Item Bias in a Test of Reading Comprehension

Robert L. Linn, Michael V. Levine, C. Nicholas Hastings, and James L. Wardrop
University of Illinois at Champaign/Urbana

The possibility that certain features of items on a reading comprehension test may lead to biased estimates of the reading achievement of particular subgroups of students was investigated. Eight nonoverlapping subgroups of students were defined by the combinations of three factors: student grade level (fifth or sixth), income level of the neighborhood in which the school was located (low and middle or above), and race of the student (black or white). Estimates of student ability and item parameters were obtained separately for each of the eight subgroups using the three-parameter logistic model. Bias indices were computed based on differences in item characteristic curves for pairs of subgroups. A criterion for labeling an item as biased was developed using the distribution of bias indices for subgroups of the same race that differed only in income level or grade level. Using this criterion, three items were consistently identified as biased in four independent comparisons of subgroups of black and white students. Comparisons of content and format characteristics of items that were identified as biased with those that were not, or between items biased in different directions, did not lead to the identification of any systematic content differences. The study did provide strong support for the viability of the estimation procedure; item characteristics, estimated with samples from different populations were very similar. Some suggestions for improvements in methodology are offered.

A common use of tests is to predict some future behavior, such as job performance or success in school or college. For the predictive use of tests, the issue of possible bias revolves around the question of whether or not identifiable subgroups perform better on the job or in college than would be predicted from their test scores (Anastasi, 1976; Cleary, 1968; Linn, 1973; Petersen & Novick, 1976).

Prediction is one of the uses made of achievement tests, but it is by no means the only use. More often, achievement tests are used to assess current status, to evaluate programs, and to diagnose problems. For the nonpredictive uses of achievement tests, strategies for assessing possible sources of bias have generally focused on the internal characteristics of the test. The goal is to identify the nonessential characteristics of test items resulting in the misinterpretation of the achievement of certain groups of students. For example, reading is a skill that is incidental to the one that is purported to be measured by a mathematics achievement test. Dependence of the test results on reading ability could lead to a biased indication of the relative competence in mathematics for two groups that differ in reading ability.

If items on a test differ in their dependence on the characteristic that is incidental to the skill being assessed, then the biasing effects of that incidental characteristic would be expected to result in an interaction between the items and the characteristics of the examinees. In other

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 5, No. 2, Spring 1981, pp. 159–173

0146-6216/81/020159-15$1.75

words, the magnitude of group differences in performance would be expected to vary as a function of the extent to which items were dependent on the incidental characteristics.

The idea of searching for item characteristics that interact with group membership in order to reduce possible bias is not new. For example, the stated purpose of the landmark study by Eells, Davis, Havighurst, Herrick, and Tyler (1951) was to "identify (a) those kinds of test problems on which children from high socioeconomic backgrounds show the greatest superiority and (b) those kinds on which children from low socioeconomic backgrounds do relatively well" (p. 6). Interactions between item content and gender were investigated by Coffman (1961), and a number of studies have been conducted to identify types of items that are unusually difficult for members of minority groups (e.g., Angoff & Ford, 1973; Cleary & Hilton, 1968).

One of the limitations of the early studies of item-group interactions is that they relied upon sample-dependent item statistics. There is no sound theoretical basis for expecting a constant difference in the proportion of people in two groups that respond correctly to various items.

A second limitation of definitions of item bias that depend on differences in the proportion correct for two groups is that proportion correct is confounded with other item characteristics such as item discriminating power (Hunter, 1975). The difference in proportion correct for two groups can be expected to vary from item to item solely as a function of differences in the discriminating power of the items; and conversely, proportion correct can be exactly the same for two groups in the presence of extreme bias. Thus, as stated by Warm (1978), "the use of classical test theory item parameters is inappropriate for and can lead to erroneous identification of item bias" (p. 128).

Lord (1977a, 1977b, 1980, chap. 14), Wright (1977), and others have suggested that latent trait theory provides a theoretically sounder approach to the problem of identifying items that interact with group membership than can be achieved using item statistics based on classical test theory. Several recent studies (e.g., Harms, 1978; Ironson & Subkoviak, 1979; Rudner, 1977; Shepard, Camilli, Averill, 1980) have compared indices of item bias based on latent trait theory with indices from several earlier approaches. It is clear that the earlier approaches, based on statistics used in classical test theory, are not substitutes for an approach based on latent trait theory.

The primary advantage of an approach based on latent trait theory is that to the extent that the model holds, the item parameters will be invariant. That is, they should not depend upon the sample of people on which the estimates are based. Thus, after an appropriate linear transformation to adjust for arbitrary differences in scale, the estimates for different groups would be expected to be the same, except for sampling error, even though the groups may differ substantially in ability level.

This study has two major purposes, one of which is methodological in nature and the other, substantive. Refinements are needed in the techniques used to detect items that lead to biased estimates of the ability of a particular group. The analyses conducted for this study were intended to provide some evaluation of an approach based upon a particular latent trait model and to contribute to the development of better methods of using latent trait models for detecting items that result in biased ability estimates.

The substantive purpose of this study, as originally conceived, was to investigate the possibility that certain features of items on a reading comprehension test may lead to biased estimates of the reading achievement level for black students as compared to white students and/or for children attending schools in low-income neighborhoods as compared to those attending schools in middle- or high-income neighborhoods. The identification of items that lead to such estimates would be of particular value if the items so identified could be characterized by some generalizable features that could be used

as a guide in constructing and editing reading comprehension tests to minimize bias against particular subgroups of students.

Method

Strategy for Identifying Bias

Birnbaum's (1968) three-parameter logistic model was used to obtain estimates of ability and item parameters in all of the analyses reported below. The LOGIST computer program (Wood, Wingersky & Lord, 1976) was used to estimate the item parameters and abilities of the students.

According to the three-parameter logistic model, the conditional probability $P_i(\theta)$ that a person randomly chosen from all those with ability θ will answer item i correctly is a function of θ and three item parameters. Each item is characterized by three parameters: the item discrimination, a; the location or difficulty of the item, b; and the lower asymptote or probability that persons with extremely low ability will respond correctly to the item, c. The graph of $P_i(\theta)$ as a function of θ is called the item characteristic curve (ICC) for item i. According to the model, the probability of getting the item right is completely determined by θ and the three item parameters. More specifically, members of different groups with equal ability (i.e., equal θ) should have the same probability of correctly answering an item. In other words, the conditional probabilities, $P_i(\theta)$, and their graphs should be invariant from one group to another.

Comparisons among ICCs estimated separately for different subgroups were made in order to identify items that functioned differently for members of different subgroups. If the ICCs of some items differed from group to group more than would be expected due to sampling error, then such items may be considered biased: The probability of correctly answering an item is not equal for persons of the same overall ability who come from different subgroups.

Such bias may be the consequence of multidimensionality. That is, the probability of correctly answering an item depends on more than one latent trait, and the groups differ in their distributions of the secondary latent traits (Hunter, 1975). Multidimensionality may still be considered a form of bias, however, in that it can lead to apparent differences in the primary ability when, in fact, there are no such differences. Indeed, bias may generally be conceptualized as multidimensionality confounding differences on a primary trait with differences on a secondary trait.

Data

Data for the analyses reported below were obtained from the Anchor Test Study (Bianchini & Loret, 1974) equating study files. Item response data on the Reading Comprehension section of Form F of the *Metropolitan Achievement Tests* (Durost, Bixler, Wrightstone, Prescott, & Balow, 1970) were obtained for students in Grades 5 and 6. Data were available for a total of 15,485 fifth-grade and 14,843 sixth-grade students. At each grade level, slightly over 16% of the students with available data were black and somewhat over 76% were white. All analyses reported below are based on samples from these two groups of students within each grade.

The sample of students was divided into eight subgroups. The subgroups were defined by grade (fifth or sixth), by race (black or white), and by income level of the neighborhood in which the sample school was located (low and middle or high). The analyses were based on all black students for whom the necessary item response data were available. Analyses for white students were based on spaced samples containing roughly the same number of students as were in the black student samples attending low income schools. Listed in Table 1 is the number of students within each subgroup upon which the parameters of the ICCs were estimated. As can be seen, group size was roughly 2,000 per subgroup for all but the subgroup of black students attending schools in middle- or high-income neighborhoods. The latter was considerably

Table 1
Number of Students Within Each Subgroup
Used to Estimate Parameters of
Item Characteristic Curves

Income	Blacks	Whites
	Grade 5	
Low	2024	2109
Middle or High	463	2111
	Grade 6	
Low	1907	2028
Middle or High	444	2137

smaller, containing approximately 22% as many students, on the average, as the other three subgroups at each grade level.

Scale Equating

Under the assumption that the three-parameter logistic model holds for all subgroups, the estimated abilities should be on essentially the same scale, regardless of the group used to obtain the estimates. The assumption implies that the different subgroups can differ only by a linear transformation from one subgroup to another. Thus, it is possible to equate the scales by means of a linear transformation and then make meaningful comparisons of the ICCs for different subgroups.

The specific steps followed to equate the scales of two groups were as follows. First, one group was arbitrarily identified as the base group and the other as the comparison group. The scale of the base group was left unchanged (i.e., no transformation was made of the θ's or item parameters for the base group). Two constants, A and B, were then found such that the weighted mean and variance of the transformed b's of the comparison group were equal to the weighted mean and variance of the base group. More specifically, if b_i^* is the item difficulty of item i in the comparison group after equating and b_i is the corresponding value prior to equating, then

$$b_i^* = A + Bb_i \qquad [1]$$

where A and B are the equating constants selected such that the weighted mean and variance of b_i^* in the comparison group are equal to the weighted mean and variance of the original b's in the base group. The i^{th} weight was the inverse of the larger of the estimated variances of the b_i computed from the comparison group and the b_i computed from the base group. Thus, items for which the difficulty parameter was poorly estimated (i.e., had a large estimated sampling variance) for either of the groups were given relatively less weight in determining the equating constants than were items for which the difficulty parameter was better estimated. Detailed formulas used in estimating the variances and covariances of the errors of estimate for the item parameters and for approximating the standard error of a point on an estimated ICC are provided by Linn, Levine, Hastings, and Wardrop (1980).

Once the A and B in Equation 1 were obtained, the comparison group ability estimates and estimates of item discrimination were converted to the base group scale. In particular, the transformed θ scale, say θ^*, for the comparison group is given by

$$\theta^* = A + B\theta \qquad [2]$$

and the transformed a's, say a_i^*, by

$$a^*_i = a_i/B \qquad [3]$$

No transformation of the c parameter estimates is required.

After the estimated abilities and item parameters of a comparison group were transformed to the base group scale, several types of comparisons were made. ICCs for each group were plotted on the common scale and compared. In order to better evaluate whether observed differences in ICCs were attributable simply to sampling error, the standard errors of the ICCs were estimated; and ICCs plus and minus two standard errors of estimate were obtained and plotted for each group.

Indices of Bias

In addition to the comparison of the ICCs and the confidence bands determined by their standard errors, several indices of item bias were computed. Three of these indices were described by Ironson (1978) and Ironson and Subkoviak (1979). They involve areas between the ICCs of a base group and a comparison group. Sums of squared differences between ICCs were also computed.

The four bias indices used for the results reported below are as follows.

1. Absolute Difference: the area enclosed by two ICCs and the two vertical lines, $\theta = -3$ and $\theta = +3$.
2. Base High Area: the area, if any, between two ICCs where the ICC for the base group is above that of the ICC for the comparison group. As before, only θ's between +3 and −3 were considered.
3. Base Low Area: the first index minus the second. In other words, the area, if any, between the two ICCs where the comparison group ICC is higher.
4. Square Root of the Sum of Squares: the square root of the sum of the squared differences between ICCs in the region of $\theta = -3$ to $\theta = +3$.

An item with a large base high area (Index 1) but small or zero base low area (Index 2) would be considered to be biased against the comparison group. Such an outcome would indicate that persons in the comparison group have a smaller probability of correctly answering the item than persons in the base group with equal estimated ability. The direction of the bias would be just the opposite for an item with a large base low area but zero or small base high area. The bias in an item with large base high and large base low areas would depend upon the distribution of ability in the groups of examinees contrasted.

Estimates of item parameters that were obtained separately for the eight subgroups defined by grade level, race, and income level of the school neighborhood were used to make a total of 12 pairwise comparisons. In each pairwise comparison, the base group and the comparison group differed in only one of the three characteristics used to define the subgroups. Thus, there were four independent comparisons of the different levels of each group characteristic with constant levels of the other two group characteristics. For example, for a fixed grade level and income level of the school neighborhood (income), comparisons were made across racial groups (race), so that four comparisons were made. Similarly, income level comparisons were made for each of four race-by-grade combinations, and grade comparisons were made for each of four race-by-income combinations. The base group and comparison group in each of the 12 comparisons are listed in Table 2.

Results

A general indication of the comparability of the ICCs for the 12 pairwise comparisons is provided by the distributions of the square root of the sum of squares bias indices. When an item has very similar ICCs for two groups, the index should be near zero. Distributions of the bias index values for the 45 items are shown for all 12 pairwise comparisons in Figure 1. The top four distributions provide comparisons of Grade 5 with Grade 6, holding race and income constant. The middle four distributions provide income level comparisons holding grade and race constant, and the bottom four distributions show the results of the racial group comparisons with

Table 2
Base Group and Comparison Group
In Each of the Twelve Pairwise Comparisons

Base Group	Comparison Group
Grade Level Comparisons	
LW5: Low income, white, grade 5	LW6: Low income, white, grade 6
LB5: Low income, black, grade 5	LB6: Low income, black, grade 6
MW5: Middle income, white, grade 5	MW6: Middle income, white, grade 6
MB5: Middle income, black, grade 5	MB6: Middle income, black, grade 6
Income Comparisons	
LW5: Low income, white, grade 5	MW5: Middle income, white, grade 5
LB5: Low income, black, grade 5	MB5: Middle income, black, grade 5
LW6: Low income, white, grade 6	MW6: Middle income, white, grade 6
LB6: Low income, black, grade 6	MB6: Middle income, black, grade 6
Racial Comparisons	
LW5: Low income, white, grade 5	LB5: Low income, black, grade 5
MW5: Middle income, white, grade 5	MB5: Middle income, black, grade 5
LW6: Low income, white, grade 6	LB6: Low income, black, grade 6
MW6: Middle income, white, grade 6	MB6: Middle income, black, grade 6

grade and income held constant. The group characteristics that are held constant for a given distribution are identified by the letters and numbers above each histogram. For example, the left-hand histogram in the first row of Figure 1 is the grade comparison for white students attending schools in low-income neighborhoods and is denoted LW. Another example is the M6 over the right-hand histogram in the bottom row of Figure 1. M6 denotes that the racial comparison in the lower right-hand histogram is for sixth-grade students attending schools in middle- or high-income neighborhoods.

An immediate observation that can be made from an inspection of Figure 1 is that there are fewer large values of the bias index for the four comparisons involving only white students than for any of the other comparisons, that is, the comparison of ICCs across grade for white students (the two left-hand distributions in the top row of Figure 1) or across income level for white students (two left-hand distributions in the middle row of Figure 1). Only one of the 180 bias indices is as large as .2 for these four distributions. None is as large as .3.

Items with indices less than .2 have quite similar ICCs. Some indication of the degree of similarity is provided by the plots shown in Figure 2 for two items. The plots in Figure 2 compare the ICCs for fifth-grade white students attending schools in low-income neighborhoods (LW5) with their sixth-grade counterparts (LW6). Item 6 (Figure 2a) had the second largest index (square root of the sum of squares bias index equals .161) of any of the 45 items. The index for Item 18 was .070, which is closer to the mean of .076 for the 45 items.

The three solid lines show the ICC, the ICC plus two standard errors of estimate, and the ICC minus two standard errors of estimate for LW5 students; the three dashed lines show the corresponding curves for the LW6 students. The ICCs in Figure 2 are strikingly similar. This provides rather strong support for the claim of invariance. Even the item with the largest sum of squares bias index has ICCs with confidence in-

Figure 1
Distributions of the Square Root of the Sum of Squares Bias Indices for the Twelve Pairwise Comparisons

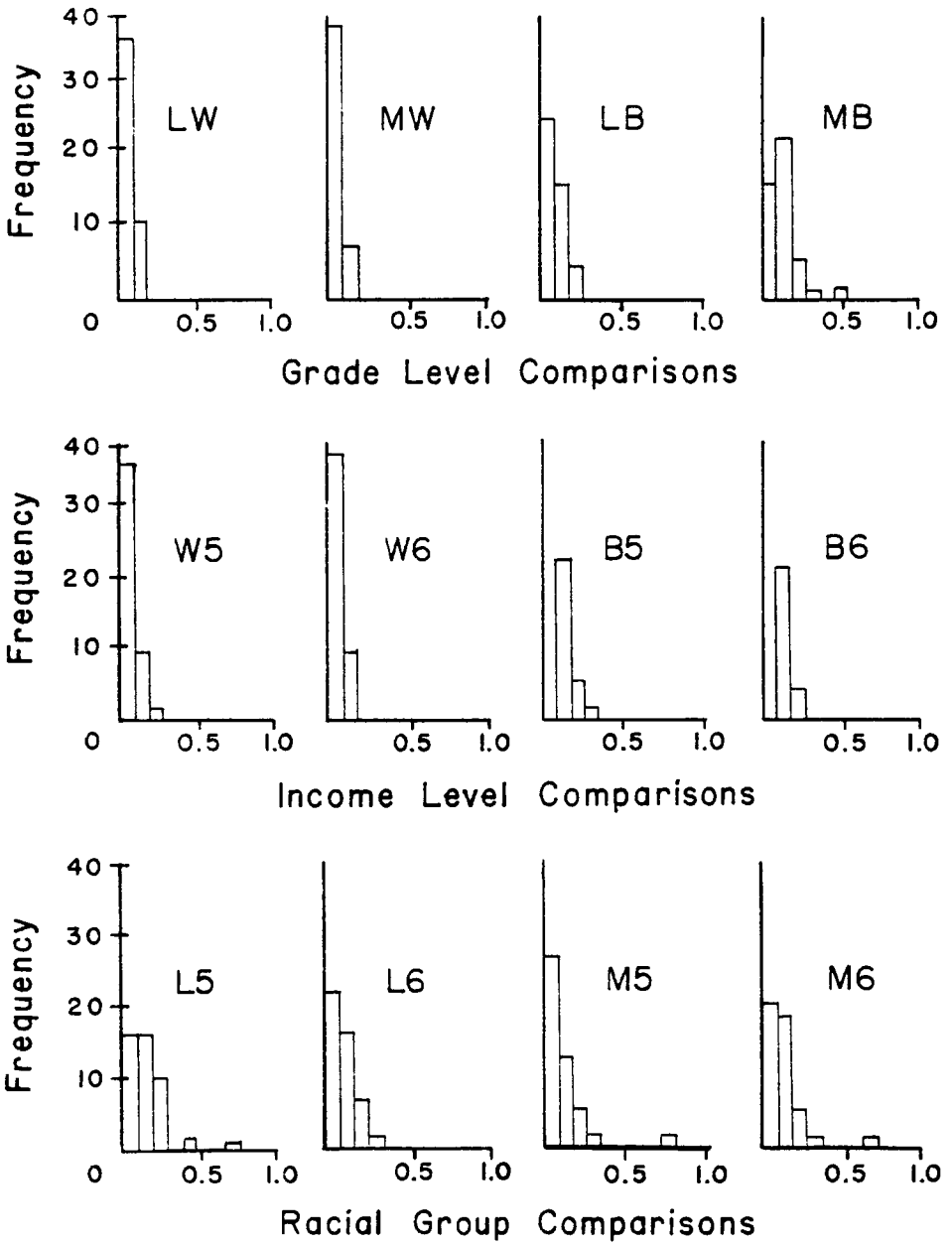
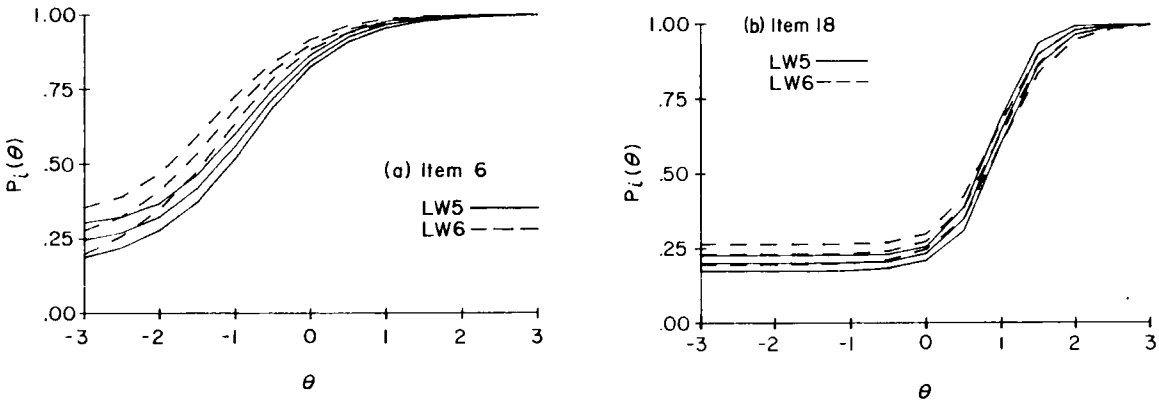


Figure 2
Item Characteristic Curves and Confidence Intervals
for Fifth- and Sixth-Grade White Students
Attending Schools in Low-Income Neighborhoods (Items 6 and 18)



tervals that overlap substantially throughout most of the range of ability. This evidence of invariance of the parameters over grade and income level for white students strengthens the case for using ICC comparisons to identify items that result in biased estimates for particular subgroups. The distributions of indices for the four pairwise comparisons of white subsamples also provide a base rate against which the indices for other pairwise comparisons can be evaluated.

Returning to Figure 1, it can be seen that the black subsamples provide less evidence of invariance across either grade or income level. Comparisons involving middle-income black subsamples might be expected to show less invariance because the estimates are all less stable due to the smaller sample sizes. The comparison of black fifth-graders attending schools in low-income neighborhoods (LB5) with black sixth-graders attending schools in low-income neighborhoods (LB6), however, involves samples sizes comparable to the white subgroup comparisons. Yet, four of the items had indices of .2 or larger for the LB5 versus LB6 comparison.

The comparisons of primary interest in Figure 1 are, of course, those between white and black subgroups of students, since it is there that the presence of biased items is most suspected. The last row of Figure 1 shows the distributions of the square root of the sum of squares bias index for the four pairwise comparisons between subgroups of white students and subgroups of black students. Large indices are clearly observed with greater frequency in the four comparisons in the last row of Figure 1 than in the across-grade or income level comparisons for white students. Only occasionally are the indices for the racial group comparisons more extreme than they are for the within-race comparisons for black students.

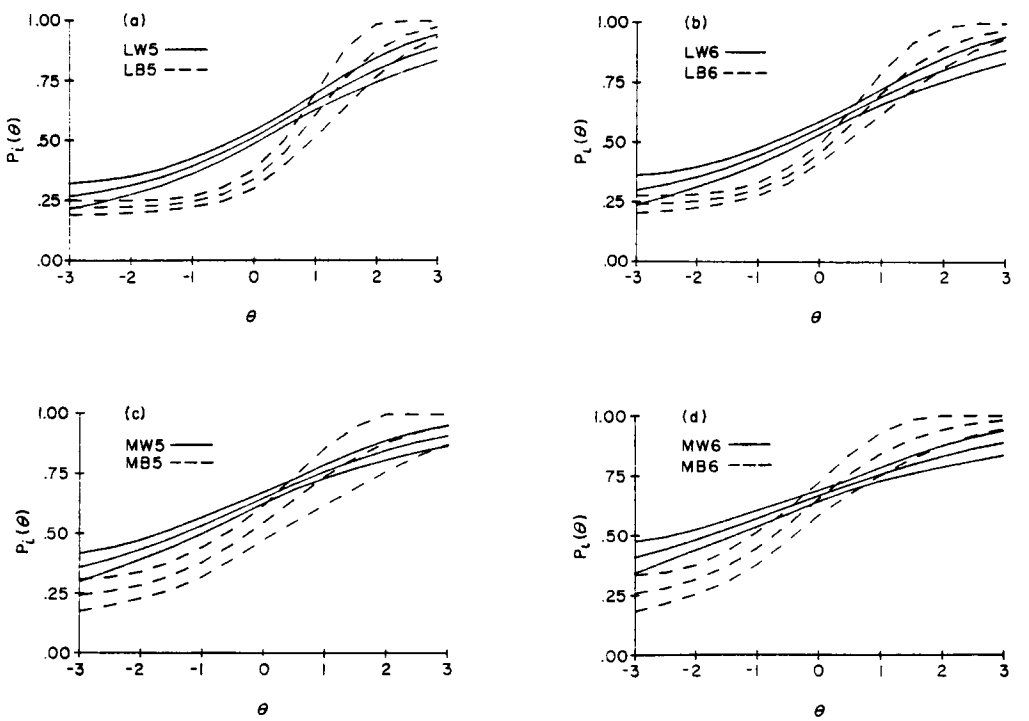
Using a cutoff of .2 to indicate a possibly biased item, 13 of the 45 items in the LW5-LB5 comparison and 7 items in each of the other three comparisons between racial groups would be so identified. The number of items identified as possibly biased obviously depends on the stringency of the criterion employed. The ICCs corresponding to the largest indices, however, are markedly different.

The agreement among the four independent between-race comparisons regarding the identification of items as possibly biased is far from perfect. On the other hand, the agreement is considerably better than would be expected if items were randomly identified by the four inde-

pendent comparisons. Using the above criterion, three of the items were identified in all four pairwise comparisons. If an equal number of items had been selected at random in each comparison, the probability that an item would be selected all four times is only .00109. Thus, the expected number of items that would be identified four times by a random process is only about .05 (i.e., $45 \times .00109$). A chi-square statistic with 2 degrees of freedom was computed to test the goodness of fit of the observed occurrences of 0, 1, and 2 or more identifications of an item to that expected if the four comparisons were independent. The resulting chi-square was 12.13, which is significant at the .01 level. The agreement is clearly better than would be expected on the basis of chance.

The ICCs for one of the items that was identified as possibly biased in all four comparisons using the square root of the sum of squares bias index are shown in Figure 3. Four pairs of ICCs plus and minus two standard errors of estimate

Figure 3
Item Characteristic Curves and Confidence Intervals
for Four Independent Racial Group Comparisons (Item 3)



are shown in Figure 3. The solid lines are the ICCs plus and minus two standard errors for the white sample and the dashed lines are the comparable figures for the black sample at the same grade and income level.

From an inspection of Figure 3, it is apparent that the four independent comparisons show a great deal of consistency. In each comparison, the ICC for the white students is above that of the black students for low and mid-range values of θ. Item 3 is less discriminating (smaller value for a) for white students than for black students in each of the comparisons, however. The ICCs thus cross and the ICC for black students is

above the ICC for white students at high values of θ. Although the direction of bias depends on the value of θ, Item 3 is generally biased against black students in the region where the majority of the black student sample falls (i.e., below a value of θ equal to the mean of the white student sample). If more items with ICCs similar to those of Item 3 were added to the test, the test performance of most black students would appear worse than it currently does in comparison to white students. On the other hand, elimination of Item 3 would tend to improve the relative standing of black students.

Figures for the other two items (Items 25 and 31) that were identified as possibly biased in all four comparisons are not presented here due to space limitations but can be found in Linn, Levine, Hastings, and Wardrop (1980). The large bias indices for Item 25 are brought about largely by its very poor discriminating power for black students. Item 25 is a difficult item for all subgroups. It discriminates well among high-ability sixth-grade white students. The discrimination of Item 25 for high-ability black students, however, is problematic. Consequently, the probability of correctly answering Item 25 is less for the blacks with high θ's than for their white counterparts. On the other hand, the probability of correctly answering the item is slightly higher for blacks with near average θ's than for whites with equal θ values.

The pairs of ICCs for Item 31 are quite similar for low values of θ; but for higher values of θ, the curve for black students is above the one for whites in all four of the comparisons. Thus, Item 31 would be considered biased in favor of black students relative to other items on the test. Inclusion of more items such as Item 31 would tend to improve the relative standing of black students on the test.

The contrasts that are found between groups for Items 3, 25, and 31 may be summarized by the four bias indices computed for each of the contrasts. In order to facilitate comparisons, the indices for the 45 items were first rank-ordered with a rank of 1 given to the item with the highest value of a particular index for a given contrast. The rank ordering was obtained separately for each index and each contrast. The rank order of the bias indices for Items 3, 25, and 31 are listed in Table 3.

Item 25 has relatively large base high bias indices in all four of the independent racial group comparisons. Indeed, in three of the four comparisons, Item 25 has the largest or second largest base high bias index. The white sample was used as the base group and the black sample as the comparison group in all four racial group comparisons. Thus, a large value of a base high bias index implies that the ICC for white students tends to be above the ICC for black students. The large base high bias indices for Item 25 accurately reflect the fact that the ICC for white students is generally above the one for blacks. The relatively smaller, but nonzero, base low bias indices for Item 25 reflect the fact that the ICCs cross in all four comparisons. Item 31, on the other hand, has either the largest or second largest base low bias indices but relatively small base high bias indices in each of the comparisons.

Item 3 has base high and base low bias indices that generally rank among the highest third of the items. Thus, the relatively large overall indices reflect a combination of moderately large base high and base low differences due to the crossover of the ICCs in all four comparisons (see Figure 3).

Items 25 and 31 are probably the two most clearly contrasting items in terms of the racial group differences in ICCs. Item 25 was consistently identified as biased against black students, and Item 31 was consistently identified as biased in favor of black students. The items are of quite different types. Item 25 asks the meaning of the word "character" as it is used in one of the reading passages on the test. Item 31, on the other hand, asks for the "best title" of a story about a fictional baron presented in another passage.

There were 11 items that asked the meaning of a word as used in a passage and 5 items that asked the best title of a story. The rank order of

Table 3
Rank Order of Bias Indices for the Three Items Identified as Possibly Biased in All Four Comparisons

Item	Comparison	Index: Base-High Area	Base-Low Area	Absolute Difference	Root Sum of Squares
3	L5	3	15	2	6
	L6	3	10	2	4
	M5	3	15	4	4
	M6	4	4	2	3
25	L5	9	20	9	10
	L6	1	11	1	1
	M5	2	3	2	2
	M6	1	3	1	1
31	L5	33	2	10	8
	L6	30	1	4	2
	M5	18	1	6	6
	M6	20	2	7	7

the base high bias index and the base low bias index is listed in Table 4 for the word meaning and best title items for each of the four racial group comparisons. The simple comparison of these two types of items does not reveal a clear tendency for one type to be biased against black students and the other biased in their favor. With the exception of Item 31, the best title items have few high rankings on either of the indices. In addition to Item 25, Item 2, "there"; Item 27, "reigning"; Item 17, "setting"; and Item 42, "speculate" tended to have fairly high ranks on the base high bias index. Some of the other word meaning items, however, have relatively low-ranking base high bias indices and may even rank higher on the base low bias index (e.g., Item 15, "rest"). Thus, generalizations based on such surface level characteristics of the items do not seem warranted.

Discussion and Conclusions

The analyses involving comparisons of students at different grade levels or who attend schools located in neighborhoods with different income levels showed that the ICCs were generally very similar. For example, the ICCs based on a sample of fifth-grade white students attending schools in low-income neighborhoods were almost indistinguishable from those for a sample of their sixth-grade counterparts. The results, showing a high degree of similarity between ICCs for the within-race comparison involving differences in the other two grouping variables, lend credence to the viability of the general approach. A basic assumption of the latent trait model is that the item parameters, and therefore the ICCs, are invariant over different groups of people. Thus, the remarkably good invariance of the ICCs over grade level and income level within racial groups suggests that the model is reasonable for the 45 items on the test that was analyzed.

The degree of invariance in the ICCs was noticeably less for the racial group comparisons than for either the grade or income level comparisons. This suggests that there are some items that function differently for black students than they do for white students. Such items may reasonably be labeled as biased. Whatever the cause of the difference in the ICCs, the effect of including a larger or smaller number of items where the ICC of one group is above that of another is the same. The relative standing of

Table 4
Rank Order of Base-High and Base-Low Area Bias Indices
For the Four Racial Group Comparisons
Involving Word Meaning and Best Title Items

Item Number	Word	Base-High Area				Base-Low Area			
		L5	L6	M5	M6	L5	L6	M5	M6
		Word Meaning							
2	there	12	5	8	12	41.5	40	24.0	33
6	rings	43	9	34	45	22.0	34	41.5	23
15	rest	21	42	21	36	11.0	5	12.0	5
17	setting	20	10	5	3	4.0	8	41.5	9
19	run	23	12	16	42	41.5	42	41.5	14
23	tribute	31	35	24	29	21.0	20	7.0	40
25	character	9	1	2	1	20.0	11	3.0	3
27	reigning	2	6	14	10	41.5	44	13.0	40
29	assumed	24	37	33	24	31.0	21	10.0	26
39	true	7	27	39	18	17.0	22	33.0	28
42	speculate	1	17	9	6	10.0	30	2.0	12
		Best Title							
5		8	20	41	21	34.0	31	8.0	40
11		18	23	27	19	36.0	39	28.0	40
18		16	31	22	9	41.5	27	41.5	15
24		13	37	20	43	29.0	14	30.0	10
31		33	30	18	20	2.0	1	1.0	2

black students would be higher on a test that had fewer items where the ICC for white students was above the one for black students.

Although a few items were consistently identified as biased in each of the four independent comparisons, the consistency of identification at different grade levels and/or different income levels was far from perfect. For example, using the criterion that the square root of the sum of squares bias index was greater than .2, seven items were identified as possibly biased in the comparison of low-income white students in grade 6 with low-income black students in Grade 6. Of these seven, Items 7, 3, and 4 were also identified as possibly biased in the other three racial group comparisons (i.e., LW5-LB5, MW5-MB5 and MW6-MB6, respectively). Only three items were identified as possibly biased in all four comparisons. The modest amount of agreement among the independent comparisons suggests that, at least for the test studied, it may be difficult to identify biased items because of the unreliability of the indices used. Furthermore, in many practical settings, it is difficult to obtain large enough samples to get stable estimates of the item parameters. Simulation studies now in progress are being used to determine the statistical power of the tests for bias with given sample sizes.

The use of .2 as a cutoff for identifying an item as biased depended on observing within-race distributions of bias indices. Such an approach may be feasible in practical situations, even though the minority group sample size is not large enough to divide into subgroups. Distributions of bias indices comparing subgroups of the majority group could still be obtained and used as a base for judging indices obtained from majority group versus minority group comparisons.

Although the ICCs were substantially different for white and black students for a few of the items in one or more of the comparisons, the

overall impression is that the ICCs were generally quite similar. Furthermore, the direction of the bias for the few items that showed a consistently large difference was not always against black students. One of the three consistently identified items was, if anything, biased in favor of black students. Thus, eliminating Items 3, 25, and 31 from the test would have only a trivial net effect on group differences.

Comparisons of the content and format characteristics of items that were identified as biased with those that were not, or between items biased in different directions, did not lead to the identification of any systematic differences. For example, items asking the meaning of a word in context sometimes appeared to be biased in one direction and sometimes in the other. Thus, no generalized principles that would be useful in avoiding items that tend to show bias can be stated for guiding the future construction of tests of reading comprehension. Instead, only a post-hoc analysis procedure that may be useful in eliminating biased items after the items have been administered can be offered.

There are important advantages in the use of comparisons of ICCs such as those in this study over approaches that simply compare estimated item parameters. It is possible, as was sometimes observed in the present study, for item parameters to be substantially different yet for there to be no practical difference in the ICCs. This can occur, for example, where the b parameter is estimated to be exceptionally high for one group. To illustrate this, consider the following pairs of hypothetical item parameters for two groups in terms of a common θ scale: group 1, $a = 1.8$, $b = 3.5$, and $c = .2$; group 2, $a = .5$, $b = 5.0$, and $c = .2$. The item difficulties and discriminations for the two groups are markedly different; but the difference in the ICCs is never greater than .05 for θ values between -3 and $+3$. Thus, the suggestion of bias based on a large difference in estimated item difficulty or discrimination might be misleading. The value of practical concern is the difference in the probability of correctly answering the item for people of equal ability from different groups. This is, of course, precisely the difference in ICCs.

Major reliance was placed on the square root of the sum of squares bias index for the initial screening of items in this study. Reliance on the absolute value index would have led to very similar results. Analyses were also conducted using bias indices that were weighted by estimated standard errors of the difference in ICCs (Linn et al., 1980). The weighted indices yielded very similar results.

The use of estimates of the standard errors of the ICCs seems potentially useful. By plotting bands of two standard errors on either side of the ICCs, it became evident that some seemingly large differences in ICC curves were occurring only in regions where one or both of the ICCs being compared were poorly estimated. The advantages of using estimated standard errors were not very apparent in terms of a comparison of the weighted and unweighted bias indices, however. It may be that better estimation procedures are needed for this purpose.

One problem that may limit the utility of the standard errors as they were estimated in this study is caused by the tendency for the LOGIST-estimated abilities of some subjects to diverge. To deal with this problem, the ability estimates were arbitrarily limited to a range of +4.0 and −4.0. For some of the groups, sizeable numbers of students had ability estimates at the lower extreme. For example, 44 of the MB5 sample students had estimated θ's of −4.0. This artificial clustering of subjects at the extreme results in estimated standard errors of the ICC at low ability levels that are too small. That is, the inflated number of examinees at the extremes makes it appear as if there is more information at that ability level than would be the case without the need to fix bounds on θ. In analyses now underway, this problem can be dealt with by estimating standard errors after deleting examinees with extreme θ values or by using estimated ability distributions.

Despite the limitations noted above and the fact that the results did not lend themselves to

making generalizations about features of items that result in biased estimates of achievement for members of a particular subgroup, there are still some noteworthy results from the study. It provides strong support for the reasonableness of the three-parameter model for data of this kind. The across-grade level comparisons revealed strikingly similar item characteristic curves. The procedures used for placing confidence bands around the item characteristic curves yielded reasonable results; and, with refinements such as those suggested above, they hold the promise of substantially improving the basis for comparing item parameters and item characteristic curves.

References

Anastasi, A. *Psychological testing* (4th ed.). New York: Macmillan, 1976.

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

Bianchini, J. C., & Loret, P. G. *Anchor test study final report. Project report and Volumes 1 through 30; and Anchor test study supplement. Volumes 31 through 33.* Berkeley CA: Educational Testing Service, 1974. (ERIC Document Reproduction Service Nos. ED 092 601 through ED 092 634).

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading MA: Addison-Wesley, 1968.

Cleary, T. A. Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement*, 1968, *5*, 115-124.

Cleary, T. A., & Hilton, T. L. An investigation of item bias. *Educational and Psychological Measurement*, 1968, *28*, 61-75.

Coffman, W. E. Sex differences in responses to items in an aptitude test. In I. J. Lehmann (Ed.), *Eighteenth Yearbook*. East Lansing MI: National Council on Measurement in Education, 1961.

Durost, W. N., Bixler, H. H., Wrightstone, J. W., Prescott, G. A., & Balow, I. H. *Metropolitan achievement tests, Form F.* New York: Harcourt, Brace, & Jovanovich, 1970.

Eells, K., Davis A., Havighurst, R. J., Herrick, V. E., & Tyler, R. W. *Intelligence and cultural differences*. Chicago: Chicago Press, 1951.

Harms, R. A. *A comparative concurrent validation of selected estimators of test item bias.* Unpublished doctoral dissertation, University of South Florida, 1978.

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, Annapolis MD, December 1975.

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

Ironson, G. H. *A comparative analysis of several methods of assessing item bias.* Paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada, April 1978.

Linn, R. L. Fair test use in selection. *Review of Educational Research*, 1973, *43*, 139-161.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. *An investigation of item bias in a test of reading comprehension*, 1980. (ERIC Document Reproduction Service No. ED 184 091).

Lord, F. M. A study of item bias using item characteristic curve theory. In Y. H. Poortingal (Ed.), *Basic problems in cross-cultural psychology*. Amsterdam: Swets & Zeitlinger, 1977. (a)

Lord, F. M. Practical applications of item characteristic curve theory. *Journal of Educational Measurement*, 1977, *14*, 117-138. (b)

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

Petersen, N. S., & Novick, M. R. An evaluation of some models for culture-fair selection. *Journal of Educational Measurement*, 1976, *13*, 3-29.

Rudner, L. M. *An evaluation of select approaches for biased item identification.* Unpublished doctoral dissertation, Catholic University of America, 1977.

Shepard, L., Camilli, G., & Averill, M. *Comparison of six procedures for detecting test item bias using both internal and external ability criteria.* Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, April 1980.

Warm, T. A. *A primer of item response theory* (Technical Report No. 941078). Oklahoma City: U. S. Coast Guard Institute, Department of Transportation, 1978. (NTIS No. AD A063072)

Wood, R. L., Wingersky, M. S., & Lord, F. M. *LOGIST: A computer program for estimating examinee ability and item characteristic curve pa-*

rameters (ETS RM 76-6). Princeton NJ: Educational Testing Service, 1976.

Wright, B. D. Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 1977, *14*, 97-116.

Acknowledgments

The research reported herein was supported in part by the National Institute of Education under Contract No. US-NIE-C-400-76-0116. We thank William Tierre for his help with the data preparation and analysis.

Author's Address

Send requests for reprints or further information to Robert L. Linn, 210 Education Building, University of Illinois, 1310 S. Sixth St., Champaign IL 61820.