ED 353 308 TM 019 372

AUTHOR	Hambleton, Ronald K.; Jones, Russell W.
TITLE	Comparison of Empirical and Judgmental Methods for Detecting Differential Item Functioning.
PUB DATE	[92]
NOTE	23p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 21-23, 1992).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150) -- Tests/Evaluation Instruments (160)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Achievement Tests; *American Indians; *Anglo Americans; Comparative Analysis; *Item Bias, Judgment Analysis Technique; Multiple Choice Tests; Occupational Tests; *Racial Differences; Reliability; Research Methodology; Research Problems; State Programs; Statistical Analysis; Testing Programs; *Test Results
IDENTIFIERS	Empirical Research; Native Americans

ABSTRACT

The purpose of this study was to improve both statistical and judgmental methods for detecting potentially biased test items in an attempt to examine the agreement between the results obtained with these methods. If greater agreement between methods can be achieved, test items can be more effectively screened using judgmental methods prior to field testing or actual test admir:strations. Several methodological shortcomings of current empirical and judgmental methods were addressed. The test data came from samples of 2,000 Native Americans and 2,000 Anglo-Americans who took a 4-choice 150-item Statewide Proficiency Test. Fifteen Native American educators provided item bias reviews. To reduce computer time and facilitate the analyses, 75 items on the test were analyzed. The Mantel-Haenszel procedure was applied to the item responses for the two subject groups. The results suggest that s somewhat higher level of agreement between methods was obtained than has been observed in other studies. The use of cross-validation in empirically identifying potentially biased items was one reason for the higher level of agreement. However, the judgmental method implemented in this study appeared to have several shortcomings. Practical implications of the findings are presented. Two tables and one figure are included. (Author/RLC)

Comparison of Empirical and Judgmental Methods for Detecting Differential Item Functioning

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

Abstract

The purpose of this study was to improve both statistical and judgmental methods for detecting potentially biased items in a test in an attempt to examine the agreement between the results obtained with these methods. If greater agreement between methods can be achieved, test items can be more effectively screened using judgmental methods prior to field testing or actual test administrations.

Steps were taken to address several methodological shortcomings of current empirical and judgmental methods. The test data came from samples of 2,000 Native American and 2,000 Anglo-American students who took a 150-item Statewide Proficiency Test. Fifteen Native American educators provided item bias reviews.

The results suggest that a somewhat higher level of agreement between methods was obtained than has been observed in other studies. The use of cross-validation in empirically identifying potentially biased items was one reason for the higher level of agreement. However, the judgmental method implemented in this study appeared to have several shortcomings. Practical implications of the findings are presented.

LR231

Comparison of Empirical and Judgmental Methods for Detecting Differential Item Functioning[1,2]

Ronald K. Hambleton and Russell W. Jones[3]
University of Massachusetts at Amherst

Paper-and-pencil tests are widely used as tools for selection, promotion, certification and licensure decisions throughout education, business, the armed services, and industry. As test use for important decisions has increased, the issue of item bias has achieved considerable significance. Test developers must now demonstrate that their tests are free of item bias. To this end, various judgmental and empirical methods for detecting potentially biased items have been proposed (see, for example, Berk, 1982; Hills, 1989; Scheuneman & Bleistein, 1989). These "DIF" studies, as they are commonly called, are designed to detect differential item functioning (DIF) between reference and focal groups.

Typically, judgmental and empirical methods for detecting differentially functioning items have shown little agreement (Plake, 1980; Engelhard, Hansche, & Rutledge, 1990). A partial explanation for this low agreement is that the judgmental review forms are sometimes focused on cultural and sexual stereotyping in items rather than on factors which may lead to differential performance between subgroups of interest (Scheuneman, 1982). As a result, many undesirable items are identified in the item bias review process, such as those which may show members of minority groups doing

[1] Laboratory of Psychometric and Evaluative Research Report No. 231. Amherst, MA: University of Massachusetts at Amherst, School of Education.

[2] Paper presented at the meeting of NCME, San Francisco, 1992.

[3] The authors wish to thank the 15 Native American educators who completed the item review task, and John Martois for his assistance in locating reviewers.

unskilled work or having problems of one kind or another. However, the items identified, although undesirable, are unlikely to function differentially in actual practice.

Another reason for the low agreement is that the statistical methods themselves are not highly reliable (see, for example, Hambleton & Rogers, 1989; Hoover & Kolen, 1984). Hambleton and Rogers showed that both IRT methods and the Mantel-Haenszel method led to somewhat unstable results even with fairly large samples (N=1,000). Agreement levels in the classification of items as "DIF" or "not DIF" ranged between 72% and 80%. They recommended that a cross-validation sample be used whenever possible, and that an item be considered potentially biased, or differentially functioning, only if flagged in both samples. In the same study, these authors also drew attention to the importance of the portion of the ability scale over which DIF is measured, and they described a method for choosing a cut-off score for interpreting DIF statistics.

The purpose of this study was to refine, in relation to common practices, both statistical and judgmental procedures for detecting potentially biased items in an attempt to improve the agreement between the results obtained with these methods. This seemed a worthy goal because if greater agreement between methods can be achieved, test items can be more effectively screened using judgmental methods prior to field testing or actual test administrations. In fact, in some small scale test development studies, item bias reviews may be as much bias identification work as can be accomplished. In other studies, empirical work can be done but the results are unstable because of small sample sizes, especially for the focal group. Also, the fewer items that are defective during field tests or test administrations, the more credible the agencies producing the tests are judged to be. Clearly, therefore, research that might lead to improvements in item bias review forms seemed desirable.

The statistical methods were refined by (1) focusing only on items which were differentially functioning in both the original sample and in a cross-validation sample, (2) carefully choosing the interval over which DIF was measured and the cut-off score for interpreting the DIF statistics, and (3) using more than one DIF statistic in the empirical analysis. The judgmental methods were refined by (1) carefully distinguishing between stereotyping of groups and factors which could differentially impact on test performance, and (2) using the findings of an earlier study by the authors to refine the item bias review form.

Method

Collection of Judgmental Item Bias Review Data

Some form of item bias review has probably been applied to test items since the inception of testing. The development of technical principles for educational and psychological tests in the form of the AERA, APA, and NCME Test Standards (since 1954), has gradually resulted in a concomitant application of more sophisticated judgmental review of items, item piloting and final item selection. Relatively recent developments, largely as a result of an increasing public, political, and judicial awareness of the impact of testing and the importance of accurate measurement, have seen increasing attention being paid to whom the judges are, the focus of the judgments, and the systematization (formalizing) of judgments (Tittle, 1982).

For this research, an item bias review form was developed which was based on the following five principles:

1. Avoidance of stereotyping, defined as the consistent representation of a given group in a particular light, which may or may not be offensive to members of that group.

2. Fair representation of women and minorities.

3. Equal familiarity or experience of subgroups with the content and language of items.

4. The opportunity to learn item content. That is, the match of overlap of items with the instructional process (Tittle, 1982).

5. Requirement that all subgroups have equal probability to respond correctly for the correct reason(s). Item formats, test structures, stimulus material (directions, graphics, etc.), response alternatives, and clues which favor or hinder a particular subgroup are avoided.

Issues of stereotyping and fair representation (i.e., 1 and 2 above) may be considered distinct from differential item functioning (i.e., 3 to 5 above). Stereotyping and fair representation of a given group in a particular light, although undesirable, would not, except in extreme cases, lead to differential performance between designated groups of interest. Conversely, DIF, by definition, does lead to differential performance. DIF may be defined as the presence of some characteristic within an item which results in differential performance for two individuals of the same ability but from different subgroups.

Stereotyping and inadequate or unfavorable representation of subgroups are undesirable properties of test items. Tests should be free of material which may be offensive, demeaning, or emotionally charged to some groups. An example of emotionally charged material would be an item dealing with the high suicide rate among Native Americans. An example of offensive material would be an item which implied the inferiority of a certain group.

Potential DIF comes in many forms. An item may be functioning differentially if it contains content or language that is differentially familiar to subgroups of examinees, or if the item structure or format is differentially familiar to subgroups of examinees, or if the item structure or format is differentially difficult for subgroups of examinees. An item may be considered to exhibit content DIF if it utilizes knowledge that is not uniformly available to all subgroups within the population of examinees. An example of content bias against females is found in an item in which students are asked to compare the weight of several common objects, including a football. Since girls are less likely to have handled a football as frequently as boys, they might find the item more difficult than males, even though they have mastered the concept measured by the item (Scheuneman, 1982).

An item may be considered functioning differentially if it uses terms that are not commonly used throughout the examinee population, or which have different connotations within different subgroups. An example of language DIF against Blacks was found in an item where students were asked to identify an object which began with the same sound as "hand." While the correct response was "heart," Black students more often chose "car" as their response because, in Black slang, a car is referred to as a "hog." The Black students had mastered the concept, but were getting the item wrong because of language differences (Scheuneman, 1982).

Also, an item may be considered to function differentially in terms of structure and/or format if the structure or format is constructed in such a way as to favor or hinder a particular subgroup of examinees. An example of this form of bias may be found in an item which contains directions or graphics that may be more familiar to a particular subgroup of examinees.

The initial step in a DIF study is to identify the subgroups of interest. For the purpose of this study, examinee test data from a statewide proficiency test was used and attention was focused on Native Americans and Anglo-Americans. The item bias review form in Figure 1 was constructed specifically to address DIF and stereotyping in Native American and Anglo-

American subgroups while adhering to the five principles of judgmental review described earlier.

- -
Insert Figure 1 about here.
- -

:n an earlier study, the State Department of Education located eight educational specialists, representing the various factions or subdivisions within the Native American community. These specialists agreed to review the 150 test items of the Statewide Proficiency Test. Judges received through the mail a copy of the test, an item bias review form, and a set of directions. Our preference was to bring the judges to a central place to permit training and group discussion; however, this was not possible because of the costs involved.

Based on the results of this initial study the item bias review form was substantially revised (see Figure 1). Specifically, the number of questions was reduced from 13 to five, and the five remaining questions were revised to improve their clarity. Seven new judges were approached from the Native American community, and agreed to review the test items. The judges included three university professors, two graduate students, and two school principals. All seven persons were working in the field of education.

Findings from the first round of reviews revealed that the review of 150 items by an individual judge was too long a task. Hence, judges involved in the second round of item reviews were asked to review only 75 items. The 75 items reported on by Hambleton and Rogers (1989) were used in this study.

Description of the Test Data and Examinee Samples

Data from the Statewide Proficiency Test was used to generate the examinee samples. The test is designed to assess skills in five major areas: Knowledge of Community Resources, Consumer Economics, Government and Law, Mental and Physical Health, and Occupational Knowledge. The data set contained the responses of approximately 23,000 students to the 4-choice, 150-item test. In the total group of students, approximately 8,000 were Anglo-American and 2,000 were Native American.

Empirical Methods

A popular definition of DIF states that an item exhibits DIF if examinees of the same ability but from different sub-groups do not have the same probability of a correct response to the item. Item response theory (IRT) based DIF detection methods are popular currently and also considered by some (e.g., Shepard, Camilli, & Averill, 1981) to be "theoretically preferred" as a consequence of their close connection to this most widely accepted definition of DIF. In essence, the study of DIF within an IRT framework simply necessitates the comparison of item characteristic curves (ICCs) for the target sub-groups (Hambleton, Swaminathan, & Rogers, 1991). The total area between the two ICCs obtained for the groups of interest is directly related to the differences in probabilities of success for the two groups at every level of ability and is thus a natural index of DIF. In this study the area between two ICCs between the lower group mean minus three standard deviations to the upper group mean plus three standard deviations was calculated. A "cut-off" value was obtained for interpreting the total area statistics by observing the performance on two randomly equivalent groups (the two Native American samples). Because there can be no DIF present between these two groups, the largest IRT area statistic obtained may then serve as an indicator of the greatest value of the statistic likely to occur by chance.

A popular method of DIF detection is the Mantel-Haenszel (MH) procedure proposed by Holland and Thayer (1988). Similar to IRT-based methods, the MH procedure compares the probabilities of a correct response in the two target groups for examinees of the same test score. This method was applied to the item responses for the two groups of interest -- i.e., Native Americans and Anglo-Americans.

Procedure

The basic data came from the 2,000 Native Americans and a random sample of 2,000 Anglo-Americans who took the test. Each ethnic sample was then subdivided in an odd/even split to obtain two non-overlapping samples of 1,000 for each ethnic group. The creation of two Anglo-American and two Native American samples enabled two independent DIF analyses to be performed (i.e., cross-validation of the results was possible). A second comparison was deemed useful to facilitate an examination of the consistency with which the empirical procedures identified differentially functioning items. Items not consistently identified as DIF by a procedure were to be eliminated from subsequent analysis.

In an earlier study investigating the agreement between IRT area based procedures, MH, and item bias review procedures, 75 items out of a possible 150 contained in the test were analyzed in order to reduce computer time and to facilitate the analyses. Items were dropped if they had unusually low item discrimination indices (point biserial correlations < .10) in the combined samples or unusual difficulty levels (p < .10 or p > .90). Such items cause problems in IRT parameter estimation. The first 75 items in the test which met the criteria were used in the analyses. These 75 items were also the items used in the second round of the item bias reviews.

Results

Empirical Methods

Following the estimation of three-parameter model estimates for items and examinees for the two Anglo-American and two Native American samples, the appropriateness of the fits between the model and test data were determined by the calculation of standardized residuals (Hambleton & Swaminathan, 1985). The fits for all four datasets were found to be excellent. In all four analyses, the distributions of standardized residuals were approximately normal with mean 0 and standard deviation 1. (For more details on the fit results, see Hambleton & Rogers, 1989.)

The cut-off value used for the MH statistic was 6.64, the tabulated value of the chi-square distribution with one degree of freedom at the .01 alpha level. A cut-off value of .468 was established for the IRT area statistic.

Of special interest in this research was the total set of items which were consistently identified as differentially functioning by the IRT area procedure, the MH procedure, or both. The 16 items which were flagged consistently (across two independent studies) by either the IRT area or MH procedure are presented in Table 1. A more detailed presentation of these results including the IRT parameter estimates and the DIF statistics was given previously in Hambleton and Rogers (1989).

- -

Insert Table 1 about here

- -

The total area procedure identified 14 items consistently across the two comparisons while the MH procedure identified nine items consistently. In total, 16 of the 75 items were identified by one or both of the procedures. Seven items were common to the two lists of DIF items: Items 46, 56, 61, 62, 68, 73, and 75. The two by two table below was used in computing the consistency of item classifications (DIF vs. non-DIF) across procedures of .88 and a kappa value of .56:

		MH Procedure		
		non-DIF	DIF	Marginals
IRT-Area Procedure	DIF	7	7	14
	non-DIF	59	2	61
	Marginals	66	9	75

After the unreliability of the individual procedures is taken into account (by focusing on only items which showed DIF in a cross-validation sample., consistency of item identification appears to be high.

An examination was undertaken of the nine items consistently flagged by only one DIF detection method. Table 1 summarizes these findings. Item 3, in addition to being consistently flagged by the MH statistic, was also flagged in one comparison by the IRT Area procedure. This suggested that the discrepancy in the results with regard to item 3 was likely due to a Type II error with the IRT area method. Conversely, items 30 and 57 were both flagged consistently by the IRT area procedure as well as by one comparison of the MH statistic. Thus, this discrepancy was likely due to a Type II error with the MH procedure. Consequently, items 3, 30, and 57 were considered to be determined consistently differentially functioning items.

ICCs for the two groups were plotted and compared for the remaining six items. Plots of four of these items (items 12, 13, 51, and 74) were found to cross markedly. It is likely that the MH statistic could not identify these items because this procedure is not designed to detect non-uniform DIF. These items, too, therefore were added to the pool of differentially functioning test items which were consistently determined. In all four instances, the items were less discriminating in the Native American samples than in the Anglo-American samples. Item 48 was found to be differentially functioning against the Anglo-American samples and therefore this item was not considered further in our work.

Finally, examination of item 32 found that the item had been flagged by the MH procedure; the ICCs were uniformly different at the low end of the ability scale, but nearly identical for high ability examinees. It was probable that the MH procedure identified item 32 as exhibiting DIF because the most pronounced differences were in the region on the ability scale where many Native American examinees scored. Thus, this item was added to the list of differentially functioning items.

The result was that all of the items listed in Table 1, except for item 48, were viewed as functioning differentially because consistency in this classification was observed (1) over parallel analyses on randomly equivalent samples and (2) over related DIF statistics. These 15 items were the ones of special interest in the analysis of the judgmental data.

Analysis of the Judgmental Data

Preliminary Results

A preliminary analysis of the item bias review data received from the seven judges using the new form was as follows:

Judge	Number of "Yes" Marks (Faults)
1	18
2	1
3	91
4	134
5	7
6	174
7	68

Two judges essentially found no problems at all, and two other judges found an average of two faults per question! These data are not really believable. The tremendous variability among the judges was disappointing, and neither extreme pattern of ratings seemed realistic. The new data were not perceived as having sufficient levels of reliability or validity by themselves to address the purpose of the study. As a result, the judges' ratings were combined with ratings collected using the earlier item review form. For the purposes of analysis, questions on the longer form were combined to match the new shorter form:

Earlier Form	New Short Form
1, 2, 3	1
4	2
5, 6	3
9, 11	4
7, 8, 10	5

The first set of ratings were not without problems either. Two of the judges tended to rate early items fairly critically and later items fairly leniently. Clearly there was an order effect in the DIF they identified. These judges ratings were removed from the analysis. Another judge was removed because he tended to find fault with every item. The final result was a set of ratings on 75 items from 15 judges (8 of 11 who used the long form, and 7 who used the short form).

Detection of DIF

Test items were identified for which (at least) six of the 15 judges found a fault (i.e., answered "yes" to at least one of the questions on the item bias review form). The analysis, though more subjective than we had planned, and based upon samples of judges using two different forms, resulted in the identification of 11 test items. No items were identified because they contained "stereotyping" (question 1). The 11 items and the reasons for their identification by the judges (supplemented by our own post-hoc judgments) are presented in Table 2. Of the 11 test items, 5 were identified by the empirical procedures.

- -
Insert Table 2 about here
- -

Of the 10 additional test items detected by empirical methods but which went undetected by the judges, a number of observations can be made. One, eight of these 10 test items (30, 32, 46, 51, 57, 62, 73, 75) included a negative word in the stem (i.e., not, except, or least) or required negative thinking (e.g., what did I do wrong?). Apparently, the meaning or readability of these items was unclear to some Native Americans. Interestingly, there were only 12 negatively worded items or items requiring negative thinking in the test -- eight of them were identified as DIF using the empirical procedures but only one (63) was spotted by the judges and that one most likely because of another obvious flaw in the item. Two, of the two additional items undetected by the judges (12 and 56), one seemed to require some prior knowledge that Native Americans did not possess. The second item could be answered more successfully by test-wise candidates. Possibly the scores of the Anglo-Americans on this item were inflated due to the influence of test-wiseness skills.

The two-by-two table below was used in determining the consistency of item classifications (DIF vs. non-DIF) between the empirical and judgmental procedures. Consistency of item classifications across procedures was .73; kappa was .28.

		Statistical Procedures		
		non-DIF	DIF	Marginals
Judgmental Procedure	DIF	10	5	15
	non-DIF	54	6	60
	Marginals	64	11	75

While agreement was not as high as might be desired, agreement in identifying DIF was over twice what was expected by chance. Of course the percents are unstable because of the small numbers identified as DIF.

Discussion

Several major points emerged from the analyses. First, both IRT-based methods and the Mantel-Haenszel method were somewhat unreliable in identifying differentially functioning items. This result helps to explain the moderate agreement reported in the measurement literature among approaches to DIF. The fact is that studies of overlap of results with methods for investigating DIF are influenced considerably by the unreliability of the statistics. There appeared to be substantial agreement between an IRT-based procedure (the IRT Area method) and the Mantel-Haenszel method in the detection of DIF when only items which showed DIF in a cross validation sample were considered in the analysis.

Second, our work appeared somewhat successful with the item bias reviews. Five of eleven items identified by the judges as potentially biased were identified as DIF by the empirical procedures. With a couple of changes in the item bias review form, the results would have been even better -- for example, ask judges to identify test items with negative words or ideas in the stem, and search for test items that require prior knowledge that may be less present in Native Americans than in Anglo-American students. Of course, the generalizability of these recommendations to other editions of the test or to other basic skills tests or to other ethnic groups, such as Hispanics who take the test used in this study in very large numbers, is unknown.

Third, of the 15 items which were consistently identified as DIF (across empirical methods and samples) to the disadvantage of the Native Americans, five of those items were identified by the judges. Thus, the use of the item bias review form in the test development process in advance of any test administrations would appear to be helpful. Still, the item bias review process was not implemented as well as would be desirable. The judges ratings used in this study had some fairly serious problems: There was evidence of (1) multiple definitions of what constitutes bias in test items and (2) lack of careful attention. Perhaps in the future we will be able to (1) standardize the training of judges, (2) provide a ratings context for judges in which they know their ratings will influence the test development process (i.e., that their ratings will be taken seriously), and (3) provide a manageably sized task that judges can carefully attend to from the beginning of the process to the end.

Finally, the results also suggest some promising revisions for the item bias review form, especially in the areas of item content and item readability. Failure to attend to these two factors appeared to explain most of the non-agreement between the empirical and judgmental procedures for DIF in this study.

The implications of the results of this study for practice seem clear. First, test developers should be reminded about the unreliability of DIF statistics. This means that they should be encouraged to use large samples in their analyses whenever possible and interpret the statistics with a fair degree of caution. Second, the evidence suggests that the Mantel-Haenszel procedure can be safely substituted for IRT-based methods if safeguards are put in place to detect non-uniform DIF. Some of these items are likely to go undetected by the Mantel-Haenszel procedure. Finally, and most importantly, there is some evidence that a judgmental process can be effective in identifying test items that may be DIF in practice. And, careful analysis of items which are identified as DIF using empirical methods may be helpful in redesigning item bias review forms. By so doing, more effective item bias reviews can be carried out. This suggestion seems especially applicable within an on-going testing program. How useful a "tailored" review form for one test will be for another, or even how useful the form will be for identifying multiple types of DIF remains to be determined.

References

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

Engelhard, G., Jr., Hansche, L., & Rutledge, K. E. (1990). Accuracy of bias review judges in identifying differential item functioning. Applied Measurement in Education, 3, 347-360.

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

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

Hills, J. (1989). Screening for potentially biased items in testing programs. Educational Measurement: Issues and Practice, 8(4), 5-11.

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.

Hoover, H. D., & Kolen, M. J. (1984). The reliability of six item bias indices. Applied Psychological Measurement, 8, 173-181.

Plake, B. S. (1980). A comparison of a statistical and subjective procedure to ascertain item validity: One step in the test validation process. Educational and Psychological Measurement, 40, 397-404.

Scheuneman, J. D. (1982). A posterior analyses of biased test items. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 180-198). Baltimore, MD: The Johns Hopkins University Press.

Scheuneman, J. D., & Bleistein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. Applied Measurement in Education, 2(3), 255-275.

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.

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

Table 1

Consistently Identified DIF by the IRT Area or MH Procedures

Items 1 to 75 (original)	IRT Area	MH	Agreement Between Procedures	Reason For Non-Agreement
3 (11)		X		Type II Error
12 (28)	X			Non-Uniform Bias
13 (30)	X			Non-Uniform Bias
30 (57)	X			Type II Error
32 (60)		X		Choice of Interval
46 (82)	X	X	X	
48 (88)[1]	X			Non-Uniform Bias
51 (92)	X			Non-Uniform Bias
56 (101)	X	X	X	
57 (102)	X			Type II Error
61 (107)	X	X	X	
62 (110)	X	X	X	
68 (122)	X	X	X	
73 (128)	X	X	X	
74 (129)	X			Non-Uniform Bias
75 (130)	X	X	X	

[1] This item was deleted from subsequent analyses because DIF favored the Native Americans.

Table 2

Items Identified by Judges as Potentially Biased

Item[1,2]	Reason
3 (11)*	Not appropriate content -- some Native Americans have limited knowledge of the telephone system.
4 (14)	Not appropriate content -- unfamiliar to Native Americans.
8 (20)	Judges were opposed to the content and item format. They also felt some of the terms would be unfamiliar to Native Americans.
13 (30)*	A number of judges felt that the item required knowledge that would not be possessed by all Native Americans.
41 (73)	The correct answer -- lobbyist -- would not be familiar to all Native Americans.
58 (104)	Several judges felt that the item required knowledge that was not part of the Native American experience.
61 (107)*	Some judges felt that some Native Americans did not have the prior knowledge to answer the question -- health care during pregnancy.
63 (112)	Generally, judges felt that cultural differences were not taken into account when preparing the item scoring key.
68 (122)*	Some judges felt that the content would be unfamiliar to Native Americans.
72 (127)	A number of judges felt that some of the vocabulary would be unfamiliar to Native Americans.
74 (129)*	Some judges felt that there was no clear correct answer to the question.

[1] Original item number appears in brackets.

[2] Items with a "*" were identified by one or both of the empirical methods.

Figure 1. Revised Item Bias Review Form.

Item Bias Review Form

Reviewer: ______________________ **Date:** __________ **Test:** High School Proficiency Examination

Directions: Read each test item and then answer the 5 questions below. Mark "Y" for Yes, "N" for No, and "?" for unsure.

Question	Item														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Does the test item contain stereotyping of Native Americans (e.g., stereotypical situations or characteristics) or contain material which may be emotionally charged for Native American students?															
2. Does the test item contain content or words that may be less familiar or have a different meaning for Native American students than for other students?															
3. Does the test item include information and/or require skills that may not be expected to be within the educational experience of Native American students?															
4 Will the item format or the stimulus material (directions, graphics, tables, figures, etc.) be more difficult for Native American students than for other students?															
5. Will any of the incorrect answers be more attractive to Native American students thanto other students for cultural reasons?															