

DETECTION OF ITEM BIAS  
USING ANALYSES OF RESPONSE PATTERNS

David L. McArthur

---


CSE Report No. 163  
1981

CENTER FOR THE STUDY OF EVALUATION  
Graduate School of Education  
University of California, Los Angeles

Appreciation is extended to Beverly Cabello for her analysis of cultural and linguistic issues.

---

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.



## ABSTRACT

Item bias, when present in a multiple-choice test, can be detected by appropriate analyses of the persons x items scoring matrix. Five related schemes for the statistical analysis of bias were applied to a widely used, primary skills multiple-choice test which was administered in either its English- or Spanish-language version at each of the two levels, to 1259 students in bilingual education programs. The results indicate that from one-fifth to one-third of the items in the tests show strong evidence of bias, corroborated by a separate analysis of linguistic and cultural sources of bias for both the biased items and those items with no statistical findings of bias.

## Introduction

A systematic but unanticipated pattern of responses to a multiple-choice test found for an entire group of test-takers is generally regarded as evidence of bias. This interpretation results from indications of one or more differences between groups on levels of knowledge and skill, or in linguistic and cultural issues related to the use of language in the test. However, the behaviors of individual respondents have important consequences for that interpretation. Whether the respondent unerringly picks the correct response, or successfully engages in elimination of incorrect answers, or guesses well, the observer scores the item "correct" and concludes that the student "knows" the required skills or material. The inference that the respondent "does not know" is made whether he/she guesses incorrectly, eliminates wrong choices badly, or chooses an attractive but incorrect alternative.

Most likely, phenomena looking like systematic patterns of bias in test items are the results of complex interactions of these group and individual factors with one another and with certain properties of the test items. What is required to make sense of the issue of bias is analysis of patterns found in these combinations of performance. The multiplicity of possible patterns suggests that the detection and interpretation of bias must be conducted along several routes.

## Goals of This Research

The first of two purposes of this paper is to investigate analyses of the persons x items scoring matrix of a test for the detection of item bias. The persons x items scoring matrix contains a significant amount of information about the patterns of responses generated by a set of examinees.

Using a few geometrical and statistical considerations, the patterns of responses from separate groups of examinees tested with the same instrument can be compared. If these patterns show that the test is not measuring the same thing -- skills, competence, thinking abilities -- in comparable groups, if the groups are responding to different aspects of the test items, or if cultural and/or linguistic issues take precedence, it may be that the test is biased.

The second purpose of this paper is to study empirically the question of bias as shown by these several techniques in the context of a widely used achievement test, the Comprehensive Tests of Basic Skills (CTBS), which has been translated from English into Spanish. The claims made about this instrument include the statement that the Spanish-language version represents a close replicate of the English-language version with careful attention having been exercised in removing all forms of unintended bias. The primary task of this analysis is to ascertain the degree of comparability of the two versions of the CTBS in the assessment of similar groups of children, and to see if any bias remains.

#### Related Literature

A substantial research literature has developed around the term "item bias" in the search for a single best all-purpose indicator which always reveals bias whenever systematic discrepancies in performance between groups are found. A large number of methods have been proposed and a large number of studies conducted (cf. reviews in Berk, in press; Subkoviak, Mack, & Ironson, 1981). Certain tests such as the Wechsler Intelligence Scale for Children have been extensively investigated (cf. Sandoval, 1979). The range of applications of the term "bias" is quite broad: studies have examined

sociocultural bias and the stereotyping of items and answers, cultural differences, and linguistic variations (cf. Jensen, 1980); construct bias and the different aspects of performance tapped in different examinee groups by the same test (cf. Evel, 1975); and contextual bias and the misuse of tests with specific groups (cf. Williams, 1971). Occasionally the word is even used to mean a conscious preference on the part of the examinee (Hudson, 1963).

Increasingly complex techniques have been set forth for the detection of bias in items. Methods have been based on analysis of variance, transformed item difficulties, factor techniques, adjusted chi square procedures, distractor analyses, "adverse impact" and item characteristic curves (Merz, 1980; Petersen, 1980; Rudner, Getson, & Knight, 1980). Many of these methods are statistically complex but, with the exception of the last, statistically inelegant (Hunter, 1975); unfortunately the most elegant solution, item characteristic curve analysis, requires large numbers of items and respondents for its computation. Few of these approaches offer convincing or useful explanations of why some items are biased and others are not (Crowder, 1979). Faced with the multiplicity of both the forms of item bias and the statistical methods that have been put forward to detect such bias, one logical place to begin is to inquire about the nature of a test which is absolutely free of bias.

#### An Unbiased Test

If a test could be created which fulfilled all of the requirements of a bias-free instrument, its items would all measure the same trait or ability and be equally reliable and equally valid for all groups (Petersen, 1980). It would also show orderly variation in the relative difficulties of the

items, and be responded to in an orderly manner by every individual. One example of the outcome of this improbable creature is the familiar perfect Guttman scale, in which persons are perfectly ordered by increments of skill level, and items within the test are perfectly ordered by increments of difficulty. No higher-level item is mastered by any respondent until each lower-level item is mastered; guessing also plays no role. The sequence of successes and failures is highly deterministic.

Figure 1A represents a ten-item test with right/wrong scores for ten respondents. These ten persons never successfully answered a more difficult item without first having succeeded on a less difficult item. An axis of

-----  
Insert Figures 1A and 1B about here  
 -----

performance can be drawn on the diagonal to separate all correct scores from all incorrect scores. While the total p-value for the test is lower for another group of ten persons tested on the same ten items, shown in Figure 1B, the performance patterns are parallel. Other than a main effect due to groups, nowhere in either diagram is any indication of a systematic unexpected difference in the pattern of responses or bias in the test.

#### A Slightly Biased Test

A somewhat less artificial example of test results from a multiple-choice test is shown in Figure 2A; the score matrix of a hypothetical ten-item

-----  
Insert Figures 2A and 2B about here  
 -----

test has been sorted by both persons, on ascending total score, and by items, on ascending level of difficulty. Neither persons nor items is perfectly ordered in the sense used above, and guessing of correct answers probably contributes by an unknown amount to the scores obtained. Not one but two

dividing lines are now required to separate the patterns of performance in this figure. The first line, a cumulative ogive representing student performance, is drawn on the matrix based on the total correct score for every respondent. The second, representing problem difficulty, is drawn as a cumulative ogive based on item p-values. Note that for a test which demonstrates exclusively random responding, the theoretical position of the student curve (S-curve) would be vertical, and of the problem curve (P-curve), horizontal.

At this juncture we introduce a second set of data obtained from the same hypothetical test. The "respondents" were slightly less capable on most items but all other considerations were held equal. A score matrix for the same set of items as shown in Figure 2A but now with the second group of examinees is shown in Figure 2B. The relative order of items is somewhat changed because of differing levels of difficulty; the second group performs less well overall than the first group. Statistical differences between the data in Figures 2A and 2B should reflect overall item and group differences, but because of the idealized symmetry between the two, there is little likelihood that a statistical indicator of bias would prove significant. An initial analysis of these figures recommended by Jensen (1980) is a two factor (group x items) nested analysis of variance. The interpretation of a significant groups effect, in the absence of other significant factors, is that the groups behave symmetrically with respect to ordering of item difficulties but that one group is consistently more capable across the trait being appraised by this test. A significant difference on both the groups and items factors, plus a significant interaction between groups and items, together suggest that the test items and examinee abilities in the two groups are



heterogenous.<sup>1</sup> However, these findings would be quite insufficient to say that the test is biased (Hunter, 1975) and, additionally, do not account for the contribution of guessing.

A second approach recommended by Jensen (1980) for understanding the differences between the two figures uses the phi coefficient, which is the correlation obtained between the group response to a given item and the same group's response to any other item in the test. Phi is a measure of joint contingency; Jensen explains its use for analysis of bias:

---

Only if the two items have the same difficulty...can phi be equal to 1...To determine the intrinsic correlation (of the items) free of the influences in item difficulty, we must divide the obtained phi by the maximum value of phi that could possibly be obtained with the given marginal frequencies (p.431).

---

The ratio of phi to maximum value of phi is summed over all possible pairs of items for each group, and then the ratios are compared. The null hypothesis for this comparison is that the difference between the obtained sums is not different from randomness, and thus there is no systematic discrepancy in group performance. In the artificial situation shown by the Guttman scale for both groups in Figure 1, this test is necessarily nonsignificant. For data which do not fit the mandates of a perfect scale, the obtained value for the comparison of ratio sums increases as the discrepancy in overall patterns of response by the two separate groups widens.<sup>2</sup> While the amount of

---

<sup>1</sup>The comparison of Figures 2A and 2B yields only a significant difference on the factor of items ( $F(9,162)=13.98, p<.001$ ).

<sup>2</sup>For the difference between Figures 2A and 2B,  $\chi^2= 8.0222, p<.01$ .

difference between groups is given by the analysis of variance and phi, the nature of patterns of response to items is not adequately explained.

Only a small number of statistically-based analyses specifically designed to study patterns of responding to multiple-choice tests have been proposed. Tatsuoka (1981) and Harnisch and Linn (1981) have been working on a norm conformity index and other parameters which address each individual's performance in the context of patterns obtained by all members of the group. Sato (1980) defines an index of disparity between actual and ideal response patterns which can be applied to individuals or to items. To unravel the problem of patterns, we now turn to Sato's system of analysis of the persons x items matrix.

#### The S-P Method and Analysis of the Person x Items Matrix

The key element in Sato's (1980) S-P method of analysis of test performance is the doubly-ordered persons x items matrix, with student curve (S-curve) and problem curve (P-curve) drawn in. In Japan, this procedure is widely used in classrooms to obtain the characteristic performance of the set of examinees, which may be compared visually to several "standard" curve functions for diagnostic purposes.<sup>3</sup>

Sato has developed an index of discrepancy to evaluate the degree to which the S and P curves do not conform either to one another or to the Guttman scale. Except in the case of the perfectly ordered sets shown in

---

<sup>3</sup>Direct interpretation of item scores, person scores, and the amount of discrepancy between the S and P curves is relatively easy to accomplish; the same holds for item analysis, individual performance analysis, and other summary statistics within a group. In Japan, this system has been automated using a microcomputer (Sato, Takeya, Kurata, Morimoto & Chimura, 1981).

Figure 1, there is always some degree of discrepancy between curves. The index is explained as follows:

$$D^* = \frac{A(N, n, \bar{p})}{A_B(N, n, \bar{p})} \quad \text{where the denominator}$$

...is the area between the S curve and the P curve in the given S-P chart for a group of N students who took n-problem test and got an average problem-passing rate  $\bar{p}$ , and  $A_B(N, n, \bar{p})$  is the area between the two curves as modeled by cumulative binomial distributions with parameters N, n, and  $\bar{p}$ , respectively (Sato, 1980, p. 15).

The denominator is a function which expresses a truly random pattern of responses for a test with a given number of subjects, given number of items, and given average passing rate, while the numerator reflects the obtained pattern for that test. As the value of this ratio approaches 1.0, it portrays an increasingly random pattern of responses. For the perfect Guttman scale as represented by Figure 1, the numerator will be 0 and thus  $D^*$  will be 0.<sup>4</sup>

Indices of discrepancy, when computed for each of two groups of examinees, may not be statistically compared because of differences in ranking of item difficulty, and/or compound differences in response patterns to several items. However, as long as the two  $D^*$  values obtained are not equivalent, it is an indication that somewhere within the matrices are one or more items which are behaving dissimilarly across groups.

#### Analysis of Respondents Above P Curve

Patterns of discrepant performance result from a mixture of random behaviors and wrong choices, except for those items which are so easy that no respondent gets them wrong. Aside from the tautology that respondents

---

<sup>4</sup>In Figure 2A,  $D^* = .2534$ ; in Figure 2B,  $D^* = .3747$ .

with less ability are less likely to answer a given item correctly, all other things being equal they are also likely to use chance responding. Analysis of those respondents who are unlikely to be answering randomly would seem a likely means to understanding patterns and bias in items. To begin constructing a simple analytic solution to this problem, suppose we take a single uncomplicated item from the S-P chart, and examine the pattern of responses for only that portion of the same group of examinees for whom the prediction of success is relatively high, i.e., those above the P-curve. These are the examinees who tended to score better overall. Specifically, respondents at the very top of this select subgroup are expected to have had a finite but small probability of having guessed their way to success. Respondents at the bottom of this select subgroup would have a finitely larger probability, while those at the very bottom of the entire S-P chart would be likely to have a more random pattern.

If the selected item, however, is one for which no individual within the sample, no matter how skilled, is able to answer knowledgeably, the response pattern among the select group of putative "masters" should be random, and should not differ from the response pattern of those examinees not included in this subgroup. For a four-choice item of this kind, the item's p-value should be about .25, and the select subgroup of putative "masters" would be correct only 25% of the time. Figure 3 illustrates a pattern of responses for a nearly random item, in contrast with an item which is fairly well-fitted to the skills of a set of respondents.

-----  
Insert Figure 3 about here  
-----

The proportions of "masters" who are indeed correct can be compared between groups. With relatively uniform variances, the test of significant

difference in independent proportions applied to this problem yields a z score; a significant z score would be an indication of possible bias separate from the difference in average passing rates for that item, if any. A comparison of nonuniform variances requires transforming the item difficulties into standard score form, then testing the size of the difference following Rudner, Getson and Knight, (1980). Within certain limits, an item which is relatively easy for one group and relatively difficult for another, may show no bias in the proportions of "masters" who are correct because those individuals who place above the P curve all have the ability to answer that item correctly. However, on another item one of two groups may not be academically equipped, or may be prevented from responding by biases in the test, curriculum, or culture; thus the proportions may differ, possibly by an amount sufficiently large to be deemed significant.

#### Analysis of Distractors

One further analysis of the potentially biased item is to examine the patterns of wrong answers made by the separate groups of respondents. Within the multiple-choice test format, differences between groups in the attractiveness of incorrect responses signal that the item's wrong choices may be differentially distracting. When a given item has attractive but incorrect responses for one group, Goodman and Kruskal's Lambda indicates whether another group shares the same proportional pattern of selecting those incorrect responses (Veale & Forman, 1976). Lambda is an index of predictive association, which shows "...how one is led to predict differentially in light of the relationship..." (Hayes, 1963, p. 610, italics original). It is calculated for a problem involving two groups by evaluating the largest discrepancy between rates of responding to similar wrong choices:

$$\lambda = \frac{\sum \max.f_{jk} - \max.f_{.k}}{N - \max.f_{.k}}$$

where  $\max.f_{jk}$  is the larger frequency of the two groups for any single wrong choice, and  $\max.f_{.k}$  is the larger marginal frequency of the two groups summed across all wrong choices.

In Goodman and Kruskal's lambda is appreciably above zero, the interpretation can be made that the pattern of distraction is different for the two groups. If the index is zero, even though the difficulty of the item and/or the proportions who select a wrong option may differ between the two groups, the pattern of selecting the wrong answers is about the same.

Another check on the relative attractiveness of a wrong answer can be made by counting the number of wrong answers which are chosen at least 10% more often than the next most popular wrong answers. These particular wrong choices constitute a class of "popular distractors," each of which can be studied further. The easiest comparison is between those items for which both groups picked the same popular distractor and those items for which both groups picked different popular distractors. Note that in this latter case, the computation of lambda will always yield a nonzero value.

A series of analyses of item bias has been described, with special attention paid to those comparisons premised on the persons x items scoring matrix, doubly sorted. The following sections describe the execution of these analyses in the context of a multi-language achievement test.

## METHOD

### Instruments

For a study of the possible bias inherent in a multi-language test, two levels of the Comprehensive Tests of Basic Skills (CTBS) published by CTB/McGraw Hill (1974, 1978) were administered in this study. Students in grades 2 and 3 were given the CTBS Level C; participating fifth and sixth grade students took Level 2. CTBS-English Level C is designed for students in grades 1.6 to 2.9; CTBS-Spanish Level C is designed for students in grade 2. CTBS-English level 2 has a target population in grades 4.5 to 6.9; the Spanish translation was designed for students in grades 5 and 6.

The CTBS-English and CTBS-Spanish tests were selected for several reasons. Test content is roughly parallel. The CTBS-Spanish was the first test at CTB/McGraw Hill to be subjected to a four-step editorial procedure designed to reduce test bias; included were studies of content validity, application of editorial guidelines in item construction, reviews for bias, and separate ethnic group pilot studies with the test. In the translation of the CTBS from English to Spanish, the test developers tried to keep the test content and measurement features intact. This, of course, meant that in some cases word-for-word translations were not possible. Nevertheless, the publishers's intent was to provide tests that are similar in rationale and in the process/content classification scheme. Thus, both the English- and Spanish-language versions used in this study purport to measure the following objective:

1. the ability to recognize or recall information
2. the ability to translate or convert concepts from one kind of language (verbal or symbolic) to another

3. the ability to comprehend concepts and their interrelationships
4. the ability to apply techniques, including performing operations
5. the ability to extend interpretation beyond stated information  
(CTBS, 1974/1978)

Test length, test time, and administration procedures are exactly the same for English and Spanish versions of each test level.

### Subjects

Five school districts in the state of California participated in the study. The total number of pupils tested was 1259, representing 81 intact classrooms.

Classrooms were selected to represent a wide range of program options. The criterion for selection of school districts was that they had bilingual-bicultural education programs funded by Title VII. Potential participants were identified from schools listed in the California State Department of Education 1979 Bilingual Program Directory. From this list, invitations were sent to schools which had at least two classes at the same grade level (grades one, two, five or six) having bilingual programs. Additionally, instruction had to be delivered in self-contained, multisubject settings; departmentalized or pull-out programs were excluded.

### Analyses

Five statistics explained above were used to evaluate the data for every item separately. Each uses a minimum threshold value, above which the result is taken as an indication of possible bias in the item. The analyses and their minimums can be summarized as follows:

- a) Test of proportions of correct scores: across groups, a difference between transformed p-values which generates a  $z > 1.96$ ;



- b) Test of proportions of correct scores for "masters": across groups, a difference between proportions of those respondents above the P-curve who make errors, which generates a  $z > 1.96$ ;
- c) Test of chance responding by "masters": within each group, a difference between the obtained proportion of those passing the item and a theoretical p-value of .25, which generates a  $z < 1.96$ ;
- d) Test of differential attractiveness of wrong answers: a Goodman and Kruskal's lambda computed on the proportions of incorrect answers by choice within item, such that  $\lambda > 0.0$ ;
- e) Test of popular distractors: a wrong choice for an item attracting at least 10% or more responses than the next most popular wrong choice for that time.

---

## Results

The number of items within each subtest by level, and the number of students in each of two language groups who were included, are shown at the top of Table 1. Item P-values indicate that items ranged from moderately easy to very difficult for both language groups, with an overall mean of somewhat over half of the items correct. While in a few items the Spanish-language

-----  
Insert Table 1 about here  
 -----

group did better, without exception the Spanish-language groups always scored lower overall on the subtests. In every instance the maximum p-values achieved by the English-language groups are slightly higher than the comparable scores for the Spanish language groups. Table 1 also shows for the corresponding number of students, the p-value needed for a significant ( $p < .05$ ) difference from chance responding to an item. This figure is obtained by reversing the usual computation for the test of independent proportions, using  $z = 1.96$  and  $P_{\text{chance}} = .25$ . For all but one of the subtests, both language groups had one or more items which appear to represent random choice of the correct answer. Except for the Passage Comprehension subtest at Level C, the Spanish-language

group appears to make random selections more often than the English-language group, an assumption which is further explored below.

For purposes of illustration, two analyses recommended by Jensen (1980) were conducted on the subtest with the smallest number of items, Level C Passage Comprehension. The two-factor nested analysis of variance for this subtest shows a significant effect due to the groups factor ( $F(1,650) = 54.91$ ,  $MS_{\text{error}} = 1.37$ ) and a significant effect due to the interaction between items and groups ( $F(17,11050) = 2.61$ ,  $MS_{\text{error}} = 0.43$ ). The ratio of  $\phi$  to  $\phi$ -max is higher for the English-language sample than for the Spanish-language sample (English mean  $\phi/\phi$ -max = .8207; Spanish mean  $\phi/\phi$ -max = .766,  $t(151) = 4.01$ ,  $p < .01$ ). This brief set of findings indicates only that the language groups are not performing the same way as one another on the subtest. It seems that the Spanish-language sample may have had more difficulty with some items than did their English-language counterparts. No further detail can be learned from these analyses, and they are not used in the study of the remaining subtests.

The S-P charts were drafted for each subtest by language group for a total of eight complete charts. The index of discrepancy  $D^*$  is presented in the last row of Table 1. The fact that the  $D^*$  values are higher for the Spanish-language groups suggests that they engaged in patterns closer to chance responding more often than did English-language groups. While the differences between pairs of  $D^*$  values are large for the Passage Comprehension subtest at both level C and level 2, these values cannot be compared further. The specific reasons why the Spanish-language versions generate larger  $D^*$  values can only be made evident with further analyses.

Results from the set of five analyses which together provide sufficient

evidence of patterns of discrepant performance are presented below and in Table 2. The table shows percentages of items for each of the four subtests in this study which exceed a critical minimum on each of the five analyses.

Test of proportions of correct scores. The first of the concise set of analyses is the test of proportions, which is applicable to percentages of correct answers expressed in standard score form, for both groups on each item of each subtest. The first two rows of Table 2 show the percent of items favoring the English- or Spanish-language groups. Six out of every ten items in the Vocabulary subtests show significant differences between

---

Insert Table 2 about here

groups; in a majority of instances the higher group is always the English-language group. Half of the items in the Passage Comprehension subtest at Level C show a significant difference and over three-quarters of the items in that subtest at Level 2 show a significant difference; in no instance are the Spanish-language groups ahead of their English-language counterparts.

Test of proportions of correct scores for "masters". Both the second and third analyses in this set are based on the selective sample of "masters", those students whose overall scoring position places them above the P-curve for each item. By evaluating the proportions of correct scores for those members of the language groups, a list of statistically significant discrepancies between "masters" is generated. The third and fourth rows of Table 2 show the percent of items within subtest for which the success rate among "masters" is significantly higher for the English-language or Spanish-language groups. The Passage Comprehension subtests at both levels appear to have different rates at which the "masters" are able to avoid the wrong answer; in the majority of instances the rate is higher for the English-language groups. In the Passage Comprehension subtests, the rate is uniformly

higher for the English-language groups.

Test of chance responding by "masters". How often the samples of "masters" are not able to choose the correct response at a rate better than chance forms a third part of the analysis. The fifth and sixth rows of Table 2 show that for the Level C subtests, no items are found for which either group responded randomly. However, for Level 2, a small number of items in both subtests elicited chance responding by "masters". These items appear to be so difficult that not even the better students could knowledgeably select the correct response. The Spanish-language group has a much larger number of chance responses among "masters" than the English-language groups on the Level 2 Passage Comprehension subtest.

Test of differential attractiveness of wrong answers. The fourth analysis in this sequence is the analysis of differential patterns of incorrect responses. Goodman and Kruskal's lambda was calculated for each item, using a 2 x 3 table of groups by incorrect response rates. Values ranged from 0.0 to .23, with a median of 0. Lambda will be 0 for any 2 x 3 table of proportions for which both groups are attracted to the same response, even if the actual dimensions of those attractions differ drastically. As there is no exact test of significance, any nonzero lambda was considered to be an indicator of possible bias. The seventh row of Table 2 shows the percentage of items within each subtest for which a nonzero lambda was found. The ratio of such items to the number of items within subtest ranges from 1:4 to 1:2, suggesting that, when wrong answers were selected the two language groups often behaved very differently.

Test of popular distractors. The concluding analysis in this series asks whether there are any incorrect choices which were sufficiently

attractive to be classed as popular distractors. In the final rows of Table 2 are shown the percentage of items which meet the 10%-or-greater criterion for the English-language groups, the Spanish-language groups, and jointly across groups. Except in Passage Comprehension at Level 2, the Spanish-language group's results show more items with popular distractors than the English-language group. Percent joint overlap is of particular interest, since that value gives another indication of the uniformity of behaviors across language groups when selecting incorrect responses. In the subtests in this study, the joint overlap of popular distractors is very small, suggesting again that many items of the English version of the test and the Spanish translation may not be as comparable as the test designers intended.

The degree of overlap between the five analyses in terms of the number of positive findings for each subtest is shown in Table 3. The percentage of

-----  
 Insert Table 3 about here  
 -----

items for which none of the preceding analyses show evidence of bias is remarkable small. Level C Passage Comprehension, for example, has only a single item which never shows a difference between the language groups. Over half of the items in that subtest have at least two positive findings, and four of the items have three positive findings. Table 3 shows that the percentage of items for which three, four, or five out of five statistical indicators yield positive results varies from about one-fifth to about two-fifths of the items within each subtest.

### Content Analysis

On the basis of the preceding evidence from the statistical approach to

bias detection in the CTBS, those items which showed agreement of three or more indicators were subjected to a careful analysis of item content. The content analysis was a search for possible linguistic, curricular, and/or cultural reasons which might explain differential performance between language groups. This portion of the study was undertaken by an educational researcher fluent in both English and Spanish, who made extensive reference to the curricular materials used by the students in the sample, and consulted with native speakers of various dialects in making an appraisal. Five categories were tabulated as possible sources of influence which item content might

---

exert on the different language groups:

- a) Mistranslation: the meaning and/or grammatical form of a key word or phrase within the item was translated from the English original in a manner which is an incorrect or inappropriate use of the Spanish language;
- b) Cultural bias: some key word or phrase within the item requires familiarity with objects, behaviors, or values which are not normally found in the Spanish and Latino cultures, or which may have very different interpretations;
- c) Linguistic bias: some key word or phrase within the item requires familiarity with an idiomatic expression or verbal allusions which, because of innate differences in language, do not translate well;
- d) Low frequency word bias: some key word or phrase within the item is not found, or rarely found, in the basal readers used for instruction by the students in our sample.
- e) Unfamiliar context bias: some key word or phrase within the item appears in a context which is quite different from that found for the word or phrase in the basal readers used for instruction.

An example of item content judged to bias respondents is shown by item number 29 of the level C Vocabulary subtest, an item for which all statistical indicators point to possible trouble. Item 29 (rated as category c, linguistic bias) requires the student to select a synonym for "happy". The English-

language version of the test yielded responses which appear significantly disadvantaged on this particular item. While the correct option for this item in the Spanish-language version, /alegre/, was selected 60% of the time by our sample, the correct option in the English-language version, / gay/, was selected only by 13% of the sample. The English-language respondents instead split their selection equally between two of the remaining options. Only one other item in the entire test set received as strong a rejection, suggesting that among second and third graders, the slang English-language meaning for 'gay' has not only rendered it useless as a synonym for 'happy' but has given it a strong pejorative flavor as well.

Table 4 shows data for items in each of the four subtests for which the content analysis identified probable sources of bias. The entries in the table

-----  
 Insert Table 4 about here  
 -----

represent tabulations of the content analysis categories for those items on each subtest which have three or more statistical indicators. For the Level C Vocabulary subtest, twelve items have at least three statistical indicators; nine of those twelve show evidence of linguistic bias, and five of the nine show evidence from an additional category of content bias as well. Three of the four items from the Level C Passage Comprehension subtest fit at least one of the categories of content bias, two of them with multiple indicators. Only four out of nineteen on the Level 2 Vocabulary subtest items with three or more statistical indicators do not have ostensible problems as shown by the content analysis procedure. Of twenty-one items in the Level 2 Passage Comprehension subtest with three or more indicators, only three cannot be corroborated by the analysis of content. None of the items in any subtest

which had no statistical indicators of bias were found to have any content indicators of bias.

Table 5 presents a summary of subtest performance by group when those items for which three or more statistical indicators turn up positive are excluded. In three of the four subtests, the adjusted scores of the Spanish-

-----  
 Insert Table 5 about here  
 -----

language groups move closer to their English-language counterparts. A substantial difference remains, however, between scores for the Passage Comprehension subtest at Level 2. The gain from initial to adjusted group mean by the Spanish-language group is quite insufficient to raise that value to the level of the English-language group. The adjusted minimum p-values achieved by both groups move upward but the English-language group pulls ahead noticeably.

#### DISCUSSION

Five relatively simple analyses have been presented which point to five related considerations in the search for bias. These are (a) overall group differences and their direction, (b) differences in performance by a select subsample of better respondents within groups, (c) differences from chance responding by those subsamples, (d) differences between groups in the selection of wrong answers, and (e) degree of distraction provided by wrong item choices. The first of these follows the well-known Anghoff delta procedure (Anghoff, 1972), without resorting to the arbitrary use of rescaling, which simply serves for added convenience. The second and third analyses make use of the select subsample of putative "masters", those students within each group whose overall performances place them above the P-curve; these



approaches are extensions of the work of Sato (1980) and colleagues. The fourth and fifth procedures examine the bias question by studying those parts of the multiple-choice item which are usually excluded from study in a right-wrong scoring context (cf. Powell & Isbister, 1974).

For purposes of this paper, the five procedures are considered jointly, with equal weights. Interpretations of bias are confirmed in the clear majority of cases where the joint indication of three or more statistics is found for an item. Certain problems remain to be solved, however, and therefore some conditions must be placed on the use of this set of approaches to the detection of item bias. It is clear, for example, that the first index, because it is based on proportion of correct items, is to be used with caution: "proportions of correct answers in a group of examinees is not really a measure of item difficulty. This proportion describes not only the test item but also the group tested" (Lord, 1980, p.35). Indeed, throughout it must be remembered that the results of this study are descriptive of this sample only, and no external criteria are available to evaluate comparability across language groups by grade.

A second objection is that the psychometric properties of the CTBS items are only partially expressed by reliance on p-values and the S-P chart, which at its core relies on the index of item difficulty. Thus, the conclusions drawn from work with that chart are only as good as the strength of the item difficulty metric. In addition, the S-P chart suffers from other metric problems. The first is that the doubly-sorted persons x items matrix treats data, in part, as interval rather than continuous data. Thus, for instance, subtle gradations of difficulty may be given the same credence as larger differences in the case where p-values are nonuniformly distributed.

Analogously, nonlinear distributions of total performance scores may contribute in unknown ways to the use made of ranking information regarding respondents: the patterns may not be as smooth as the chart makes them appear. Moreover, as the S-P chart approaches randomness and its index of discrepancy,  $D^*$ , approaches 1.0, increasingly complex but hidden interactions between the properties of the items in the test and the attributes of the sample are likely. Thus, the second and third statistics in the analytic set depend upon certain assumptions about the nature of performance patterns, violations of which bear rather unclear consequences. Related problems appear in item characteristic curve analysis (Linn, Levine, Hasting, Wardrop, 1980), and in the "adverse impact" approach (Merz, 1980).

A third objection to the procedures used in this study centers on issues of guessing. In the absence of an externally valid explicit criterion, correction for guessing does not seem feasible (Choppin, 1974). Yet assumptions about the occurrence and distribution of guessing affect all aspects of the analysis, particularly statistics which address incorrect responses. Volitional bias, quite likely contributing to the anomalous response by the English-language group to item 29 on the Level C Vocabulary subtest, is nowhere adequately considered. How much of a role guessing plays is not well treated by the assumption that chance responding is represented by  $p = .25$ . In the very likely event that some members of any group will engage in guessing some of the time on some items, only the most general and simplistic conclusions can be drawn from the data presented here. One problem of particular note is the strong possibility that guessing assumes a gradient distribution within the person x items matrix. That is, from the most capable to the least capable person, the contribution of guessing on any item may move from relatively low probability to relatively high probability, thus potentially interfering with diagnosis of problems inherent

in the item. But such diagnosis lies at the heart of the effort to decipher and describe item bias. Until the gradient problem is separated from the bias problem, only partially satisfactory conclusions can be drawn about either.

On the positive side, the high level of match between content analysis and the aggregate of statistical evidence suggests that this simple approach to bias detection may have as much viability as more laborious and unwieldy procedures. The ease of computations and interpretations, and the parsimony of explanation are also favorable points (Merz, 1980). While some attempt is made in the preceding pages to demonstrate the use of multiple indicators, ~~more possibilities can be pursued within this framework.~~ The explanatory power of the five-part procedure appears to exceed that offered by analysis of variance or phi/phi-max, and the assumptions required about the configuration of persons and items are fewer in number than those required by the modified chi-square analyses which recently have been challenged as inadequate (Marascuilo and Slaughter, in press).

Comparison of the present set of results with those of more complex analytic procedures conducted on the same data set awaits further study. However, unlike the results reported by Linn, Levine, Hastings and Wardrop (1981), in which item characteristic curve analyses for a hypothetical data set "...did not lend themselves to making generalizations about features of items..." (p. 38), the findings of the present study suggest at least one concluding observation. Many signals point to a primary conclusion that a number of items in the English-language and Spanish-language versions of the CTBS do not seem to be comparable. Across a spectrum of indicators, the Spanish-language groups regularly produced lower scores. In three of four subtests, removing those items for which three or more statistical indicators pointed to difficulty gave adjusted

scores which were very similar between groups. In the fourth subtest, that correction did not yield significant improvement, suggesting that the Spanish-language sample at grade 6 may be disadvantaged in some respect unrelated to the CTBS itself.

---

Table 1  
Summary of Performance by Subtest by Group

Subtest	Level C				Level 2			
	Vocabulary		Passage Comprehension		Vocabulary		Passage Comprehension	
Group	English	Spanish	English	Spanish	English	Spanish	English	Spanish
n items	33		18		40		45	
N students responding	364	286	363	280	378	231	377	203
$\bar{p}$ value	.6570	.6212	.6254	.5924	.5599	.4302	.5225	.3832
s.d.	.1619	.1775	.0874	.1139	.1473	.1506	.1254	.1022
maximum p	.8571	.8542	.7356	.7128	.8568	.7662	.7507	.6321
minimum p	.1395	.1538	.4826	.4088	.2892	.2078	.2366	.1272
minimum re- quired p greater than chance res- ponding	.2969	.3033	.2970	.3039	.2960	.3096	.2961	.3138
n items less than minimum required p	1	2	0	0	1	11	2	11
index of dis- crepancy $D^*$	.3408	.3568	.2353	.4690	.4416	.4980	.4741	.6288

TABLE 2

Percentage of Items Exceeding  
Critical Minimums in Five Analyses

Subtest	Level C		Level 2	
	Vocabulary	Passage Comprehension	Vocabulary	Passage Comprehension
<u>Analysis</u>				
a) Test of proportions of correct scores				
English significantly higher	45%	50%	55%	76%
Spanish significantly higher	18%	0%	8%	0%
b) Test of proportions of correct scores for "masters"				
English significantly higher	33%	44%	40%	60%
Spanish significantly higher	22%	0%	5%	0%
c) Test of chance responding by "masters"				
in English	0%	0%	3%	7%
in Spanish	0%	0%	3%	16%
d) Test of differential attractiveness of wrong answers between groups	36%	50%	43%	27%
e) Test of popular distractors				
in English	9%	11%	13%	29%
in Spanish	30%	17%	30%	24%
Overlap between groups	6%	0%	10%	13%

TABLE 3

Percent of Items Showing Statistical  
Indicators of Differential Performance

Subtest	Level C		Level 2	
	Vocabulary	Passage Comprehension	Vocabulary	Passage Comprehension
No indicators	9%	6%	23%	4%
One indicator	33%	39%	18%	20%
Two indicators	21%	33%	18%	40%
Three indicators	27%	22%	33%	34%
Four indicators	6%	0%	8%	2%
Five indicators	3%	0%	0%	0%

Sources of Content Bias for Items with Three or More Statistical Indicators of Differential Performance, by Subtest

Key: a) test of proportions  
 b) test of proportions of correct scores for "masters"  
 c) test of chance responding by "masters"  
 d) test of differential attractiveness of wrong answers  
 e) test of popular distractors

1) mistranslation  
 2) cultural difference  
 3) linguistic difference  
 4) low frequency word or phrase  
 5) unfamiliar context for word or phrase

Level C Vocabulary		Level C Passage Comprehension		Level 2 Vocabulary		Level 2 Passage Comprehension	
item	2 a b e;	4 item 1 a b d ;	item 1 a b d ;	item 1 a b d ;	item 1 a b e; 1	2 a b e; 1	2 a b e; 1
6 a b d ;	3	4 a b e; 1	3	6 a b e; 1	3	2 a b e; -	2 a b e; -
7 a d e;	2 3	6 a d e; 2	2	8 a b d ;	-	3 a b d e; 1	3 a b d e; 1
12 a b e;	3	7 a b d ; 2 3 4	2 3 4	9 a b e;	3 4	7 a b d ;	7 a b d ;
14 a b e;	4			11 a b e;	1 2	9 a c e;	9 a c e;
15 a b c e; 1	3			12 a b d e; 1	1	15 a d e;	15 a d e;
16 a b e;	3 5			13 a b e;	1 2 3	17 a b e;	17 a b e;
20 b d e;	2 3			15 a b e;	2	18 a b c e;	18 a b c e;
23 a d e;	2 4			19 a b c e; 1	3	21 a b d ;	21 a b d ;
29 a b c d e;	3			20 a b d e;	-	22 a b d ;	22 a b d ;
30 a b d e;	2 3			23 a d e;	2	24 a b c d ;	24 a b c d ;
32 a b e;	3			25 a b c d e; 1 2	1 2	25 a b c d ;	25 a b c d ;
				26 a b c e;	-	28 a b c e; 1	28 a b c e; 1
				32 c d e;	1	29 a c e; 1	29 a c e; 1
				34 a c d ;	-	34 a d e;	34 a d e;
				35 a b c e;	4	36 a b c ;	36 a b c ;
				36 b c e;	3	37 b c d ;	37 b c d ;
				39 a b c d ;	3	38 a b c e;	38 a b c e;
				40 a b d ;	3	39 a b c e;	39 a b c e;
						41 a b d ;	41 a b d ;
						45 a c d ;	45 a c d ;



TABLE 5

Revised Summary of Performance by Subtest Group, Deleting  
Items with Three or More Statistical Indicators

Subtest	Level C				Level 2			
	Vocabulary		Passage Comprehension		Vocabulary		Passage Comprehension	
Group	<u>English</u>	<u>Spanish</u>	<u>English</u>	<u>Spanish</u>	<u>English</u>	<u>Spanish</u>	<u>English</u>	<u>Spanish</u>
adjusted n items	21		14		21		24	
adjusted mean	.6804	.6606	.6216	.6061	.5818	.5322	.5431	.4067
change from original	.0234	.0394	-.0038	.0137	.0219	.1020	.1230	.0969
adjusted s.d.	.1298	.1502	.0936	.1039	.1418	.1476	.0206	.0235
adjusted maximum	.8571	.8542	.7356	.7128	.8568	.7662	.7507	.5707
adjusted minimum	.4104	.3004	.4826	.4343	.3344	.3005	.2366	.1272

## Figure Captions

Figures 1A and 1B: 1A) Perfect Guttman scale for a hypothetical ten-item test scored right (1) and wrong (0). Persons and items are uniformly ordered, by total correct score and level of difficulty, respectively. 1B) Perfect Guttman scale, showing uniform ordering with lower overall performance.

Figures 2A and 2B: 2A) Hypothetical score matrix for a ten-item test sorted by respondents on descending total score and by items on ascending level of difficulty. S- and P-curves reflect cumulative ogives of performance, and lead to an appraisal of the characteristic performance of the group. 1B) Hypothetical score matrix for the same test with a different group, again sorted by respondents and items.

Figure 3: Hypothetical patterns of response to two items by ten persons, showing a poorly-fitted and a better-fitted item.

1A)

	Items	1	2	3	4	5	6	7	8	9	10	Total score
Persons	A	1	1	1	1	1	1	1	1	1	1	10
	B	1	1	1	1	1	1	1	1	1	0	9
	C	1	1	1	1	1	1	1	1	0	0	8
	D	1	1	1	1	1	1	1	0	0	0	7
	E	1	1	1	1	1	1	0	0	0	0	6
	F	1	1	1	1	1	0	0	0	0	0	5
	G	1	1	1	1	0	0	0	0	0	0	4
	H	1	1	1	0	0	0	0	0	0	0	3
	I	1	1	0	0	0	0	0	0	0	0	2
	J	1	0	0	0	0	0	0	0	0	0	1
% correct		100	90	80	70	60	50	40	30	20	10	

$\bar{p} = .5500$   
s.d. = .3028

1B)

	Items	1	2	3	4	5	6	7	8	9	10	Total score
Persons	K	1	1	1	1	1	1	1	0	0	0	7
	L	1	1	1	1	1	1	0	0	0	0	6
	M	1	1	1	1	1	0	0	0	0	0	5
	N	1	1	1	1	0	0	0	0	0	0	4
	O	1	1	1	0	0	0	0	0	0	0	3
	P	1	1	0	0	0	0	0	0	0	0	2
	Q	1	0	0	0	0	0	0	0	0	0	1
	R	0	0	0	0	0	0	0	0	0	0	0
	S	0	0	0	0	0	0	0	0	0	0	0
	T	0	0	0	0	0	0	0	0	0	0	0
% correct		70	60	50	40	30	20	10	0	0	0	0

$\bar{p} = .2800$   
s.d. = .2616

2A)

Items	2	4	1	5	3	9	10	6	8	7	
Persons E	1	1	1	0	1	1	0	1	1	1	8
A	1	1	1	1	1	1	0	0	1	0	7
G	1	1	1	1	1	1	1	0	0	0	7
C	1	1	1	1	0	0	0	1	0	0	5
F	1	1	1	1	1	0	0	0	0	0	5
B	1	1	1	0	1	0	0	0	0	0	4
J	1	1	1	0	0	0	1	0	0	0	4
D	1	1	1	0	0	0	0	0	0	0	3
H	1	0	0	1	0	0	0	0	0	0	2
I	1	0	0	0	0	0	0	0	0	0	1
p-value	1.0	.8	.8	.5	.5	.3	.2	.2	.2	.1	

S-curve  
P-curve

2B)

Items	2	1	4	3	5	9	10	6	7	8	P-curve	Total score
Persons M	1	1	1	1	1	1	0	1	0	0		7
K	1	1	1	1	1	0	1	0	1	0		7
P	0	1	1	1	0	1	1	0	0	0		5
L	1	1	0	1	1	0	0	0	0	0		4
N	1	1	1	1	0	0	0	0	0	0		4
O	1	1	1	0	1	0	0	0	0	0		4
S	1	1	1	0	0	0	0	0	0	0		3
T	1	0	1	0	0	0	0	0	0	0		2
R	1	0	0	1	0	0	0	0	0	0		2
Q	1	0	0	0	0	0	0	0	0	0		1
p-value	.9	.7	.7	.6	.4	.2	.2	.1	.1	.0		

S-curve  
P-curve

3)

		<u>Poorly-fitted item</u>	<u>Better-fitted item</u>
Persons	U	0	1
	V	0	1
	W	1	1
	X	0	1
	Y	0	0
	Z	0	0
<hr/>			
	a	0	0
	b	1	1
	c	0	0

-----P-curve  
crosses here

-----P-curve  
crosses here

References

- Anghoff, W.H. A technique for the investigation of cultural differences. Paper presented at the Annual Meeting of the American Psychological Association, Honolulu, 1972.
- Berk, R.A. (Ed.) Handbook of methods for detecting test bias. Baltimore: Johns Hopkins University Press, in press.
- Choppin, B.H. The correction for guessing on objective tests. Stockholm, International Association for the Evaluation of Educational Achievement, 1974.
- Crowder, C.R. An investigation of item bias occurring at different ability levels for Anglo and Mexican-American students. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1979.
- 
- Comprehensive Tests of Basic Skills (CTBS) Examiner's Manual and Spanish Examiner's Manual. Monterey: CTB/McGraw-Hills, 1974/78.
- Ebel, R. L. Constructing unbiased achievement tests. Paper presented at the National Institute of Education Conference on test bias, Baltimore, 1975.
- Harnisch, D.L., & Linn, R.I. Analysis of item response patterns: Consistency indices and their application to criterion-referenced tests. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, 1981.
- Hayes, W.L. Statistics. New York: Holt, Reinhart and Winston, 1963.
- Hudson, L. The relation of psychological test scores to academic bias. British Journal of Educational Psychology, 1963, 33, 120-131.
- Hunter, J.E. A critical analysis of the use of item means and item test correlations to determine the presence or absence of content bias in achievement test items. Paper presented at the National Institute of Education Conference on test bias. Baltimore, 1975.
- Jensen, A.R. Bias in mental testing. New York: Free Press, 1980.
- Linn, R.L., Levine, M.V., Hastings, C.N., & Wardrop, J.L. Item bias in a test of reading comprehension. Applied Psychological Measurement, 1981, 5, 159-173.
- Lord, F.M. Applications of item response theory to practical testing problems. Hillsdale, New Jersey: Lawrence Erlbaum, 1980.
- Marascuilo, L.A., & Slaughter, R.E. Statistical procedures for analyzing item bias based on chi-square statistics. Journal of Educational Measurement, in press, 1981.

- Merz, W.R. Methods of assessing bias and fairness in tests. ARC Technical Report #121-79, Sacramento, Applied Research Consultants, 1980. (ERIC Document Reproduction Service No. ED 198 145).
- Petersen, N.S. Bias in the selection rule, bias in the test. In van der Kamp, L.J.T., Langerak, W.F., & deGruiter, D.N.M. (Eds.). Psychometrics for educational debates. Chichester, G.B.: John Wiley and Sons, 1980.
- Powell, J.C., & Isbister, A.G. A comparison between right and wrong answers on a multiple choice test. Educational and Psychological Measurement, 1974, 34, 499-509.
- Rüdner, L.M., Geston, P.R., & Knight, D.L. Biased item detection techniques. Journal of Educational Statistics, 1980, 5, 213-233.
- Sandoval, J. The WISC-R and internal evidence of test bias with minority groups. Journal of Consulting and Clinical Psychology, 1979, 47, 919-927.
- 
- Sato, T. The S-P chart and the caution index. NEC (Nippon Electric Company, Japan), Educational Informatics Bulletin, 1980.
- Sato, T., Takeya, M., Kurata, M., Morimoto, Y., & Chimura, H. An instructional data analysis machine with a microprocessor--SPEEDY. NEC (Nippon Electric Company, Japan) Research and Development, 1981, No. 61, 55-63
- Subkoviak, M.J., Mack, J.S., & Ironson, G.H. Item bias detection procedures: empirical validation. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, 1981.
- Tatsuoka, K. An approach to assessing the seriousness of error types. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, 1981.
- Veale, J.R., & Foreman, D.I. Cultural variation in criterion referenced tests: A global item analysis. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1976.
- Williams, R.L. Abuses and misuses in testing black children. Counseling Psychologist, 1971, 2, 62-77