, Camilli's chi-square procedure can be used and bias indices should be used as the sole basis for accepting or rejecting an item.

Strategies Used by Test Publishers to Eliminate Bias Items

Since the issues of bias in testing are major concerns for educators, test publishers also take steps to "debias" their material. Green (1982) stated that CTB/McGraw Hill pay its primary concern at content validity. The draft of tests are made by item writers who follow the guidelines provided by CTB. The CTB editors and outside reviewers them follow the guidelines to review the test. After that, the ethnic groups separately analyze the items that appear to be undesirable for one or more groups. In addition to the judge mental review, CTB also integrates the point biserial technique as item bias analysis.

The electic approach, representing both subjective judgements and statistical analysis, is adopted by Riverside Publishing Company (Coffman, 1982). Five stages are involved in this approach: (1), applying prefessional judgements at the item-writing stage; (2), inviting representative panels of test users to systematically review the test contents; (3), conducting statistical analysis based on a variety of bias models; (4), making comparisons among the results of stages 2 and 3, and (5) , designing special follow-up studies to seek answers raised by the comparisons at stage 4.

Although various "debiasing" models have been used to conduct statistical analysis, no one method, Coffman stated (1982), proves to be superior, and the results are not comparable. The goals for Riverside Company in debiasing now include learning about the nature of the tests themselves, about the performance characteristics of subgroups, and about the models being used for assessing bias.

For the American College Testing (ACT) Company, a review of each content area is consistently conducted to monitor possible changes. The item writers are chosen from a variety of backgrounds and are periodically provided with training services. A pre-test is administered to the representative samples and analyzed for sex, minority groups, or regional differences in performance.

In addition to the routine procedure, the use of statistical techniques such as chi-square as well as ANOVA is also periodically employed. The future researches on bias that ACT plans to focus on include studying the internal consistency of different methods for detection bias, and investgating the influence of various test construction with test situations on the performance of examinees (Handrick & Loyd, 1982).

Like the CTB-McGraw Company, a combination of judgemental and statistical approaches are employed by Science Research Associates. In developing the Achievement Services, for instance, the effort to minimize bias began with providing detailed specifications for item writers, designed to make them sensitive to all the possibilities for editorial bias. Use of unbiased language and balance in ethnic and sex work roles were also emphasized.

During the item development process, items were reviewed by content and testing specialists and by knowledgeable representives of minority groups. Acceptable items were then pre-tested by both majority and minority groups. The recent development plans for SRA, as Raju (1982) stated, focus on the study of latent trait models, especially concerning the three-parameter model on the development of multi-level achievement series with a large sample size. Because to increase the test reviewers' sensitivity is one of the salient features of the Educational Testing Service's current policy the reviewers, therefore, are selected from a pool of volunteers, coming from different backgrounds.

A training program is then implemented during which the process is explained, criteria are discussed and examples of acceptable and unacceptable practices are worked. In evaluating the Test of English as a Foreign Language (TOEFT) , for instance, a unique procedure of sensitivity was initiated to meet the need for cultural sensitivity on a broader range. A group of eight reviewers were assembled, each of whom had had a minimun of 10 years of living and working abroad. The training sessions then were conducted at the beginning of the review process.

The delta-plot method is the procedure most commomly used by ETS for detection bias. Other approaches such as the latent trait theory analysis, and distractor analysis have also been employed. For research on bias, identifying why items are biased and investigating the effectiveness of various methods in detectng biased items now are two major goals for ETS (Carlton & Marco, 1982).

Summaries

Allen and Yen (1979) said:"... It is difficult to imagine any one who has not taken hundreds of tests."(p.1). Therefore, whether a test can truly reflect what it is purported to measure receives wide attention from society. To say a test is invalid can be interpreted from three aspects: inadequate contents, different predictive validity, and measurement of different construct. These three types of invalidity can be translated into three types of bias: content bias, predictive bias, and constructive bias.

Since the "bias" in tests can be interpreted from various aspects, it does not have a unique definition. For instance, Cleary (1968) suggested that if the scores for a subgroup are consistently predicted too high or too low, the test is biased. Defining from item level, Shepard, Camilli and Averill (1980) stated that an item is biased if two individuals with equal ability but from different group do not have the same probability of success on the item.

Based on different rationales, a variety of methods for detection bias have been proposed by the researchers. In order to determine which proposed methods can successfully accomplished that end, a number of comparative studies have been conducted. The data being used are drawn either from the simulated or empirical data. Although the data are not the same, some similar results are found:

1. The item characteristic curves approach and chi-square technique have the highest correlation in detection the biased item (Subkoviak, Mack, & Craig, 1984; Monaco, 1986; Shepard, Camilli & Averill, 1980; Ironson & Subkoviak, 1979; Runder & Convey, 1978). It may be, as Osterlind (1983) said, because both of these methods are derived from the item response theory.
2. The item characteristic curves approach is most sensitive to bias in item discrimination (Runder, Geston, & Knight, 1980; Merz and Groosen, 1979; Runder & Convey, 1978; Shepard, Calmill & Averill, 1980).
3. The item characteristic curves method is restricted by the large sample size, cost and complicated computer analysis. For practical consideration, the chi-square method and transformed item difficulties are recommended for most users (Runder, Geston & Keight, 1980; Shepard, Camilli, & Averill, 1980 ; Ironson & Subkoviak ,1979).
4. The point biserial procedure does not function well in detection biased items (Merz & Grossen, 1979; Shepard, Camilli & Averill, 1980).

Although the judgemental review procedure is criticized by many researchers due to its poor validity, most test publishers integrate it with statistical methods for identifying biased items. This controversial fact may be because: (1) subjective reviews of test items are sensitive to any ambiguous items and to any pervasive reduction in performance caused by perceived unfairness in questions; (2) there is no fool proof statistical bias detection method, and (3) item bias techniques themselves require validation (Shepard, 19883).

In the statistical analysis process, the debiasing methods vary markedly from publisher. They range from the transformed item difficulties approach, analysis of variance method, or point biserial procedure to item characteristic curves methods.

Conclusion and Suggestions

From the review of methods for detection test bias, it appears that neither the judgemental reviews nor statistical procedures can function well enough in item bias detection. For the judgemental review, it lacks of objective validity, but it is useful for identifying aspects of the test format that may be offensive to the minority group. Due to lack of external criterion, statistical techniques can not detect pervasive bias. Meanwhile, each of them has its own specific limitations. Therefore, in order to eliminate biased items, it seems necessary (1) to include the judgemental review and item bias studies as integral phases of test development and (2) to integrate bias analysis into the test construction process rather than to "debias" the test after it has been used for some time or after charges of bias have been directed at the test.

Reviewing the current literature about statistical item bias methods, most comparative studies focus on the comparisons of item characteristic curves approach, transformed item difficulties methods, chi-square procedure and/or analysis of variance. In addition to these methods, other procedure such as the loglinear approach, distractor analysis procedure, and partical correlation index have been suggested recently as possible techniques for identifying biased items. For future comparative study, it seems that these methods could also be included for evaluation.

Reference

Allen, M. J., & Yen, W. H. (1979). Introduction to measurement theory. California: Brooks/Cole Publishing Co.

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10, 95-105.

Berk, R. A. (1982). Handbook of methods for detection test bias. Baltimore , Ma: The Johns Hopkins University Press.

Camilloi, G. A. (1979). A critique of the chi-square method for assessing item bias. Unpublished paper. Labortory of Educational Research, Univer sity of Colorado.

Carlton, S. T., & Marco, G. L. (1982). Methods used by test publishers to "debias" standardized tests. In R. A. Berk (ED.) Handbook of methods for detection test bias . Baltimore, Ma: The Johns Hopkins University Press.

Cleary, T. A. (1968). Test bias: Prediction of grades of negro and white students in integrated college. Journal of Educational Measurement, 5, 115-124.

Coffman, W. E. (1982). Methods used by test publishers to "debias" standardized tests. In R. A. Berk (ED.) Handbook of methods for detection test bias . Baltimore, Ma.: The Johns Hopkins University Press.

Cole, N. S. (1979). Approaches to examing test bias in achievement test items. Califonia, Los Angles. (ERIC Document Reproduction Service No. 200 163).

Flaughter, R. C. (1978). The many definitions of test bias. American Psychologist, 33, 671-679.

Green, D. R. (1975). What does it mean to say a test is bias? Educational and Urban Society, 8 , 33-52.

Green, D. R. (1982). Methods used by test publishers to "debias" standardized tests. In R. A. Berk (ED.) Handbook of methods for detection bias. Baltimore, Ma: The Johns Hopkins University Press.

Groome, M. L., Groome, W. R. (1979). Item bias identification: A comparison of two approaches . Maryland: Washington, D. C. (ERIC Document Reproduction Service No. 174 685.)

Handrick, F. A., & Loyd, B. H. (1982). Use of Chi-square and latent trait approaches for detection item bias. In R. A. Berk(ED.) Handbook of methods for detection test bias. Baltimore, Ma.: The Johns Hopkins University Press.

Ironson, G. H., & Subkoviak, M. J. (1979). A comparison of several methods of assessing item bias. Journal of Educational Measurement, 16, 209-225.

Jense, A. R. (1979). An examination of cultural bias in the wonderlic personal test. Intelligence, 1, 51-64.

Marascuilo, L. A., & Slaughter, R. E. (1981). Statistical procedures for identifying possible sources of item bias on chi-square statistics. Journal of Educational Measurement, 18, 229-248.

Merz, W. R., & Groosen, N. (1979). An empirical investigation of six methods for examing test item bias. California: San Francisco. (ERIC Document Reproduction Service No. 178 566).

Merz, W. R., & Runder, L. M. (1978). Bias in testing: a presentation of selected methods. Tornoto: Ontario. (ERIC Document Reproduction Service No. 164 610).

Monaco, L. (1985). A comparison of three methods of detecting test item bias. Unpublished doctoral dissertation. North Texas State University, Denton, Texas.

Osterlind, S. J. (1983). Test item bias. California: Sage Publication Ltd.

Plake, B. S., & Hoover, H. D. (1979). An analysis method of identification biased test items. Journal of Educational Measurement, 48, 153-154.

Reynold, C. R. (1982). The problems of bias in psychological assessment. In R. A. Berk (ED.) Handbook of methods for detection test bias. Baltimore, Ma.: The Johns Hopkins University Press.

Runder, L. M., & Geston, P. R., & Keight, D. L. (1980). A Monte Carlo comparison of seven bias item detection techniques. Journal of Educational Measurement , 7, 1-10.

Runder, L. M., & Convey, J. J. (1978). An evaluation of selected approaches for biased item identification . Toronto, Canada. (ERIC Document Reproduction Service No. 157 942).

Scheuneman, J. D. (1975). A new method for assessing bias in test items. Maryland:

Washington, D. C. (ERIC Document Reproduction Service No. 106 359).

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

Shepard, L. A.; Camilli, G., & Averill, M. (1980). Comparison of six procedures for detection test item using both internal and external criteria. In R. A. Berk (ED.). Handbook of methods for detection test bias. Baltimore, Ma.: The Johns Hopkins University Press.

Shepard, L. A. (1982). Definition of bias. In R. A. Berk (ED.). Handbook of methods for detection test bias. Baltimore, Ma.: The Johns Hopkins University Press.

Strassberg-Rossenberg, B. C., & Donlon, T. F. (1975). Content influences on sex differences performance and aptitude. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, D. C.

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.

Wooten, K. L. (1982). The foundation of equality. In S. B. Anderson, & L. V. Coburn (ED.) Academic testing and the consumer.

試題偏差檢驗方法之比較研究

林惠芬*

中文摘要

本研究係透過文獻探討方式比較各種檢驗試題偏差之方法。所探討的方法有：轉換試題難度法、試題特徵曲線法、卡方法、變異數分析法、點二系列法、以及專家判斷法。從有關之比較研究的結果得知：(1)雖然專家判斷法受到許多學者的質疑，但大部份的測驗出版商仍採用此種方法來檢驗試題是否有偏差；(2)在各種以心理計量方式來探討試題偏差的方法中，以試題特徵曲線法在檢驗試題是否有偏差最爲敏感，但是此種方法須使用相當大的樣本，以及複雜的電腦分析程式；(3)從實用觀點考量，則以卡方法以及轉換試題難度法較爲適用。依據研究之結果，本研究並提出未來探討的方向。

*國立彰化師範大學特教系 副教授