DOCUMENT RESUME

ED 268 157	TM 860 179

AUTHOR	Scheuneman, Janice
TITLE	Exploration of Causes of Bias in Test Items.
INSTITUTION	Educational Testing Service, Princeton, NJ. Graduate Record Examination Board Program.
SPONS AGENCY	Graduate Record Examinations Board, Princeton, N.J.
REPORT NO	ETS-RR-85-42; GREB-81-21P
PUB DATE	Dec 85
NOTE	78p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	Black Students; *College Entrance Examinations; Graduate Study; Higher Education; *Item Analysis; Latent Trait Theory; Multiple Choice Tests; *Performance Factors; Racial Differences; *Racial Factors; Research Design; Scores; *Test Bias; Test Format; Testing Problems; Test Items; White Students
IDENTIFIERS	*Graduate Record Examinations

ABSTRACT

A number of hypotheses were tested concerning elements of Graduate Record Examinations (GRE) items that might affect the performance of blacks and whites differently. These elements were characteristics common to several items that otherwise measured different concepts. Seven general hypotheses were tested in the form of sixteen specific hypotheses, most with five or six items. Items were developed in pairs, some with and others without the hypothesized element, and were administered within the GRE analytical, verbal, and quantitative sections. Log linear analyses of the sixteen hypotheses showed interactions between group membership and item version, indicating a differential effect on black and white examinees' performance. A latent trait approach was also used to examine individual item bias for eight of the hypotheses. The manipulations included context clues, item format, vocabulary, selecting the false rather than the true answer in a multiple choice item, understanding the item writer's inference, test wiseness, location of the correct response, and ambiguity versus structure. It was concluded that the item manipulations did have differential effects on blacks' and whites' performance. However, the complexity of the effects suggested that other uncontrolled factors affecting performance were present. (Author/GDC)

GRE

GRADUATE RECORD EXAMINATIONS

FXPLORATION OF CAUSES OF BIAS IN TEST ITEMS

Janice Scheuneman

GRE Board Professional Report GREB No. 81-21P
ETS Research Report 85-42

December 1985

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

EDUCATIONAL TESTING SERVICE, PRINCETON, NJ

Exploration of Causes of Bias in Test Items

Janice Scheuneman

GRE Board Professional Repɔrt No. 81-21P

December 1985

Acknowledgments

My thanks to my colleagues at ETS who prepared the test items for this study and without whom the study could never have been done. They are:

Verbal:	Mari Pearlman
	Edward Shea
Quantitative:	Patricia Kuntz Klag
	Beverly Whittington
Analytical:	Clark Chalifour
	Richard Evert
	Elizabeth McGrail
	Erich Woisetschlaeger

Carole Slaughter and Paul Ramsey also participated in the planning phases and in review of the items following analysis. Paul Holland provided advice and counsel on the log linear analyses. The item bias analyses were planned and executed by David Thissen and Lynne Steinberg at the University of Kansas in consultation with Howard Wainer of Educational Testing Service.

Abstract

Although the study of item bias has been an active area of research for the past few years, little progress has been made in understanding what factors in a test might create bias. To advance this understanding, a number of hypotheses were developed concerning elements of the items that might affect the performance of Blacks and Whites differently. The elements investigated were characteristics that would be common to several items that otherwise measured different concepts or content.

Seven general hypotheses were developed and tested in the form of 16 specific hypotheses, most with five or six items. For purposes of this study, items were developed in pairs, with a hypothesized element present in one item and missing or modified in the other. The item pairs were assembled into two separate sections and administered as part of the Graduate Record Examinations General Test. Pairs of test sections were developed for each of the three areas tested in the General Test: verbal, quantitative, and analytical ability. These sections were spiralled for administration to yield randomly equivalent groups of examinees.

The 16 specific hypotheses were evaluated separately using log linear analyses. Results showed that 10 of the 16 hypotheses showed interactions between group membership and item version indicating a differential effect of the manipulation on the performance of Black and White examinees. Eight of these hypotheses were also analyzed for item bias at the individual item level, using a procedure based on item response theory.

The conclusion was that item manipulations of the type hypothesized did have a differential effect on the performance of Black and White examinees. The complexity of the effects, however, suggested that other uncontrolled factors affecting performance were also operating. A number of suggestions for future research based on these results were provided.

To date, research on test bias has done little to cast light on the possible sources of bias in testing. If, however, bias is in fact a factor in tests that works to inflate score differences between two population groups, some understanding of the causes of bias is essential if we are, one day, to take effective remedial action. A theory of bias now being developed provides a perspective broader than the perspectives suggested by previous research, which has tended to focus on other bias issues. From the perspective of this bias theory, questions concerning sources or causes of bias can be more readily formulated and addressed. (Details of the bias theory and its relation to the results of test and item bias research are provided by Scheuneman 1981, 1984.)

In this theory, bias is defined as a multifaceted component of observed test scores that, like the usual measurement error, causes an observed score to be different from the "true" score, but, unlike the usual measurement error, is associated with membership in a particular group and, for members of that group, has an expected value other than zero. That is, a biased score is a systematic over- or underestimate of ability for members of the group. This bias is postulated to stem from two major sources of bias--individual difference characteristics of examinees, which are differently distributed within groups, and characteristics of tests or test items, which have different effects for persons in different groups.

Of primary interest in this study are sources of bias that lie in characteristics of test items. In this respect, the study relates to item bias research, where the identification of items that contain such characteristics is clearly an objective. The item bias methodologies, however, have focused on detecting "outliers," items that are functioning most differently for two groups. Perhaps this is why the interpretation, the attribution of possible causes of the obtained result, has also tended to focus on the peculiarities of the items identified rather than on communalities that may exist among them. The bias theory from which this study is derived argues that the causes of bias that are likely to be important, in the sense of a greater distortion of score differences between groups, are those that lie in characteristics that are common to several items in a test.

According to the theory, bias may even exist in tests where no outliers are found. For example, if all the items in a test use a single format and that format is more familiar to the members of one group than to those of another, the test scores for the two groups may suggest that the differences in ability are larger than they really are, that is, the test would be biased. Further, several such characteristics may be operating in a single test. Each such characteristic may, by itself, have relatively little impact, but the cumulative effect of several such characteristics may be significant. Items detected as outliers in statistical studies of item bias may, therefore, be those where other item features interact with the biasing characteristic to produce an unusually large effect or two or more biasing characteristics may be operating in the same item. If so, such effects would be unlikely to be readily discernible by an investigator, particularly if items were examined in isolation, perhaps explaining why item bias results have so frequently been found to be uninterpretable.

If these more general but subtler effects exist, they should be detectable. The thesis of this work, therefore, is that characteristics of items can be identified that are not peculiar to single items and that have a differential effect on the performance of Black and White examinees. In evaluating this thesis, however, a practical trade off was necessary. Prior to this work, little but speculation existed as to what such characteristics might be. A choice needed to be made with regard to how many hypotheses concerning the nature of these characteristics could be evaluated within the scope of this study. If few were evaluated, more items could be developed for each and hence more power would be available for identifying effects expected to be subtle. As a first effort in this direction, however, a decision was made to look at several possible hypotheses to gain more breadth of results at the risk of less certainty about the outcome. In this respect, this study should be seen as exploratory. The major questions to be addressed here are two: Can such effects be demonstrated in at least some instances, and what avenues might fruitfully be explored in more depth and with better control in the future?

Method

To evaluate the effects of the hypothesized characteristics, items were constructed in pairs with the characteristic or item feature present in one item and absent or modified in the other. Items were then assembled into different test books and administered as part of the GRE General Test. Test books were spiralled for administration, yielding randomly equivalent samples of Black and White examinees for each item ot a pair. The differences in performance on items with similar characteristics were then evaluated to determine if the effect of the item manipulation was the same for both groups.

The Test

The GRE General Test consists of seven separately timed sections. The operational portion of the test, from which examinees' scores are obtained and reported, consists of two sections each of verbal, quantitative, and analytical ability items. The seventh section (which may or may not be in the seventh position in the test book) serves one of several purposes for a given administration. It may be used for equating, pretesting new items, or various experimental purposes.

The verbal sections of the test are made up of four item types: sentence completions, antonyms, analogies, and reading comprehension sets. The quantitative sections involve concepts from arithmetic, algebra, and geometry. In addition to the usual math-type item format, these sections include data interpretation items and quantitative comparison items. Quantitative comparisons require the examinee to compare the quantities given in two columns and to decide whether one quantity is greater than the other, whether the two quantities are equal, or whether the relationship cannot be determined from the information given. The analytical ability sections consist of two types of questions, analytical reasoning and logical reasoning. Analytical reasoning items focus on the ability to analyze a given structure of arbitrary relationships and to deduce new information from that structure. Logical reasoning items focus on the ability to understand and analyze relationships among arguments or parts of an argument.

The Subjects

The experimental items were administered in six different forms of the variable section of the GRE General Test. Forms were spiralled for administration yielding random samples of examinees from the December 1982 Saturday administration of the test. Ethnic background was self-identified on registration material submitted prior to the testing and matched later with the score records.

The Black and White examinees from this administration were found to differ in background characteristics other than racial or ethnic group. In particular, the Black examinees were more often female and had different concentrations in areas of academic preparation than White examinees. Since performance is known to be different in at least some instances for men and women and for students from different academic backgrounds, these differences between the Black and White groups could cause difficulties in attributing differences to racial or ethnic group membership, although clearly a number of other background factors had been left uncontrolled. Analyses were hence performed using all examinees who identified themselves as Black and a sample of Whites selected to be similar to the Blacks in these background characteristics.

The White sample was selected randomly within categories defined by sex and by broad major field of undergraduate work: physical sciences, biological sciences, humanities, and social sciences. A spaced sample of two of every three men was taken first and pooled with all women for each form. Spaced samples were then taken to select one of three humanities majors, one of two biological science majors, and one of three physical science majors. All social science majors were retained. Characteristics of the samples for each of the three pairs of tests are given in Tables 1 and 2. Differences between the proportion of males and females and among the proportions in each major field were not significant for the four samples (Black samples and White samples for each of the two tests) for any of the three content areas.

Table 1

Number of Examinees by Race
and Academic Major Subject
(All Black Examinees/White Sample)

		Verbal		Quantitative		Analytical	
	Major	Black	White	Black	White	Black	White
Test 1	Humanities	27	187	31	162	32	179
	Social Sciences	178	1219	178	1220	154	1188
	Biological Sciences	51	353	60	366	50	322
	Physical Sciences	24	188	33	162	36	213
	Total	280	1947	302	1910	272	1902
	χ^2 Between Groups	.34		3.79		3.83	
	Probability	.95		.28		.28	
Test 2	Humanities	30	184	30	196	35	181
	Social Sciences	160	1233	188	1248	170	1208
	Biological Sciences	51	343	56	342	50	356
	Physical Sciences	34	169	33	157	35	189
	Total	275	1929	307	1943	290	1934
	χ^2 Between Groups	5.19		2.70		4.02	
	Probability	.16		.44		.26	
	χ^2 Between Tests	2.80	1.18	2.70	4.16	.36	3.04
	Probability	.42	.76	.44	.25	.95	.38

Table 2

Number of Examinees by Race and Sex
(All Black Examinees/White Sample)

		Verbal		Quantitative		Analytical	
	Major	Black	White	Black	White	Black	White
Test 1	Male	92	653	103	610	93	669
	Female	188	1294	199	1300	179	1233
	Total	280	1947	302	1910	272	1902
	χ^2 Between Groups	.05		.56		.10	
	Probability	.82		.45		.75	
Test 2	Male	82	621	106	626	97	668
	Female	193	1308	201	1317	193	1266
	Total	275	1929	307	1943	290	1934
	χ^2 Between Groups	.62		.64		.13	
	Probability	.43		.42		.72	
	χ^2 Between Tests	.60	.80	.03	.64	.03	.17
	Probability	.44	.37	.85	.42	.85	.68

Development of the Experimental Items

A series of general hypotheses was initially generated that suggested possible sources of bias in test items of the type appearing in the GRE General Test. These hypotheses were based on previous work (Scheuneman, 1979, 1982) and on suggestions from ETS staff and GRE Board advisory committees. Meetings were then held with the respective ETS test development staff members who would be preparing the items for each of the three ability areas measured by the test: verbal, quantitative, and analytical. At each meeting, these hypotheses were discussed to determine which ones might be evaluated within that particular area and the types of items that might be used in each case. Out of this process, 16 specific hypotheses were selected that could be subsumed by seven more general hypotheses. These hypotheses are described in the following section.

For each of the hypotheses, items were developed in pairs that were kept as similar as possible except for the hypothesized factor. Items were then arranged into two test forms with one item from each pair appearing in a particular form with the same serial position in each test. Two paired forms were developed in this way for each of the three ability areas. For quantitative and analytical tests all the items for a given version (A or B) of a hypothesis appeared in the same test book. Except for Hypothesis 1.1, some items from both versions of the verbal hypotheses appeared in each test book.

The Hypotheses

Throughout this report, the discussion will be clarified if a few terms are defined here. Each item pair occurs in the same serial position in a test section. Hence, verbal "Item 1" refers to the pair of items that appear in the first position in the two test sections with verbal items. Each "item" thus has two versions. These two versions differ according to some manipulation of the item elements dictated by the expectations of a given hypothesis. Although the two versions of an item pair were kept as similar as possible, in most instances other item elements also changed as a necessary consequence of the intended manipulation if the items were to appear sensible and to function properly. Such "unintentional" changes were probably the source of much of the noise that became apparent in the results.

In the following section, a brief description is provided of each hypothesis and the two item versions. A summary of the hypotheses and versions is provided in Table 3 for reference in reading this report. Table 4 shows the content area--verbal, quantitative, or analytical--in which a hypothesis was evaluated and the number of items in each of the hypotheses and areas.

Table 3

Summary of Hypotheses

<u>Hypothesis 1.0</u>

Item format affects the performance of Blacks more than that of Whites.

1.1A	Vocabulary out of context (antonyms)
1.1B	Vocabulary in context (sentence completions)
1.2A	Quantitative comparisons
1.2B	Standard multiple-choice
1.3A	Standard multiple-choice
1.3B	Roman numeral format

<u>Hypothesis 2.0</u>

Blacks are more apt than Whites to be misled by vocabulary items that depend on using secondary meanings or altering word meaning with prefixes or suffixes to produce item difficulty.

2.1A	Less common word meaning
2.1B	More common word meaning
2.2A	Suffix/prefix with less common meaning
2.2B	Suffix/prefix with more common meaning

<u>Hypothesis 3.0</u>

Items that ask for the one false answer rather than the one true answer are more apt to be confusing to Blacks.

3.1A	One false answer
3.1B	One true answer
3.2A	One true answer
3.2B	One false answer

<u>Hypothesis 4.0</u>

Performance of Blacks is more likely to be affected by items calling for inferences, since such items are more apt to require inferring the intent of the item writer.

4.1A	Inference from reading passage required
4.1B	Material directly stated in passage

Table 3 (cont.)

4.2A	Implies most likely response is required
4.2B	States most likely response is required
4.3A	Implies all possible responses
4.3B	States "a complete and accurate list"

Hypothesis 5.0

White examinees are more likely than Blacks to capitalize on the information often provided unintentionally in test items.

5.1A	Test-wiseness cues absent
5.1B	Test-wiseness cues present
5.2A	Test-wiseness cues on distractor
5.2B	Test-wiseness cues on key

Hypothesis 6.0

Blacks and Whites will be differently affected by the placement of the key among the distractors.

6.1A	Strongest distractor before key
6.1B	Key before strongest distractor
6.2A	Key in center of option sequence
6.2B	Key not in center of sequence

Hypothesis 7.0

Ambiguity in a task is more apt to be misleading to Blacks. Hence, Blacks will tend to perform better when the task is structured in concrete terms.

7.1A	Use numbers in problem
7.1B	Use symbols in problem
7.2A	Use diagrams
7.2B	Use verbal descriptions

Table 4

Number of Hypotheses and Items Tested for Each Content Area

Content Area	Hypothesis Number	Items per Hypothesis	Items per Area
Verbal	1.1	5	
	2.1	7	
	2.2	5	
	3.1	5	
	4.1	5	
	5.1	5	
	6.1	[illegible]	42
Quantitative	1.2	6	
	5.2	6	
	6.2	6	
	7.1	6	
	7.2	6	30
Analytical	1.3	5	
	3.2	5	
	4.2	5	
	4.3	5	20

Hypothesis 1.0. Hypothesis 1.0 concerns the way which the task required by the item is presented. Five items were developed for each of three specific hypotheses.

Hypothesis 1.1 contrasts antonym and sentence completion items. Since Black examinees as a group have frequently been demonstrated to have less well developed vocabularies, it was expected that the context provided by the sentence completion items would help compensate for this difference. Sentence completion items were expected to be easier than the antonym items and the difference in difficulty was expected to be larger for Blacks than for Whites. The stimulus word for the antonym versions was the key term in the sentence completion items. The stimulus words were arbitrary, dichotomy, naivete, ingenuity, and diaphanous. This last term comes from a disclosed item and is used as Example 1.*

Example 1

Hypothesis 1.1

Version A

DIAPHANOUS:

(A) inherent
*(B) substantial
(C) barbarous
(D) obsolete
(E) repetitive

Version B

By its very strength and sharpness the sunlight of Greece forbids the shifting, melting, _____ effects which give so delicate a charm to the French or Italian scene.

(A) insipid (B) barbarous
(C) sustaining *(D) diaphanous
(E) ostentatious

* Notice that unless otherwise indicated examples were developed specifically for this paper since one version of the items used in the study remains in the active item pool in most instances and hence is still secure.

Hypothesis 1.2 contrasts two quantitative item types--quantitative comparisons and the standard problem format. It was thought that the less familiar quantitative comparisons items might be more difficult for Blacks and cause their performance to differ more than that of Whites. The problem posed was the same in both versions, but the form in which the answer choices were presented differed. See Example 2.

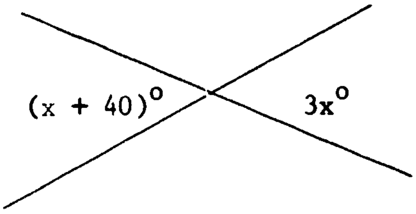
Example 2

Hypothesis 1.2

Version A

A if the quantity in Column A is greater;
B if the quantity in Column B is greater;
C if the two quantities are equal;
D if the relationship cannot be
  determined from the information given.

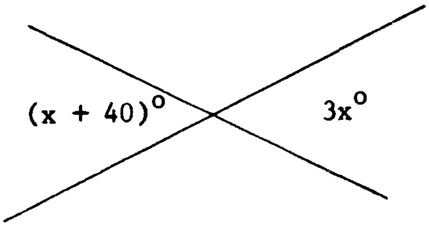
Column A	Column B



x	20

(kev = C)

Version B



In the figure above, x=

(A) 10 (B) 12.5
*(C) 20 (D) 35 (E) 40

Hypothesis 1.3 concerned the way in which the options were presented. Analytical items were developed where the situational set was the same for both versions, but the options were presented in either a standard set of multiple-choice options or a "Roman numeral" format. In the latter, some number of statements is made (usually three) each designated with a Roman numeral. In response to a question concerning which of these are true, the options would be of the form "I only," "I and II", etc. Again, it was expected that the standard format would be comparatively easier for Blacks. Example 3 illustrates this hypothesis.

Example 3

Hypothesis 1.3

All M's are Q's.
All F's are Q's.
All Q's are F's or M's, but not both.
Some Q's are Y's.
Not all Y's are Q's.
All W's are Y's.

Version A

If a W is an M, it must also be

(A) a Y only
(B) a Q only
(C) an F only
(D) an F and a Q
*(E) a Y and a Q

Version B

If a W is an M, which of the following must a W also be?

I. Y
II. Q
III. F

(A) I only (B) II only
(C) III only *(D) I and II
(E) II and III

Hypothesis 2.0. Again, on the assumption that, as a group, Blacks have less extensive vocabularies, the hypothesis was that manipulation of vocabularies in antonym items to increase the difficulty would have a greater impact on the performance of Black examinees than on that of White examinees.

Hypothesis 2.1 concerned the effect of increasing the vocabulary level of the term intended to be the key in order to make the item more difficult. Seven items were developed using the following words as stimuli: accretion, conventional sap, grave, requisite, curb, and dampen. The first of these was disclosed and this item pair is shown as Example 4. For each version, different words were provided as the correct response, though generally the distractors remained the same.

Example 4

Hypothesis 2.1

Version A	Version B
Accretion:	Accretion:
(A) disinterest	(A) disinterest
(B) disagreement	(B) disagreement
(C) exactitude	(C) exactitude
(D) adhesion	(D) adhesion
*(E) attrition	*(E) diminution

In Hypothesis 2.2, the item difficulty was altered through the addition, deletion, or modification of prefixes or suffixes. The following were the stimulus words with the alternate version in parentheses: (un)gainly, (un)impeachable, (im)palpable, (in)determinate, and guile(ful/less). Example 5 illustrates this hypothesis.

Example 5

Hypothesis 2.2

Version A	Version B
Ineffectual:	Effectual:
*(A) powerful	*(A) powerless
(B) energetic	(B) pathetic
(C) knowledgeable	(C) ignorant
(D) economical	(D) costly
(E) skillful	(E) awkward

<u>Hypothesis 3.0</u>. The standard multiple-choice item essentially asks the question, Which of the following options is true? A common variant reverses this by stating that all of the following are true EXCEPT... (or by implication, Which of the following is false?). Such items hav. often been identified as problematic in previous bias studies. It was hypothesized here that the impact of this negative phrasing would be greater on the performance of Black examinees. Five item pairs of each type were developed, five with verbal items (Hypothesis 3.1) and five with analytical items (Hypothesis 3.2). The passage to which the items referred was the same for an item pair while the statements of the question (which of the following is true or false) and the alternative statements constituted the two item versions. A single example is adequate for Hypotheses 3.1 and 3.2 since the difference between these is in the type of content, verbal or analytical, not the form. Example 6, an analytical item, is illustrative.

Example 6

Hypothesis 3.1 and 3.2

Offshore blasting in oil exploration does not hurt fishing; blasting started this year, and this year's salmon catch has been the largest in a long time.

Version A

All of the following statements, if true, are valid objections to the argument above EXCEPT:

(A) The salmon is only one of many species of fish that might be affected by the blasts.

*(B) The rapid changes of water pressure caused by the blasts make salmon mate more frequently.

(C) The noise of the blasts interferes with the feeding habits of salmon.

(D) Vibrations from the blasts destroy fish eggs.

(E) Factors that have nothing to do with the well-being of salmon may significantly affect the size of one year's catch.

Version B

Which of the following statements, if true, is a valid objection to the argument above?

*(A) The salmon is only one of many species of fish that might be affected by the blasts.

(B) The rapid changes of water pressure caused by the blasts make salmon mate more frequently.

(C) Salmon are particularly vulnerable to the effects of underwater blasting.

(D) Salmon spawn in fresh water rather than the ocean.

(E) Oil exploration is essential to the nation's economy.

Hypothesis 4.0. To the extent that the task demanded by the item is ambiguously stated or is open to interpretation, White examinees are hypothesized to be better able to infer the intent of the item writer and less apt to be confused or misled. Consequently, Black examinees were expected to be influenced more by different statements concerning the task requirements. Five item pairs were developed for each of three specific hypotheses.

Hypothesis 4.1 deals directly with inference, assuming that material not stated may be less accessible to Black than to White examinees. Here the two items of the pair were basically the same. The different versions were reflected by whether or not the material to be inferred was directly stated in the passage with a corresponding change in the statement of the question. Illustrations are provided in Example 7.

Example 7
Hypothesis 4.1

Version A

There is little agreement concerning the way in which kinds of avalanches should be classified. Some classification systems depend on the kind of snow involved, others are concerned with the type of movement, and one scheme includes both, as well as several other criteria. Existing descriptive terms, most of them German, are deeply rooted in avalanche parlance: they are expressive, but they are often untranslatable into other languages and lack precision in their own. Furthermore, as Dr. Quervain has pointed out,* avalanches are not only concrete objects capable of being photographed; they are also events. As events, they include, for example, the development of the avalanche through the influences of weather; the incident that starts the snow moving; and the type of movement. The description of the avalanche as an object includes information about the depth, physical consistency, and stratification of the snow, the features of the terrain, and the type and the dimension of the break.

It can be inferred that the author mentions Dr. Quervain's observations (lines 8-10) primarily in order to

(A) indicate the temporary nature of an avalanche
(B) illustrate the imprecision of the terms used in avalanche classification
(C) introduce his own avalanche classification system
*(D) further emphasize the complexity of avalanche classification
(E) further explain why it is important to classify kinds of avalanches

*Underline shows the section of the passage that differs between versions. The underline did not appear in the original item.

Version B

There is little agreement concerning the way in which kinds of avalanches should be classified. Some classification systems depend on the kind of snow involved, others are concerned with the type of movement, and one scheme includes both, as well as several other criteria. Existing descriptive terms, most of them German, are deeply rooted in avalanche parlance: they are expressive, but they are often untranslatable into other languages and lack precision in their own. Other difficulties in establishing classification schemes for avalanches are illustrated by Dr. Quervain when he points out that avalanches are not only concrete objects capable of being photographed, they are also events. As events, they include, for example, the development of the avalanche through the influences of weather; the incident that starts the snow moving; and the type of movement. The description of the avalanche as an object includes information about the depth, physical consistency, and stratification of the snow, the features of the terrain, and the type and the dimension of the break.

According to the passage, Dr. Quervain's observation (lines 8-12) is intended primarily to

(A) indicate the temporary nature of an avalanche
(B) illustrate the imprecision of the terms used in avalanche classification
(C) introduce his own avalanche classification system
*(D) further emphasize the complexity of avalanche classification
(E) further explain why it is important to classify kinds of avalanches

Hypotheses 4.2 and 4.3 concerned analytical items. In both cases the item pairs were identical except for a change in the wording of the question asked. For Hypothesis 4.2, one version asks, Which is the "best" or "most" likely response? In the alternate version the "best" or "most" was deleted. In Hypothesis 4.3, the alternate wordings were similar to "Which of the following could be true?" and "Which of the following is a complete and accurate list of what could be true?" Illustrations are shown in Example 8.

Example 8

Hypothesis 4.2

It is clear from trends in the 1970's that economic classes in the United States have been growing farther apart from one another rather than becoming more nearly equal. The weekly spendable earnings of private-sector nonsupervisory workers have declined by 10 percent since 1970, and by 1979 had returned to their level of 1964. The claim of equality in the United States is more and more difficult to sustain.

Version A

Which of the following, if true, would support the argument of the passage above?

(A) The rate of unemployment fluctuated greatly in the 1970's.

(B) Public-sector workers also suffered a decline in their spendable income in the 1970's.

(C) The spendable income of workers declined in the 1950's.

*(D) Supervisors and owners maintained the level of their spendable income in the 1970's.

(E) By 1964 there was already a sizeable gap between economic classes in the United States.

Version B

Which of the following, if true, would best support the argument of the passage above?

(A) The rate of unemployment fluctuated greatly in the 1970's.

(B) Public-sector workers also suffered a decline in their spendable income in the 1970's.

(C) The spendable income of workers declined in the 1950's.

*(D) Supervisors and owners maintained the level of their spendable income in the 1970's.

(E) By 1964 there was already a sizeable gap between economic classes in the United States.

Hypothesis 4.3

Seven bottles of chemicals are arranged on a shelf in seven spaces numbered 1 through 7 consecutively from left to right. Each bottle occupies one space. Three bottles are filled with sulfate, two are filled with hydroxide, and two are filled with chloride.

No bottle of sulfate is next to another bottle of sulfate.

None of the bottles of sulfate is in space 3.

Neither bottle of chloride is in space 5.

Version A

If the two bottles of chloride are next to each other and the two bottles of hydroxide are next to each other, which of the following chemicals could occupy space 2?

*(A) Only chloride
(B) Only hydroxide
(C) Either chloride or hydroxide
(D) Either chloride or sulfate
(E) Either hydroxide or sulfate

Version B

If the two bottles of chloride are next to each other and the bottles of hydroxide are next to each other, which of the following is a complete and accurate list of the chemicals that could possibly occupy space 2?

*(A) Chloride
(B) Hydroxide
(C) Chloride, hydroxide
(D) Chloride sulfate
(E) Hydroxide, sulfate

Hypothesis 5.0. Just as White examinees might be better able to guess the intent of predominantly White item writers, so might they be better able to make use of cues inadvertantly provided that might signal which of the options presented is the intended key. The core for these items was the item stem. In one version, changes were made deliberately to point to the key in ways that would normally be avoided by item writers. These changes were absent in the alternate version.

Hypothesis 5.1 concerned verbal items and the cues provided were of the type that test developers know about and that have been discussed in the literature. Hypothesis 5.2 concerned quantitative items and specifically evaluated an option elimination strategy of a type identified by Smith (1982) for verbal items and generalized to quantitative items by Kuntz (1982). These hypotheses are illustrated in Example 9.

Example 9

Hypothesis 5.1

Khrushchev's gift to history is, and always was, himself. Khrushchev's greatest qualities, those that distinguished him from all other Soviet leaders, were his energy, his enthusiasm, his confidence in himself and in others. It was his prodigal personality, his ability to confess a mistake and reverse himself, his explosive unpredictability that did more than anything else to spring the genie of spontaneity out of the bottle of repression in which Stalin had contained the Russian spirit for thirty years.

Version A

According to the passage all of the following describe Krushchev's personality EXCEPT

(A) energetic
(B) unpredictable
(C) crafty
(D) clever
*(E) inflexible

Version B*

According to the passage, all of the following describe Krushchev's personality EXCEPT

(A) energetic
(B) spontaneous
(C) enthusiastic
(D) clever
*(E) inflexible

*Notice that in this item version the key is the only one of the distractors with a negative tone.

Hypothesis 5.2

If $x = 3z + 9$ and $z = y + 5$, what is x in terms of y?

Version A	Version B
(A) $y + 14$	(A) $y + 24$
(B) $3y - 6$	(B) $3y - 6$
(C) $3y + 4$	(C) $3y + 4$
(D) $3y + 14$	(D) $3y + 14$
*(E) $3y + 24$	*(E) $3y + 24$

Hypothesis 6.0. Effective test-taking strategies will positively influence performance. One hypothesis is that Black examinees have less well developed test-taking skills. One possible outcome of this is that the position of the key will affect their performance differently than that of White examinees.

Hypothesis 6.1 used 10 verbal analogy pairs. In each item pair, the stimulus words and the options were the same. The two versions differed only in the option order. Similarly for Hypothesis 6.2 the problem posed was the same for the six quantitative item pairs. Because options were placed in order of ascending (or descending) magnitude, key placement was influenced by replacing an option of magnitude less than the key with one larger (or vice versa). Item pairs illustrating these hypotheses are given in Example 10. The analogy item is a disclosed item actually used in this study.

Example 10

Hypothesis 6.1

Version A	Version B
Parchment: Paper::	Parchment: Paper::
(A) ink : quill	*(A) clavicord : piano
(B) radiator : thermostat	(B) radiator : thermostat
(C) embroidery : broadloom	(C) embroidery : broadloom
(D) citrus : juice	(D) citrus : juice
*(E) clavicord : piano	(E) ink : quill

Hypothesis 6.2

If $15 \leq x \leq 25$ and $y - x = 3$, what is the greatest possible value of $x + y$?

Version A	Version B
(A) 60	(A) 56
(B) 56	*(B) 53
*(C) 53	(C) 47
(D) 47	(D) 40
(E) 28	(E) 28

Hypothesis 7.0. Ambiguity in an item increases as the content of an item becomes more removed from the experience of the examinee. Again it is hypothesized that Black examinees are more apt to be misled by ambigious item content. Consequently, they would be expected to do relatively better where the content is less abstract and more concrete. This hypothesis was evaluated with two quantitative hypotheses with six item pairs each.

Hypothesis 7.1 concerned the use of algebraic symbols or variables such as x and y rather than numbers. The problem to be solved was the same in both versions, but the versions differed in whether this problem was posed using numbers or letters. Example 11 gives an illustration.

Example 11

Hypothesis 7.1

Version A

A ladder that is 10 feet long is leaning against a wall. The base of the ladder is on the floor 6 feet from the wall. What is the distance in feet, from the bottom of the wall to the point where the ladder touches the wall?

(A) 6
*(B) 8
(C) 9
(D) 10
(E) $2\sqrt{34}$

Version B

A ladder that is x feet long is leaning against a wall. The base of the ladder is on the floor y feet from the wall. What is the distance in feet from the bottom of the wall to the point where the ladder touches the wall?

(A) $\sqrt{x^2 + y^2}$

*(B) $\sqrt{x^2 - y^2}$

(C) $\sqrt{x + y}$

(D) $\sqrt{x - y}$

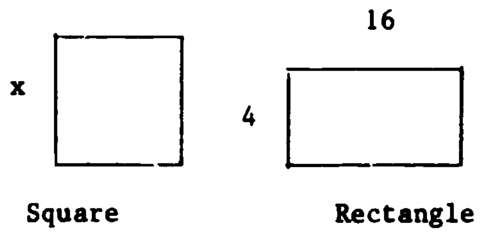
(E) $\sqrt{x} - \sqrt{y}$

Hypothesis 7.2 concerned the use of figures or diagrams to illustrate the problem posed. Again the problem to be solved was the same for the item pair. The version depended on whether a diagram or verbal description was provided in illustration. See Example 12.

Example 12

Hypothesis 7.2

Version A



In the figure above, if the area of the square is equal to the area of the rectangle, then x =

*(A) 8 (B) 7 (C) 6 (D) 5
(E) 4

Version B

If a square with each side x units in length has the same area as a rectangle 4 units by 16 units, then x =

*(A) 8 (B) 7 (C) 6 (D) 5
(E) 4

Data Analysis

Each hypothesis of a differential effect on the performance of Black and White examinees due to an item manipulation was tested using log linear procedures. The models chosen were "logit models," which are analogous to multiple regression procedures in which the variables are all categorical. The item response--correct or incorrect--was treated as the dependent variable while the item pair, the item version, and group membership were the independent variables (or predictors). In these analyses, the effects of most interest for the purposes of this study were the interaction between group and version and the three-way interaction with group, version, and item pair. Both of these interactions indicate that the effect of the item manipulation was different for Blacks and Whites. The three-way interaction indicates that this difference in effect varied among item pairs.

In addition to the log linear analyses of the effects of the independent variables on performance, item bias analyses were conducted that also took into account differences in item discrimination and the relative abilities of the two groups. These two sets of analyses differed in their perspective on the problem and the questions addressed by each. Where the focus of the log linear analyses was on the differences in difficulty between the two item versions within each group, the focus of the item bias analyses was on the differences between the two groups within an item version.

The item bias analyses used methods based on item response theory models. For the purposes of these analyses, an unbiased item was defined as one for which the probability of a correct response is the same for persons of a given level of the ability measured by the test regardless of the'r group membership. In the terms of item response theory, this defi...ion may be stated: the item characteristic curves of an unbia.ed item must be the same for two groups of interest (Lord, 1977; Scheuneman, 1980). In practice this means that the three parameters that serve to define these curves are the same for both Blacks and Whites. The method chosen to estimate these parameters was developed by Thissen (1982a) and is appropriate for this application where one sample is relatively small and the other much larger.

Both the log linear analyses and the bias analyses are based on successive fittings of a mathematical model to the data. A given model is first applied to the data and a statistic is computed that reflects the fit of that model. When an assumption is relaxed or an element added or deleted from the model, a new value for the statistic is obtained. The difference between successive statistics represents the change in the overall fit of the model to the data. The value of this difference gives a test for the significance of the interactions of interest in the log linear analysis or of the assumption that a given item parameter is equal for the two groups in the bias analysis. The details of these procedures are given in Appendixes A and B, respectively.

Results

The equivalence of the samples for a given pair of tests was evaluated first by scoring the operational sections of the content area corresponding to the experimental sections. The results are shown in Table 5. Both means and standard deviations were very similar for the two samples within a group; hence they were treated as equivalent in subsequent analyses.

Log Linear Analyses

The results for the relevant interaction effects are shown in Table 6. The chi square value shown for the group-by-version interaction is the difference between the chi square of fit for the three main effects (item pair, item version, and group) and the chi square of fit for these effects plus the group-by-version effect. The three-way interaction is evaluated by determining the chi square

Table 5

Performance of Samples

on Operational Tests[a,b]

	Blacks		Whites	
	Test 1	Test 2	Test 1	Test 2
Verbal				
Mean	32.00	31.4	46.3	46.0
SD	11.3	11.3	10.8	11.0
N	300	280	1940	1920
Quantitative				
Mean	27.1	26.5	38.6	38.5
SD	10.4	9.7	9 7	9.7
N	315	320	1900	1935
Analytical				
Mean	19.1	19.0	27.3	27.4
SD	6.7	6.5	7.1	7.2
N	285	300	1895	1925

[a] Mean scores are given for the operational sections in the same content area as the experimental section.

[b] Ns will generally be different from those shown in Tables 1 and 2. In the case of Black examinees, some were included in the analysis who did not provide sex or major subject on their registration form. In addition, the computer program from which these data were taken rounds down to an even multiple of 5.

Table 6

Chi Square Values for Interactions of Group and Version

Log Linear Analyses

	Group by Version			Group by Version By Item			
Hypothesis	χ^2	df	P	χ^2	df	P	
1.1	11.30	1	***	11.84	4	**	
1.2	.50	1		2.68	5		a
1.3	.11	1		3.44	4		
2.1	3.48	1	*	28.93	6	***	
2.2	12.76	1	***	8.00	4	*	
3.1	.50	1		14.76	4	***	
3.2	8.26	1	***	6.81	4		
4.1	1.23	1		3.81	4		
4.2	3.27	1	*	1.70	4		a
4.3	1.14	1		3.85	4		
5.1	8.65	1	***	13.97	4	***	
5.2	.90	1		.89	5		
6.1	.66	1		8.31	9		a
6.2	2.50	1		12 36	5	**	
7.1	2.80	1	*	27.27	5	***	
7.2	.94	1		11.37	5	**	

* $P < .10$
** $P < .05$
*** $P < .01$

a Version effects were not significant for these hypotheses. That is, when the two groups were pooled, the two item versions did not differ in difficulty.

for all effects <u>except</u> the three-way interaction (group-by-version-by-item pair). If the chi square is significant, it indicat⌄s that the three-way interaction must be taken into account if the model is to fi: the data. Results for all effects are given in Appendix A.

Four of the hypotheses had highly significant group-by-version interactions: Hypothesis 1.1, antonyms versus sentence completion; 2.2, the effects of adding (or modifying) prefixes/suffixes to the stimulus word in antonym items; 3.2, one true versus one false response in analytical items; and 5.1, test wiseness cues in verbal items. Marginal group-by-version interaction effects were noted with three other hypotheses: 2.1, substitution of a more difficult word for the correct response; 4.2, presence or absence of the word <u>most</u> or <u>best</u>; and 7.1, the use of numbers versus symbols in quantitative items. Of these, Hypotheses 1.1, 2.1, 2.2, 5.1, and 7.1 also had three-way interactions. Three other hypotheses had significant three-way interactions, but nonsignificant group-by-version interactions. These were Hypothesis 3.1, one true versus one false response for verbal items; 6.2, the position of the key in quantitive items; and 7.2, diagrams versus verbal descriptions in quantitative items.

In general, the three-way interaction indicates that the group-by-version effect is different for the different items making up a hypothesis. In order to understand the nature of this interaction more clearly separate analyses were performed for each item in the eight hypotheses that had significant three-way interactions. In the analysis of an individual item, the contingency table was a two-by-two-by-two table with group, version, and response (right/wrong). The chi square value was the fit statistic for a model with the two main effects only. If this value was significant, it indicated that the main effects alone did not fit the data and hence the interaction was required.

Table 7 provides the results for those items where the probability of the obtained chi square was less than .20. Higher levels of significance are indicated with asteriks. In order to achieve a clearer understanding of the different effects, however, further information is required. For this purpose, Table 7 also provides "odds ratios" and an indicator of which group or version had the larger effect. The odds ratio is the ratio of right to wrong responses for Version B divided by the ratio of right to wrong responses for Version A for each of the two groups. In both instances, if the effect of the manipulation is zero the odds ratio will equal one. A ratio greater than one indicates that Version B is easier than Version A; a ratio less than one indicates the reverse. Once the value of the odds ratio departs from one, however, the ratios for the two groups are no longer stricty comparable since the magnitude of the ratio is related to the relative difficulty. Hence, the indicator of the larger effect given in the table is based on item difficulties (percent of correct

Table 7

Analyses for Individual Items in Hypotheses with Significant Three Way Interactions

				Odds Ratio		Largest Effect	
Hypothesis	Item	χ^2 (df=1)	P	Black	White	Group	Version
1.1	1	1.70		.96	1.24	Wh	B
	2	2.09		11.37	15.80	Wh	B
	3	3.41	*	1.19	1.66	Wh	B
	5	3.66	*	.21	.14	Bl	B
2.1	7	32.13	***	9.32	34.97	Wh	B
	9	1.76		1.97	2.63	Bl	A
	11	1.67		2.63	3.38	Bl	A
	12	3.89	**	1.11	1.70	Wh	B
2.2	13	3.80	**	1.21	1.72	Wh	B
	14	2.80	*	1.29	1.79	Wh	B
	15	13.84	***	.69	1.39	Bl	B
	17	2.04		.88	1.16	-	B
3.1	18	5.35	**	1.20	1.88	Wh	B
	26	4.07	**	.73	.51	Wh	A
	32	4.97	**	.45	.71	Bl	B
5.1	23	1.66		.61	.58	-	-
	25	4.14	**	.77	1.19	-	B
	29	12.54	***	.69	1.32	Bl	B
6.2	13	3.36	*	.88	1.21	-	B
	21	4.50	**	.66	.98	Bl	B
	25	1.62		1.20	.96	Bl	A
	29	6.30	***	.54	.86	Bl	B
7.1	1	3.60	*	.69	.49	Wh	A
	17	23.78	***	.41	.16	Wh	A
	27	3.90	**	.40	.56	Bl	B
7.2	3	2.31		.69	.91	Bl	B
	6	3.32	*	.48	.66	Bl	B
	15	2.23		.95	.73	Wh	A
	23	2.89	*	.65	.44	Wh	A

* P < .10
** P < .05
*** P < .01

responses). This indicator is often, but not always, in agreement with the relative magnitude of the ratios for Blacks and Whites.

The three-way interactions may have resulted from a lack of consistency in three possible dimensions. First is the relative difficulty of the two versions. If one version is always more difficult than the other, the odds ratios will all be greater than one or less than one. For example, in Hypothesis 2.1, Version B was always easier than Version A, while in Hypotheses 7.1 and 7.2, Version A was always easier than Version B. Second is a consistency in the group for which the effect is larger. Hypothesis 1.1 showed larger differences for Whites in three of four cases while Hypothesis 6.2 similarly showed larger differences for Blacks. Third is the version that showed the greater group differences. For example, both Hypotheses 1.1 and 2.2 consistently showed larger effects in Version B.

A brief summary of the log linear results is as follows:

Hypothesis 1.1. A significant group-by-version interaction was found. The effect was larger for Whites and the difference between groups larger for Version B (sentence completion). The three-way interaction appears to be due primarily to item 5 (ingenuity), which showed larger differences for Blacks.

Hypothesis 1.2. This hypothesis showed no significant interactions but also showed no version effect. That is, the item manipulation resulted in little change in difficulty for either group.

Hypothesis 1.3. No interactions of interest were found here.

Hypothesis 2.1. The group-by-version effect for this hypothesis was marginal, perhaps due to the off-setting effects of the items contributing to the three-way interaction. Although Version B (more common meaning) was consistently easier, two items (grave and curb) showed larger effects for Blacks and Version A and two (conventional and dampen) for Whites and Version B.

Hypothesis 2.2. This hypothesis showed a strong group-by-version interaction with larger differences between groups for Version B (more common meaning). The three-way interaction appears to have resulted from an inconsistency in the relative difficulty of the two item versions for Blacks. Items 13 and 14 (unimpeachable and ungainly) both showed Version B easier than Version A and larger effects for Whites. Items 15 and 17 (impalpable and guileless), however, showed Version A (less common meaning) easier than Version B for Blacks. Item 15 also showed a larger effect size for Blacks, though for item 17, effects were nearly equal for Blacks and Whites, but in opposite directions.

Hypothesis 3.1. This hypothesis did not show the expected group-by-version interaction. Again, however, the patterns of the items contributing to the three-way interaction suggest that the effects were in different directions, canceling each other out when all items were combined. All three items with significant effects showed different patterns. Item 18 showed larger effects for Whites and Version B (one true response), item 26 for Whites and Version A, and item 32 for Blacks and Version B.

Hypothesis 3.2. This hypothesis showed a significant group-by-version interaction with Version A (one true response) showing the larger difference. The effect size was not consistently larger for Blacks or Whites, however, and Version A was consistently easier only for Whites.

Hypothesis 4.1. This hypothesis showed no interactions of interest.

Hypothesis 4.2. This hypothesis showed a marginal group-by-version interaction that is particularly interesting since the overall version effect is not significant. This seems to have occurred since the manipulation had virtually no effect on Whites, much the larger sample. The effect for Blacks was small, but sufficient to produce the small interaction. Version A ("best" or "most" absent) was easier than Version B.

Hypothesis 4.3. No interact! ns of interest were found for this hypothesis.

Hypothesis 5.1. Both a group-by-version interaction and a three-way interaction were found for this hypothesis. Version A (cues absent) was easier for Blacks, but Version B was more often easier for Whites and had the larger effects. For items 25 and 29, the effect sizes for Blacks and Whites were nearly equal and in opposite directions.

Hypothesis 5.2. No interactions of interest were found for this hypothesis.

Hypothesis 6.1. This hypothesis showed no interactions of interest.

Hypothesis 6.2. The group-by-version interaction was not significant although the probability of the obtained chi square was less than .20. The three-way interaction was significant, however. Generally, the effect was larger for Blacks and differences were larger for Version B (key is C). The exceptions, however, were probably the more important contributors to the three-way interaction and may have been sufficient to cancel out some of the overall effect. While Version A tends to be easier for both groups, Version B is easier for one item for each group, but not the same item. One item showed a larger group difference for Version A.

Hypothesis 7.1. This hypothesis had a marginal group-by-version interaction but a highly significant three-way interaction. Since Version A (numbers) is consistently easier, the interaction appears to be due to the change in both the version and group showing the larger effect. Items 1 and 17 had larger differences for Whites and Version A, while Item 27 showed a larger difference Blacks and for Version B.

Hypothesis 7.2. This hypothesis shows only a three-way interaction. Of the four items with relatively large effects two show results nearly opposite and similar in magnitude, essentially canceling out the group-by-version interaction. Although Version A (diagrams) is uniformly easier, items 3 and 6 show larger effects for Blacks and for Version B. The other two items show larger effects for Whites and for Version A.

Analysis by Reading Passage

The complexity of the interactions for Hypotheses 3.1 and 5.1 is particularly interesting if one realizes that the items for these hypotheses, as well as those for Hypothesis 4.1, all relate to the same three reading passages. The content of these reading passages may be one source of the differences in effect of the item manipulations. In order to evaluate this possibility, the five items for a given passage were analyzed separately without consideration for the particular hypothesis that dictated the two item versions. The results are shown in Table 8.

For Passage 2, the group-by-version interaction was not significant, although the three-way interaction was. Examination of the results for the five items related to this passage does not reveal any particular consistency likely to be related to an independent effect of the passage. The three-way interaction appears to be the result of different effects for the different item manipulations associated with the three hypotheses.

Passage 1, on the other hand, shows a strong group-by-version interaction but no three-way interaction. The strength of the interaction effect is particularly interesting since three of the five items for this passage are associated with the nonsignificant Hypothesis 4.1. In contrast with the results for the other passages, the effect of most interest is the larger effect for Whites in all items although Version B also consistently was easier and had the larger effect. The similarity of patterns for all five items may be due in part to influences of the passage content.

Passage 3 has both a small group-by-version effect and a three-way interaction. The interaction appears to be due to differences associated with item version and hence was probably related to the item manipulations associated with the different hypotheses. The effect sizes were, however, consistently larger for Blacks. This is particularly striking since the verbal items generally showed larger effects for Whites. Thus, this consistency may again be associated with the passage content.

Considering these results in light of the complex three-way interactions for Hypotheses 3.1 and 5.1, passage content does appear to

Table 8

Results by Reading Passage

	Group by Version			Group by Version by Item					Odds Ratio		Larger Effect	
Passage	χ^2	df	P	χ^2	df	P	Item	Hyp.	Black	White	Group	Version
1	6.57	1	***	3.52	4		18	3.1	1.20	1.88	Wh	B
							19	3.1	1.89	2.05	-	-
							20	4.1	.99	1.17	Wh	B
							21	4.1	1.03	1.36	Wh	B
							22	4.1	1.42	1.55	Wh	B
2	.27	1		11.38	4	**	23	5.1	.61	.47	-	-
							24	4.1	1.40	.176	Wh	B
							25	5.1	.77	1.19	-	B
							26	3.1	.73	.51	Wh	A
							27	5.1	.86	.85	-	-
3	3.57	1	*	15.14	4	***	28	4.1	1.16	.98	Bl	A
							29	5.1	.69	1.32	Bl	B
							30	5.1	.86	1.07	Bl	3
							31	3.1	1.63	1.38	Bl	B
							32	3.1	.45	.71	Bl	B

* $P < .10$
** $P < .05$
*** $P < .01$

have affected the results for Hypothesis 3.1 to the extent to which that interaction reflected the switch of the larger effect from one group to the other. This hypothesis had two items each related to Passages 1 and 3 that showed opposite effects. The results for Hypothesis 5.1, however, do not appear to be clearly linked with passage content.

Item Bias Analyses

Item bias analyses give even more detail about the results for individual items than the log linear analyses. This information only pertains to comparisons of the two groups within an item version, however; hence, it must be considered a supplementary analysis. These analyses go beyond the log linear analyses presented here in that item properties other than difficulty are considered and the abilities of the two groups are taken into account.

The method chosen assumes that an unbiased item functions in the same way for all examinees, regardless of group membership. Hence, the probability of a correct response at a given level of ability can be determined without reference to any particular sample of examinees. The probability of correct responses over a range of abilities is represented by the item characteristic curve. If an item is unbiased, this curve will be the same for both groups of examinees. Equivalence of the curve is established if each of the three parameters of the mathematical function defining the curve are the same. These three parameters correspond roughly to the item difficulty, discrimination, and probability of correctly guessing at low levels of ability. The method used here finds parameters to be equal if the fit of the model to the data is not significantly different when the parameters are assumed to be equal from when they are permitted to vary. More detail concerning item response theory and the method used are given in Appendix B.

Because of the time and expense involved in using item response theory methods, however, bias analyses were performed only for those hypotheses where results seemed most likely from preliminary analyses of the data. The analyses were performed for verbal Hypotheses 1.1 2.2, 3.1, and 5.1; quantitative Hypotheses 6.2, 7.1, and 7.2; and analytical Hypothesis 3.2. Details of the results are provided in Appendix B.

One might speculate about a finding of bias in relation to the log linear results for particular items. Items biased solely in difficulty should also show group-by-version interactions in the log linear analyses of individual items, with the version in which the bias occurs showing larger effects if the bias favors Whites and smaller effects if the bias favors Blacks. Likewise, where an item is sufficiently difficult to reject a hypothesis of equal probabilities of correctly guessing, the difference should be reflected in the log-linear analyses. If bias exists in discrimination, however, it might be expectec to manifest itself differently, depending on the ability distributions of the two groups. In instances where the group favored is different depending on the ability range, nearly any resı lt, including no effect, might be expected from the log linear analyses.

The items detected in the bias analysis are also likely to be outliers, items functioning somewhat differently from the others, and hence likely to be among those contributing to a three-way interaction. Conversely, a hypothesis without a three-way interaction is unlikely to contain a biased item unless the bias is unassociated with the manipulation. In this case the effect of the bias would be about the same for both items of a pair. The bias would then contribute to the item-by-group interaction, but not the three-way interaction. Unfortunately, the prelimary analyses were focused primarily on the effects of the group-by-version interaction so that Hypothesis 3.2, which does not have a three-way interaction, was selected for these analyses while Hypothesis 2.1, which has a large three-way interaction, was missed.

The results for items where bias was detected are shown in Tables 9 and 10. The three parameters are indicated, where a reflects item discrimation, b reflects item difficulty, and c is the probability of guessing correctly. A more discriminating item has a higher a value, and, as difficulty increases, the value of b increases (items with negative values are easier than items with positive values). Bias is indicated when the values of a given parameter are not the same for Blacks and Whites. Notice that while some of the differences in parameter values seem small, these values provided a significantly better fit to the data than when the values were held equal.

A brief summary of the results and a comparison with the log linear results are presented for each hypothesis as follows:

Hypothesis 1.1. This hypothesis showed one item biased, item 5 Version B (sentence completion). This agrees with the log linear analyses where item 5 was the item that appeared to account for the three-way interaction for this hypothesis. This was the only item where the effects were larger for Blacks.

Hypothesis 2.2. The analyses for this hypothesis yielded two biased items, item 15 Version A (less common meaning) and item 17, both versions. Item 15 also showed a particularly large effect in the log linear analyses. Again, this was the only item of the set where the effect was larger for Blacks although it was also larger ror the unbiased Version B. Item 17 showed only a small effect. This may be because the two item versions appeared to be biased in very similar ways, both in item difficulty and discrimination, suggesting the bias may not have been related to the item manipulation. Both items, however, appeared to be contributing to the three-way interaction for

Table 9

Summary of Items
Identified by Bias Performance

Verbal Hypotheses

				Parameter Value		
Hypothesis	Item	Version	Group	a	b	c
1.1	5	B	Whites	.3	-.5	.2
			Blacks	.3	.5	.2
2.2	15	A	Whites	.5	.7	.25
			Blacks	.5	.7	.15
2.2	17	A	Whites	1.2	1.0	.25
			Blacks	1.0	.3	.15
		B	Whites	1.2	1.0	.20
			Blacks	1.0	.4	.20
3.1	18	A	Whites	.4	.4	.14
			Blacks	.5	.9	.14
		B	Whites	.8	-.3	.14
			Blacks	2.0	-.3	.14
5.1	25	A	Whites	1.3	2.0	.22
			Blacks	1.3	2.0	.23
		B	Whites	.4	2.9	.19
			Blacks	.4	2.9	.15

Table 10

Summary of Items Identified by Bias Procedures

Quantitative Hypotheses

				Parameter Value		
Hypothesis	Item	Version	Group	a	b	c
6.2	13	B	Whites	1.5	.5	.63
			Blacks	.9	-.5	.26
6.2	28	A	Whites	1.2	.02	.12
			Blacks	1.4	.5	.12
		B	Whites	.9	.1	.05
			Blacks	1.5	.3	.06
7.1	17	A	Whites	1.2	- 8	.27
			Blacks	1.0	-.8	.08
		B	Whites	1.5	.7	.14
			Blacks	1.5	.7	.16
7.1	19	A	Whites	6	-2.3	.33
			Blacks	1.3	-2.0	.37
		B	Whites	.5	-.8	.14
			Blacks	.7	-.3	.16
7.1	27	A	Whites	.8	-.3	.33
			Blacks	.8	-.3	.37
7.2	2	B	Whites	.9	-.5	.01
			Blacks	.9	-.5	.14
7.2	3	B	Whites	.4	-.4	.01
			Blacks	.65	.25	.14
7.2	6	B	Whites	.7	-.6	.01
			Blacks	1.6	-.3	.14
7.2	15	A	Whites	1.8	.5	.40
			Blacks	.9	-.1	.16
7.2	23	A	Whites	1.0	-.3	.02
			Blacks	1.7	.2	.07

this hypothesis, as the "more common" word was actually more difficult for Blacks in both instances.

Hypothesis 3.1. This hypothesis showed only one biased item, item 18, both versions. In this instance, however, Version A (one false answer) was biased in difficulty, and to a very small extent, in discrimination. Version B showed only large differences in discrimination. Perhaps this bias was yet another contributor to the complex th.ee-way interaction for this hypothesis.

Hypothesis 3.2. The hypothesis was not found to have any biased items, not surprisingly in view of the nonsignificant three-way interaction in the log linear analyses.

Hypothesis 5.1. Bias was found for this hypothesis only when an assumption that the probability of correctly guessing by lower ability examinees was the same for both groups was found not to fit the data for either version. Since item 25 was the only item of this set difficult enough for sufficient data to be available to reject this hypothesis, this was the item assumed to be biased. This result may not be very meaningful, however, as the probability of a correct response by guessing (the value of c) actually equals the proportion of correct responses for Blacks on Version A and is only slightly lower than the proportion of correct responses for Version B. This suggests that that item was so difficult for Blacks that the responses were nearly random.

Hypothesis 6.2. For this hypothesis, two items were found to be biased, item 13 Version B and item 28, both versions. The result of item 13 showed an unusually large bias in the probability of correctly guessing that favors White examinees. This is consistent with log linear analyses where item 13 is the only item to show Version B (key not "C") easier than Version A for Whites and the only item where the effect for Whites is as large as that for Blacks. Item 28 did not show an effect in the log linear analyses. Again, that result may be associated with aspects of the item that were the same in both versions rather than with the item manipulation.

Hypothesis 7.1. Three items were biased for this hypothesis, items 17 and 19, both versions, and item 27, Version A. The results for item 17 were different for the two versions, however, with only a small difference in the probability of guessing correctly for Version B (symbols). Version A (numbers) showed a small difference in discrimination but a fairly large difference in the probability of guessing correctly. This version might therefore be expected to show a larger effect for Whites the group favored, as it did in the log linear analyses. Item 27 shows a fairly small difference favoring Blacks in the probability of guessing correctly. The log linear analyses show larger effects for Blacks and for Version B, as would be expected if the bias acted to reduce differences on Version A. Item 19, however, did not show an effect in the log linear analyses, again suggesting the bias was not associated with the item manipulation.

Hypothesis 7.2 This hypothesis showed bias in all but one of the items. Items 2, 3, and 6 showed bias in Version B (descriptions) and 15 and 23 in Version A (diagrams). For items 3, 6, 15, and 23, the bias is in all three parameters, making the effects somewhat difficult to predict, although in fact the version showing the bias for each was also the version showing the larger effect in the log linear analyses. Thus, the two sets of analyses are fully consistent for this hypothesis.

Discussion

The overall hypothesis of this study was that differences in the performance of Blacks and Whites could be demonstrated by systematically varying particular aspects of small sets of test items. Seven of the 16 hypotheses showed the group-by-version interaction expected if the item manipulation suggested by each hypotheses had a different effect for the two groups. Three other hypotheses showed a significant three-way interaction in which effects in different directions for different items appeared to cancel each other out so that, when the data were combined across items, the group-by-version interaction was not significant. The overall hypothesis that such differences can be demonstrated would thus appear to be supported by the results of this study.

Supportive evidence for the particular hypothesized source and the proposed rationale for the expected effect was, however, much less clear. These issues will be discussed with regard to each hypothesis below, but a few general observations can also be made. First the picture created in evaluating many of the specific hypotheses was far more complex than had been anticipated. The elements of interest seemed sometimes to interact with content and sometimes with other elements of the item. The existence of three-way interactions and numerous biased items for several of the hypotheses suggests that the manipulated elements were not the only ones acting to produce the observed differences between groups or between versions. These interactions, however should serve as a source of additional hypotheses concerning differential effects on performance.

Another finding was that particularly for verbal items, the performance of Whites tended to differ more between the two item versions than did that of Blacks. This result was contrary to expectation and was further unlikely since sample sizes were such that, by chance alone, changes between versions would be expected to be larger for Blacks for whom the standard error would be larger. On the other hand, the theory suggests that bias may effect either group. That is, the same difference between the observed performance of Whites and Blacks would result regardless of whether a particular item feature made an item easier for Whites or made the item more difficult for Blacks. Since these hypotheses were in several instances drawn from the results of studies where such a distinction between items favoring Whites and those disfavoring Blacks could not be made, a mistake in direction is not surprising. While the results of this study support the hypothesis that such factors do affect the performance of Blacks and Whites differently, the rationale for why such factors differentially affect performance may need to be reexamined in some instances and perhaps discarded. This distinction between whether the effect occurred and why it occurred should be considered in the following discussion of the results for the hypotheses in this study. (In reading these summaries, it may be helpful to refer to the descriptions of the hypotheses in Table 3.)

Hypothesis 1.0

Hypothesis 1.0 was concerned with item format or the structural elements of the form in which a particular concept is to be measured. This hypothesis was evaluated in all three areas: verbal Hypothesis 1.1, quantitative Hypothesis 1.2, and analytical Hypothesis 1.3.

Hypothesis 1.1 compared performance of Blacks and Whites on antonyms versus sentence completions. The rationale was that, to the extent that sentence completion items depended on vocabulary knowledge, the context provided by the sentence would help Blacks more than Whites. The results, however, showed larger effects for Whites and for the sentence completion items. Moreover despite the expectation that sentence completion items would be easier than antonyms, this was the case for only three of the five item pairs. An alternative hypothesis might be that the cues provided in sentence completion items are of a nature to provide more help to White than to Black examinees. For example, the one biased item appeared, on examination, to require an unstated value judgement in order for the intended response to be correct. This suggests that the contextual cues provided by the sentence as well as the level of vocabulary were a source of variation in the differences between groups. This possibility deserves further exploration.

Hypotheses 1.2 and 1.3, for which the item versions were more similar to each other than those used with the verbal items, showed little evidence of differences between groups. Hypothesis 1.2 concerned the quantitative comparison format used in the quantitative sections of the GRE General Test. Although results here were negative, the outcome for Hypothesis 7.2 discussed below sugge ts that this format may interact with the type of content measured. In particular, the four items showing an effect for Hypothesis 7.2 were all geometry items. For Hypothesis 1.2, the item with the largest difference was one of two geometry items in the set of six developed for this hypothesis. This item had a pattern of results very like items 3 and 6 in Hypothesis 7.2. which are quantitative comparison items. Hence, a different conclusion might have been reached with a different set of items. The results from this study should therefore not be considered conclusive evidence that this format has no differential impact on the performance of Black and White examinees.

For Hypothesis 1.0, then, some differential impact was shown for antonyms versus sentence completions, although apparently not for the reasons anticipated. No evidence was found that quantitative comparisons or the Roman numeral format presented special difficulties for Black examinees. Other common item types might be examined in further studies.

Hypothesis 2.0

Hypothesis 2.0 was concerned with the practice of manipulating item difficulty by requiring less common word meanings or by adding or deleting prefixes or suffixes. This practice was expected to have more impact on the performance of Blacks than on that of Whites. Both Hypotheses 2.1 and 2.2 were verbal hypotheses and were evaluated using antonym items.

For Hypothesis 2.1, the stimulus word was the same for both versions with the item difficulty manipulated by the word provided as the correct answer. In most of the items, all other options remained the same. This hypothesis did not show the expected group-by-version interaction, although this was probably due to the opposite effects of two pairs of items revealed in the three-way interaction. The manipulation of difficulty through the particular antonym chosen as the correct response does appear to affect the two groups differently, but results suggest that some elements other than vocabulary are also contributing to the effect on performance. Careful study and analysis of the items with opposing results may suggest what such elements might be, but nothing is readily apparent as an explanation.

Hypothesis 2.2 was originally defined according to the presence or absence of a suffix or prefix with the (unstated) assumption that the version without the suffix or prefix would be easier than the one with. With the particular words chosen, however, the "with" version was sometimes the more common usage. The results were found to be more consistent if redefine in terms of common or uncommon rather than present or absent. revision did not change the major anomaly, however, that the version that was easier for the White group was more difficult for the Blacks for two items. These two items were those identified by the bias analysis.

Hypothesis 2.2 was one of the four hypotheses that showed a strong group-by-version effect in the log linear analyses. The larger between-group differences, however, were associated with the more common rather than the less common usage and the magnitude of the version effect was larger for Whites in two of the five items. A genuine effect appears to have been detected. but the interpretation suggested by the hypothesis as stated does not appear adequate to explain the results. Because the prefix or suffix formed the antonym of the original word, the options were also changed. Characteristics of the distractors and key might be analyzed in more detail for other sources of the observed effect, but again none is immediately obvious.

The results for Hypothesis 2.0 do not suggest that manipulations of vocabulary of the type hypothesized are uniformly more detrimental to the performance of Blacks than to that of Whites. Nonetheless, differential effects from the manipulation of vocabulary in antonym items were observed for reasons that are not yet apparent. These reasons are likely to be subtle, however since antonym items are so spare, with very little information provided beyond the six words of the stimulus and options. Nevertheless, these factors clearly have some impact that can be observed in the results found here and their identification would be a worthy object of further research.

Hypothesis 3.0

Hypothesis 3.0 considered the effect on performance of items where the requirement was to select the one option that is false rather than the more usual choice of the one that is true This hypothesis was evaluated in both the verbal and the analytical sections, Hypotheses 3.1 and 3.2 respectively.

For Hypothesis 3.1, the group-by-version effect was not significant, although the effects of the manipulation were quite different as reflected in a highly significant three-way interaction. Three different patterns of effects were shown. evidentally cancelling each other out in sum. The supplementary analyses suggested that the reading passage interacted with the results in such a way that the change between versions was larger for Whites on passage 1 and for Blacks on passage 3. The first passage was a reading from physics, the second from economics, and the third from biology. The biology passage however concerned skin pigmentation and although no reference was made to humans it may have been of greater interest to Black examinees. One item from Hypothesis 3.1 was found to be biased. It is interesting to note that in the "one-false-answer" version of this item, the option that was the key on the alternate form appeared to have had a stronger attraction for Blacks than for Whites, suggesting that more Blacks than Whites may have misread the stem for this item.

Hypothesis 3.2 was one of the four hypotheses with a strong group-by-version interaction. The differences between groups were smaller, however, for the "one-false-answer" version than for the conventional version in four of the five item pairs, the opposite of the expected effect.

The change from "one true answer" to "one false answer" affected the performance of Blacks and Whites differently. Other factors must also have been operating with these items, however. In the case of the verbal items, the passage content appeared to be one such factor. Further investigation of the effects of this item type is to be encouraged.

Hypothesis 4.0

Hypothesis 4.0 concerns the effects of inferences required of the examinees as part of the task of the item. This was tested with verbal items in Hypothesis 4.1 and with analytical items in Hypotheses 4.2 and 4.3.

Hypothesis 4.1 showed no interactions of interest in the log linear analyses. The inference required by these items, however, was not far removed from the passage, and an effect might have been observed had the inferential leap required been larger.

Hypothesis 4.2 showed a marginally significant group-by-version interaction. The change in version however, was very slight, consisting only of the insertion of the word "most" or "best" into the item stem. The rest of the item was identical. The change had essentially no effect at all for White examinees, but Blacks found the version with "most" or "best" more difficult for four of the five items.

Hypothesis 4.3 did not show any interactions of interest in the log linear analyses.

In general, the support for Hypothesis 4.0 shown here is weak. The particular choices of types of inference to investigate may not have been the best choices however, and the case here may best be left a[illegible]ly unproven, with a suggestion that other types of inference be i[illegible]gated. Differential effects were observed in the items for analytical Hypothesis 4.2, however. Although the effect was small the item manipulation was well defined and probably deserves further attention. Again, the importance of small differences in wording are highlighted by these results. The fact that a change in a single word produced an observable effect is impressive.

Hypothesis 5.0

Hypothesis 5.0 is concerned with the effects of test-wiseness cues on performance of Blacks and Whites. It was evaluated with verbal Hypothesis 5.1 and quantitative Hypothesis 5.2.

Hypothesis 5.1 was one of the four hypotheses to show a strong group-by-version effect with greater differences where test-wiseness cues were present. The results also showed the version with cues tended to be easier for Whites while the version without cues was easier for Blacks. Although the three-way interaction suggests that other factors are operating with this hypothesis, as with others, the observed effects are generally in line with expectations.

Hypothesis 5.2 showed little difference between versions but the cues used were based on option elimination strategies (Kuntz, 1982) that may have been more sophisticated than most examinees actually use.

In verbal items, therefore test wiseness does appear to affect the performance of Blacks and Whites differently. For quantitative items, this difference appears to be negligible. Either quantitative items do not lend themselves as well to these strategies or the strategies chosen to be evaluated are not the ones that matter. The results for Hypothesis 6.2 discussed below, where key placement may affect the useiulaess of a guessing strategy, do suggest that simpler strategies are being used, at least by Black examinees. This hypothesis definitely warrants further elaboration and refinement, probably with greater specificity in the particular test-wiseness strategy expected to be employed.

Hypothesis 6.0

Hypothesis 6.0 concerned the effect of key placement on performance and was evaluated in two specific hypotheses: verbal Hypothesis 6.1 and quantitative Hypothesis 6.2.

Hypothesis 6.1 showed no interactions of interest in the log linear analyses. This may be because the analogy items used were so difficult that even the strongest distractor did not offer the kind of immediate pull required for a person to fail to evaluate the full set of options. Further recent investigation of changes in item difficulty due to key placement found that effects were larger when the key was moved more positions. For example, a change from position A to E produced a larger effect than from A to B (Golub-Smith. 1984). The shifts in this study were most often a single position or only the distractor position was changed and not the key. Further the strongest distractor and the key were usually in adjacent option positions. Another factor was the definition of "strongest." Although effects were not significant, somewhat different patterns resulted for Blacks if the distractor was most popular than if it drew mainly the highest scoring people. Hence, these items may not have provided a good evaluation of the hypothesis.

Hypothesis 6.2 did show a nonsignificant group-by-version interaction, but the three-way interaction was significant in the log linear analyses. Bias analyses found two biased items, one of which appeared to be unrelated to the item manipulation. The other had effects somewhat different from the other items; hence it may have served to attenuate the group-by-version effect as well as contribute to the three-way interaction. If Black examinees are finding the material more difficult than Whites however this may just mean that Blacks are more likely to be guessing at these items.

Hypothesis 6.0 was supported only in part. A differential effect was observed for quantitative items, although Black examinees appeared more often than White to be taking advantage of a guessing strategy of selecting the center option. The results with the analogy items suggest that if this effect occurs, it is with simpler items or where the strongest distractor and the key are further removed in the list of options. If this hypothesis were refined somewhat and explored further, more interesting results might be obtained.

Hypothesis 7.0

Hypothesis 7.0 concerned an abstract or concrete dimension that was evaluated in two quantitative hypotheses one contrasting numbers and symbols (7.1) and the other contrasting diagrams and verbal descriptions (7.2).

The first of these, Hypothesis 7.1, showed a marginally significant group-by-version effect but a very large three-way interaction. Bias was also found in three of the five items. Two of these also had large group-by-version effects in the log linear analyses. In contrasting the three biased with the three unbiased items, an obvious distinction emerged. The three unbiased items were relatively straightforward and simply stated, although they were not the easiest items in the set of six. The other three items were "story problems", requiring the problem to be extracted from a verbal description. In some sense, therefore, a different kind of abstract or concrete dimension was laid down across the intended one. One of the story problems also contained an obvious error that might be made if the problem were not read carefully. This option drew both groups strongly in the symbolic version B, but drew Whites much less strongly than Blacks in the numeric version.

Hypothesis 7.2 had no group-by-version effect, but again had a significant three-way interaction. Both the bias analyses and the log linear analyses indicated effects in opposite directions, however, presumably cancelling each other out when effects were summed across items. Results showed that for the quantitative comparison items, the version without a diagram appeared to be biased and showed a larger effect for Blacks; in the standard format, the item with the diagram appeared to be biased and showed a larger effect for Whites. This is an intriguing finding suggesting the possibility of an interaction with item format. The individual items should be examined in more detail to see if other plausible explanations can be found

The manipulations of the items making up Hypothesis 7.0 definitely had a differential effect on the performance of Blacks and Whites. Of the 15 items identified as biased by the item response theory procedures, eight were items from this hypothesis. The interactions, however, were striking, and the simple expectation stated in Hypothesis 7.0 seems inadequate to account for the results. Other important sources of variation in the differences between groups are clearly at work here. One possible source of such variation is an interaction with the subject matter content. Geometry items, taken without regard to hypothesis, are associated with larger between-group differences than algebra or arithmetic items. However, Hypothesis 7.1 contains only one geometry item and Hypothesis 7.2 only one item that is not geometry, making it nearly impossible to separate the subject matter content from the hypothesis except on the weight of a single item. Format may make a difference, although this seems in conflict with the negative result for Hypothesis 1.2 which looks directly at the format effect. Possibly format differences occur only with certain types of items. For example, the item that has the largest effect for Hypothesis 1.2 is a geometry item. Possible effects of verbiage were again observed with the items in Hypothesis 7.1. The relationships between group performance and the abstract or concrete dimension in quantitative items seems definitely to warrant further investigation.

Conclusions and Recommendations

In this research, the most basic question was whether differential performance of Blacks and Whites could be demonstrated through the manipulation of relatively stable characteristics of test items. The answer to this question appears to be "yes". For several of the hypotheses the effects of the item manipulations were demonstrated to be different for the two groups. Beyond that level of generality however, conclusions were less clear. Although the degree of support for the seven general hypotheses varied, perhaps none was clearly unequivocally supported as stated. Nonetheless, the results are rich with information. A number of directions for further investigation have emerged and more should follow from additional scrutiny of the items for some of the hypotheses.

One issue raised by the results concerns the interpretation of the effects. The hypotheses were stated in terms of how the manipulation was expected to influence the performance of Blacks. The results, however showed a larger effect for Whites in several cases. The question is why should this be? If the difference between groups becomes larger because of the effect on the performance of Whites the suggestion is that something has been done to enhance that performance. In Hypothesis 1.1 the difference between groups is larger for sentence completion items than for antonyms. What cues are we providing in these items that point to the correct response? Which of these are intentional? Are all of these helpful to both groups or are some relatively obscure to Black examinees? If the cues provided are not equally helpful, would avoiding those cues found to be more difficult for Blacks adversely affect the validity of the test for either group?

Similarly, items of the one-false-answer type have been identified as biased in previous studies more often than would be expected by chance. The results here (Hypothesis 3.0) supported a hypothesis of differential effects on performance, yet there was a suggestion that in some cases cues were being provided that were not equally accessible to both Black and White examinees. The interaction with content also bears further exploration. These items and the passages to which they refer need to be examined in much greater detail to see if possible reasons for the difference in effect for different items can be determined.

A related issue is that of test wiseness. The hypothesis specific to test wiseness was Hypothesis 5.0 but the results for the key placment Hypothesis 6.2 are also related to a possible test-wiseness type of strategy. A possibility suggested by these results is that Blacks and Whites differ, not in whether or not test-wiseness strategies are employed, as is commonly assumed, but in which strategies are employed. Further, some test-wiseness cues may be less accessible to Black examinees so that the difficulty of the test-wiseness task is not equivalent for both groups. Again, this should be a fruitful area for further research.

Hypothesis 2.2 also raises some issues. Many might question whether the differences observed as a result of the manipulations of antonym items actually constitutes bias. Vocabulary is, after all, part of what the test purports to measure. Ultimately this is a question of construct validity. Vocabulary is known to be associated with academic performance and has traditionally been included in academic aptitude tests. The use of a prefix or suffix to change an item from a more familiar to a less familiar form, for example, or other strategies such as making a verb from a word commonly used only as a noun, may require a reasoning about words different from that required in other antonym items. Worse, for some examinees the word may be recognized readily in either form, while for other examinees the meaning of the word must be inferred. If an inference must be made more often by members of one group than by those of another, this difference in process might arguably be a source of bias. The relevant question may be whether the validity of the test would be harmed if such items were not used.

Hypothesis 4.0 concerning inferences was clearly supported in only one of three specific hypotheses tested, yet the need for making inferences is implicit in the explanation of the results for the other hypotheses and specific items identified by the item response theory procedure as biased. Possibly the need to control the degree of inference by changing the passage or specific wording of the stem prevented the kind of inferences that would have shown a difference from being used in this study. The challenge here may be in controlling the degree of inference in a meaningful yet measurable way.

At the same time, Hypothesis 4.2 yielded the only result that could be implemented immediately if that were desirable. The use of the word "most" or "best" in the item stem was meant to be clarifying. Instead it appeared to introduce a slight confusion for some Black examinees, perhaps suggesting that something was wanted in these item beyond what was immediately understood. This change, however had virtually no effect for White examinees. Deleting these superlatives would not appear to harm most examinees and might remove a source of confusion for some. If the evidence seems too weak to make a change in practice on the basis of this study alone, it might at least be replicated in anticipation of such an action.

A final recommendation for future study concerns the experimental design. One of the more salient results of this study was the large number of interactions. Rather than try to more carefully control a single element of interest, two or more elements might be varied within the same items in such a way that the effects could be separated statistically. Such a separation was possible to some extent with the reading passages, but the number of items per passage per hypothesis should have been larger for more satisfactory analysis, and the variations to be expected according to passage should have been specified in advance Such an approach would add to the complexities of item preparation, but should greatly increase the usefulness of the results. The quantitative area may be particularly suitable for such an approach since the item elements are probably more easily definable.

What emerges clearly from this study is how little we know about the mechanisms that produce differential performance between Blacks and Whites. Still, the study has demonstrated that item elements exist that are common to some number of items measuring different content that do affect differently the performance of Blacks and Whites. Further investigation of these elements should be fruitful in increasing our knowledge concerning the causes of bias and their eventual remedy.

References

Golub-Smith, M. L. (April 1984). The effects of option scrambling on listening comprehension items: An application of item response theory. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Goodman, L A. (1978). Analyzing qualitative/categorical data. Log linear models and latent-structure analysis. Cambridge. Abt Books.

Hambleton, R. K., & Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 14, 75-96.

Kuntz, P. (1982). Test-wiseness cues in the options of mathematics items. Paper presented at the annual meeting of the American Educational Research Associations, New York.

Lord, F. M. (1977). A study of item bias using item characteristic curve theory. In N. H. Poortinga (Ed.), Basic problems in cross-cultural psychology. Amsterdam: Swits and Vitlinger.

Scheuneman, J. D. (1979). Academy of Certified Social Workers: Report on minority performance. Unpublished research report.

Scheuneman, J. D. (1980). Latent trait theory and item bias. In L. J. Th. van der Kamp, W. F. Langerak, & D. N. M. de Gruijter (Eds.). Psychometrics for Educational Debates. London: John Wiley & Sons.

Scheuneman, J. D. (1981). A new look at bias in aptitude tests. In P. Merrifield (Ed.), Measuring human abilities (New Directions in Testing and Measurement, No. 12). San Francisco: Jossey Bass.

Scheuneman, J. D. (1982). A posteriori analyses of biased items. In R. A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore: Johns Hopkins University Press.

Scheuneman, J. D. (1984). A theoretical framework for the exploration of causes and effects of bias in testing. Educational Psychology 19, 219-225.

Smith, J K. (1982). Converging on correct answers A peculiarity of multiple choice items. Journal of Educational Measurement 19, 211-220.

Thissen, D. (1982a). CULT: Compleat univariate latent trait program. Unpublished manuscript.

Thissen, D. (1982b). Marginal maximum likelihood estimation for the one-parameter logistic model. Psychometrika, 47, 175-186.

Thissen, D., Steinberg, L., & Wainer. H. (1983). On the measurement of item bias: A statistically rigorous methodology using item response theory. Unpublished manuscript.

Appendix A

Log Linear Analyses

Logit models are a class of log linear models that are analogous to multiple regression analyses. In general, regression models are additive models in which the weighted effects are summed to obtain a predicted value of the dependent variable, which is typically a co inuous variable. Log linear models are multiplicative; that is, the predicted value is the product of the weighted effects or the sum of the logarithms of the effects. In the logit models, the dependent variable is the odds of a dichotomous variable. In this study, the odds of making a correct response to an item are predicted from group membership of the examinees, the particular item pair, and the two item versions that make up the pair. The advantages of logit models over standard multiple regression procedures for the analysis of categorical data are discussed in Goodman (1978).

For each set of items corresponding to the 16 specific hypotheses, a series of analyses were performed in which successive models were fitted to the data. To get some sense of the meaning of the different effects included in the models, consider the following example for one item. (For ease of conceptualization, percent of correct responses will be used in this example rather than the odds ratio--the number correct divided by the number incorrect--actually used in the computation.)

	Black	White	
Version A	p_{11}	p_{12}	$p_{1.}$
Version B	p_{21}	p_{22}	$p_{2.}$
	$p_{.1}$	$p_{.2}$	

The objective is to predict the performance of Blacks and Whites. If group membership has no effect on performance, the four cell values, p_{ij}, will be predictable from only the information on pnrformance on the two versions, $p_{1.}$ and $p_{2.}$, the difficulty of each version combining the two groups. Similarly if the item ·ersion has no effect, the cell values, p_{ij}, will be predictable from only the performance of the two groups combining data from the two versions, $p_{.1}$ and $p_{.2}$. If both group and version are important and the effect of the item version is the same for the two groups, the cell values will be adequately predicted from the four "marginal" values, $p_{1.}$, $p_{2.}$, $p_{.1}$ and $p_{.2}$. If, however, this model with two main effects (group and version) does not fit the data, it must be assumed that an interaction between group and version exists; that is, the effect of the item versions is not the same for the two groups. Notice that the model that includes the interaction predicts the cell values from the p_{ij}, which are the cell values. This is called the "saturated model", which has no degrees of freedom since the expected values in all cells will equal the obtained values and no possibility of variation exists.

This simple example can be extended to the case where there are several items. Here, if no effect exists for different item pairs, the cell values can be predicted from the item marginals, that is, from the mean performance across all the items of one version for one group. This would mean that there was no item main effect. Likewise, the difference between versions may be the same for all item pairs, or the group performance differences may remain the same for the set of items. In these instances, there would be no item-by-version or group-by-item interaction, respectively.

In this study, the interactions of interest are the two-way, group-by-version interaction and the three-way, group-by-version-by-item interaction. The two-way interaction indicates that the two item versions are differentially difficult; that is, the difference in difficulty between versions is larger for one group than for the other. Hence knowing the difference in difficulty for Versions A and B and the difference in performance for Blacks and Whites is not sufficient to predict the difficulty of each version for each group. The three-way interaction further specifies that these differences between groups and versions are not the same for the different item pairs in a set. If the three-way interaction occurs without a two-way interaction, one of two explanations is most likely. Either, one, the differential effect occurs for only some items or, two, effects are in different directions, that is, effects are larger for Black examinees for some items and for White examinees for others. (See Table 7 on page 30.)

The models analyzed in this study are shown in Table 11. The saturated model is M_0 which includes all effects--the three main effects, the three two-way interactions, and the three-way interaction. Model M_1 includes only the main effects. If this model fits the data, the various interactions are not required. Models M_2, M_3, and M_4 evaluate the effect of each of the separate main effects. The difference between the chi square statistic for model M_1 and that for one of these models tests the significance of the effect not specified. For example, if the fit of M_1 is significantly better than that of M_2, the effect of the item pair (the effect not specified) provides a significant improvement in the prediction of performance. In models M_5, M_6, and M_7, the separate interactions are evaluated.

Table 11
Log Linear Models Estimated

Model	Effects Included*
M_0	All (G, V, I, GV, GI, VI, GVI)
M_1	Main Effects (G, V, I)
M_2	G, V
M_3	G, I
M_4	V, I
M_5	G, V, I, GV
M_6	G, V, I, GI
M_7	G, V, I, VI
M_8	G, V, I, GV, GI, VI

* G group main effect
V item version main effect
I item pair main effect
GV group-by-version interaction
GI group-by-item interaction
VI item-by-version interaction
GVI three-way interaction

Again, the significance of the interaction specified is tested by taking the difference between the result for one of these hypotheses and that for model M_1. The final model, M_8, tests for the three-way interaction. If this model does not fit the data, the three main effects and the three two-way interactions together are inadequate to explain the results. Hence, the three-way interaction, a different effect of item version for the two groups in different item pairs, is significant. A more formal and detailed presentation of the log linear logit models is also provided by Goodman (1978).

The data were analyzed using the SPSS-X log linear program. The results are given in Table 12. The first column gives the reference value chi square for model M_1, where the chi square value is the likelihood ratio chi square. The last column for the three-way interaction is the result obtained for model M_8. The other results are for the differences between the result for the relevant model and the reference value for M_1 with the degrees of freedom for the difference. Unless otherwise indicated, the probability of the obtained chi square is less than .01.

For none of the hypotheses did the model consisting of the three main effects only (model M_1) fit the data. For three hypotheses, 1.2, 4.2 and 6.1, however, the version main effect was not significant. For Hypothesis 1.2, significant interactions of the item version with the item pair suggested that, while on the average the item version had no effect, the effect was different and apparently opposite for different item pairs. Hypothesis 4.2 showed a marginally significant group-by-version interaction.

For three of the hypotheses, one of the models with a single interaction was found to fit the data. For Hypotheses 6.1 and 4.2, the model consisting of main effects and the group-by-item interaction (model M_6) was appropriate.

Table 12
Results of Log Linear Analyses
All Effects

Hypothesis	All Main Effects			Individual Main Effects				Two-Way Interactions				Three-Way Interactions		
	χ^2	df	P		χ^2	df	P		χ^2	df	P	χ^2	df	P
1.1	2176.5	13		I	1805	4		GV	11.3	1		11.8	4	**
				V	149	1		GI	25.8	4				
				G	460	1		VI	2152.4	4				
1.2	208.1	16		I	4934	5		GV	0.5	1	ns	2.7	5	ns
				V	0.07	1	ns	GI	77.9	5				
				G	743	1		VI	126.2	5				
1.3	172.7	13		I	3442	4		GV	0.1	1	ns	3.4	4	ns
				V	4.4	1	**	GI	46.9	4				
				G	322	1		VI	126.2	4				
2.1	1036.8	19		I	3898	6		GV	3.5	1	*	28.9	6	
				V	1369	1		GI	52.9	6				
				G	828	1		VI	968.6	6				
2.2	103.9	13		I	1657	4		GV	12.8	1		8.0	4	*
				V	185	1		GI	38.8	4				
				G	372	1		VI	40.7	4				
3.1	424.7	13		I	958	4		GV	.5	1	ns	14.8	4	
				V	18	1		GI	44.5	4				
				G	430	1		VI	367.0	4				

* P < .10
** P < .05
ns P > .10

Al' other effects are significant beyond the .01 level.

Table 12 (cont.)
Results of Log Linear Analyses
All Effects

Hypothesis	All Main Effects			Individual Main Effects				Two-Way Interactions				Three-Way Interactions		
	χ^2	df	P		χ^2	df	P		χ^2	df	P	χ^2	df	
3.2	117.3	13		I	1006	4		GV	8.3	1		6.8	4	n
				V	63	1		GI	9.3	4	*			
				G	323	1		VI	93.7	4				
4.1	52.6	13		I	2063	4		GV	1.2	1	ns	3.8	4	n
				V	76	1		GI	6.1	4	ns			
				G	525	1		VI	41.2	4				
4.2	22.5	13	**	I	1347	4		GV	3.3	1	*	1.7	4	n
				V	1.9	1	ns	GI	17.4	4				
				G	506	1		VI	0.1	4	ns			
4.3	113.1	13		I	5840	4		GV	1.4	1	ns	3.9	4	n
				V	67	1		GI	68.3	4				
				G	327	1		VI	39.7	4				
5.1	144.8	13		I	2967	4		GV	8.7	1		14.0	4	
				V	9.1	1		GI	21.2	4				
				G	302	1		VI	103.3	4				
5.2	95.4	16		I	2880	5		GV	0.9	1	ns	0.9	5	n
				V	21	1		GI	57.8	5				
				G	952	1		VI	36.8	5				

* $P < .10$
** $P < .05$
ns $P > .10$

All other effects are significant beyond the .01 level.

Table 12 (cont.)
Results of Log Linear Analyses
All Effects

Hypothesis	All Main Effects			Individual Main Effects				Two-Way Interactions				Three-Way Interactions	
	χ^2	df	P		χ^2	df	P		χ^2	df	P	χ^2	df
6.1	124.8	28		I	1252	9		GV	0.7	1	ns	8.3	9
				V	2.4	1	ns	GI	106.4	9			
				G	772	1		VI	9.4	9	ns		
6.2	102.3	16		I	1996	5		GV	2.5	1	ns	12.4	5
				V	18	1		GI	55.0	5			
				G	815	1		VI	30.5	5			
7.1	321.7	16		I	841	5		GV	2.8	1	*	27.3	5
				V	1747	1		GI	7.4	5	ns		
				G	841	1		VI	274.5	5			
7.2	374.5	16		I	1347	5		GV	0.9	1	ns	11.4	5
				V	98	1		GI	49.1	5			
				G	957	1		VI	307.3	5			

* P < .10
** P < .05 All other effects are significant beyond the .01 level.
ns P > .10

The chi square values were 18.4 with 19 degrees of freedom and 5.1 with 9 degrees of freedom for the two hypotheses respectively. (The results for Hypothesis 4.2 are discussed further in the body of the paper.) For Hypothesis 4.1, none of the interactions with group membership was significant, so that model M_7, with the three main effects and the item-by-version interaction, was found to fit the data. The chi square value for this model was 11.3 with 9 degrees of freedom.

Three Hypotheses, 1.3, 4.3, and 5.2, were found to have significant group-by-item and item-by-version interactions. Although a model with these effects was not specifically evaluated, the nonsignificant group-by-version interaction and three-way interaction suggest such a model would fit the data.

The remaining seven hypotheses show either the expected group-by-version interaction or the three-way group-by-version-by-item interaction. The results for these hypotheses are discussed in more detail in the body of the paper.

APPENDIX B

Item Bias Analyses

In this method, an unbiased item is defined as one where all examinees of a given level of ability have the same probability of a correct reponse regardless of their group membership; or, in the terms of item response theory, an unbiased item is one for which the item characteristic curves for two groups, and hence all three parameters that define those curves, are the same.

Basic Concepts of Item Response Theory

The item characteristic curve represents the probability of a correct response as a function of a unidimensional ability, θ. The curve is an ogive, or s-shaped curve, that begins as a line along or parallel to the axis on the left and rises to a value of 1 on the right where it again becomes parallel to the axis. The dimension from left to right represents increasing levels of ability; the height of the curve represents the probability of a correct response. At any given ability level, that is, at a given point along the axis, the height of the curve at that point is the probability that a person with that ability will get the item correct. Examples of item characteristic curves are shown in Figures 1-6.

The exact shape of the item characteristic curve is mathematically defined by a logistic function with three parameters: a, b, and c. The b parameter is the inflection point of the curve and represents the difficulty of the item. If the lower asymptote (the portion of the curve on the left before it begins to rise) is zero, b will be the ability level where the probability of a correct response is .50. The _a_ parameter is the slope of the curve at that point and represents the discrimination of the item. The third parameter, _c_, is the height of the lower asymptote and represents the probability of a correct response essentially in the absence of knowledge. For multiple choice items this asymptote is usually assumed to be non-zero and is often referred to as a guessing parameter. (For additional discussion of the basics of item response theory, see Hambleton & Cook, 1977.)

The Method

Item response theory (IRT) has been seen as particularly useful for studies of item bias because, in theory, the item characteristic curve is not dependent on the distribution of ability in the sample used to determine the parameters. Hence, if the parameters of an item characteristic curve are estimated separately for two samples drawn from the same population, the resultant curves should be the same (except for scale) even though the samples may differ in the distribution of abilities within them. If the curves are not the same, the conclusion of bias may be warranted. In practice, however, parameters may be more difficult to estimate in some samples than in others, and the estimation itself is not perfectly precise. Hence, some indicator of the degree of agreement between the two curves is needed. None of the indicators suggested in the literature to date has been found fully satisfactory. Perhaps more important, however, is the requirement for very large sample sizes in order to estimate the parameters well. In many applications, too few minority examinees are available.

A new method developed by Thissen for using item response theory provides a different approach that resolves some of these problems. The method is based on a marginal maximum likelihood estimation procedure (Thissen, 1982b). Its major difference from other procedures is the ability to impose equality constraints and to evaluate the significance of the difference between the fit of the model to the data with and without these constraints. The method is implemented through use of the CULT computer program (Thissen, 1982a).

In this study, the first step in the procedure for the analysis of a given hypothesis was to obtain ability estimates for all examinees. This was done on a selected set of items from the operational sections of the test as well as the experimental items, constraining the item parameters to be the same for both groups. In subsequent analyses, the items from the operational test were always assumed to have the same parameters, providing an anchor test for the analyses. The data were then refit, releasing certain of the equality constraints in a systematic fashion. At each step of the process, the fit was evaluated with a negative log likelihood statistic (G^2). Twice the difference between the G^2 for two successive analyses was examined as a test of significance. The degrees of freedom for this statistic were the difference between the numbers of parameters estimated in the more and less constrained models. If the difference G^2 is not significant, the unconstrained parameters cannot be assumed to be different from one another (Thissen, Steinberg, & Wainer, 1983).

Item Bias Results

Because of the cost and time requirements of the IRT analyses, these procedures were applied only for items from those hypotheses where preliminary results suggested outliers might be detected. The item bias analyses were performed for the items in verbal Hypotheses 1.1, 2.2, 3.1, and 5.1;

quantitative Hypotheses 7.1, 7.2, and 6.2; and analytic Hypothesis 3.2.

In these analyses, ability estimates for examinees were developed from items in the operational sections, which were taken by all examinees. Only items of the same type as the experimental items were used for this purpose in order to increase the likelihood that a single ability was being measured. That is, for Hypotheses 1.1 and 2.2, only antonym items were used to estimate the ability parameters; only reading comprehension items were used for Hypotheses 3.1 and 5.1. With the math items, parameters for quantitative comparison items were estimated separately from the items of the more conventional type. For Hypotheses 7.1 and 7.2, both formats had been used so two separate sets of analyses were performed within each of the hypotheses based on the different ability estimates. Hypothesis 6.2 items were evaluated in terms of ability based only on the conventional item type.

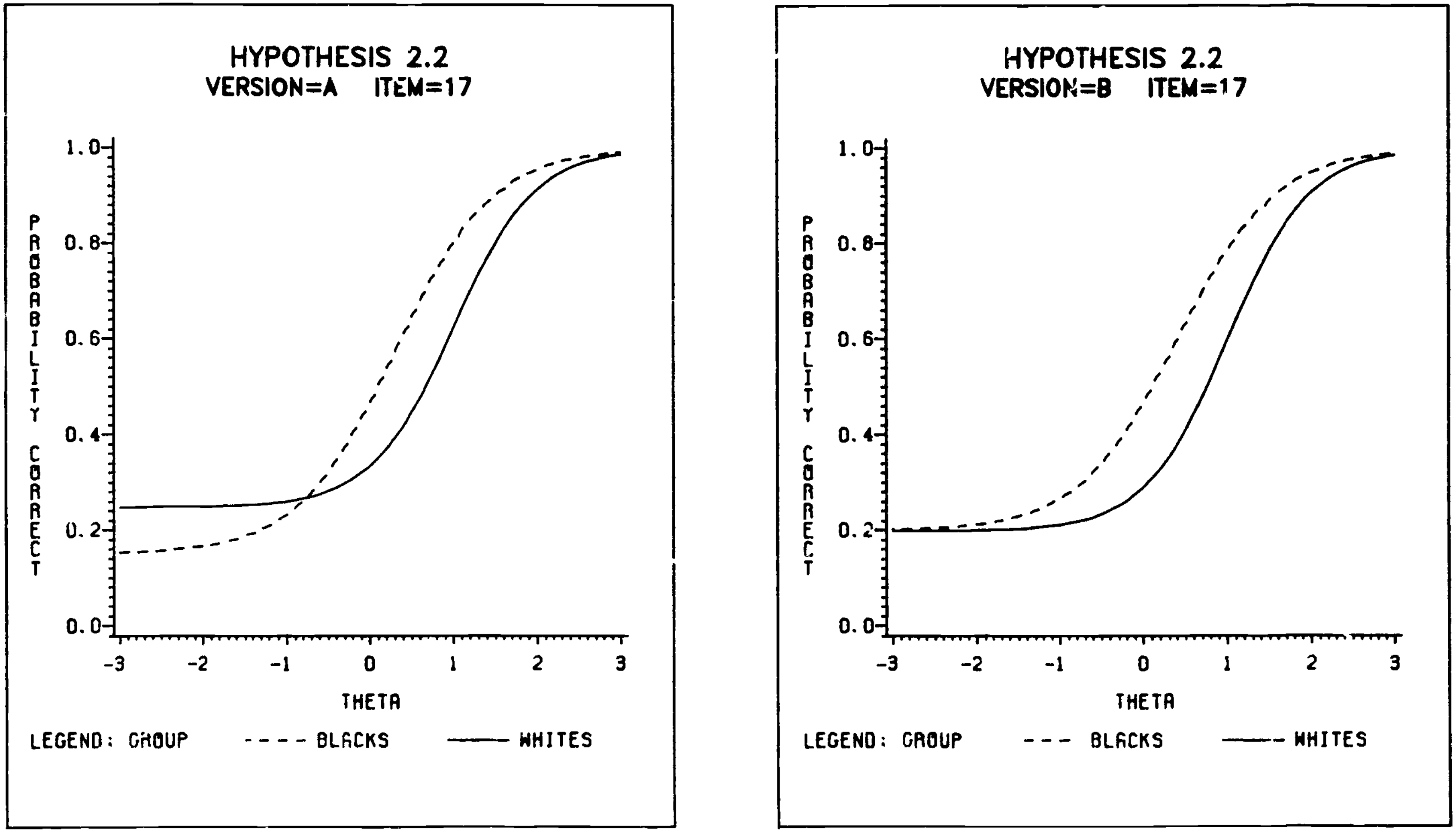
After the analysis of Hypothesis 1. it became apparent that the cost could be greatly reduced by using a smaller sample of Whites. Therefore, a series of runs were performed to determine how far the sample size could be reduced without undue loss of power. The final sample for the remaining hypotheses consisted of a one-third spaced sample of the Whites used in the analyses described earlier. This final sample was approximately twice che size of the Black sample.

For Hypothesis 1.1, the $\underline{c}$ parameters were estimated first, assuming the value of the parameter would be the same for Blacks and Whites for all five items of a given item version. The value of the $\underline{c}$ parameter was, however, allowed to be different for the antonym version and the sentence completion version. Acceptance of the same $\underline{c}$ parameter value for all items in a set only means that insufficient data exist to distinguish different c's. Estimating a common value helps stabilize the estimate of other parameters and facilitates interpretation. The $\underline{a}$ and $\underline{b}$ parameters were then estimated for each item. The data were found to be consistent with equal parameters for Blacks and Whites for all of the antonym items and for four of the five sentence completion items. For the remaining item, the difficulty parameters were found to be significantly different, with the item more difficult for Blacks. The verbal items that showed parameter differences are summarized in Table 9, appearing in the body of this paper.

Hypothesis 2.2 was found to be somewhat more complex. For these items the hypothesis of equal $\underline{c}$ parameters for Blacks and Whites could not be confirmed for one form of the test. The five items were estimated to have a $\underline{c}$ of .25 for Whites and .15 for Blacks. For relatively easy items, however. there was little data available in the ability regions where the asymptotes occur. That is, for relatively easy items, even the least able examinees were responding at a level well above chance. This hypothesis of equal $\underline{c}$'s was only apt to be rejected, therefore, where items were sufficiently difficult for data to be available in this range. The observed difference in $\underline{c}$ parameters was thus probably the result of the most difficult item, 17, and perhaps item 15. Item 17, however, also showed differences in both the $\underline{a}$ and $\underline{b}$ parameters in both item versions. In Version B (more common meaning), this item favored Blacks, but in the alternate version, the item characteristic curves crossed; hence, the relative performance of the two groups would be dependent on their respective abilities. The item characteristic curves for the two versions of this item are shown in Figure 1.

The items for Hypothesis 3.1 again appeared to be unbiased except for item 18. The version with one false answer (Version A) was clearly biased in favor of Whites, with a moderate difference in difficulty and a small difference in slope. The other version did not differ in difficulty, but had a large dif-

Figure 1
Item Characteristic Curves
Hypothesis 2.2

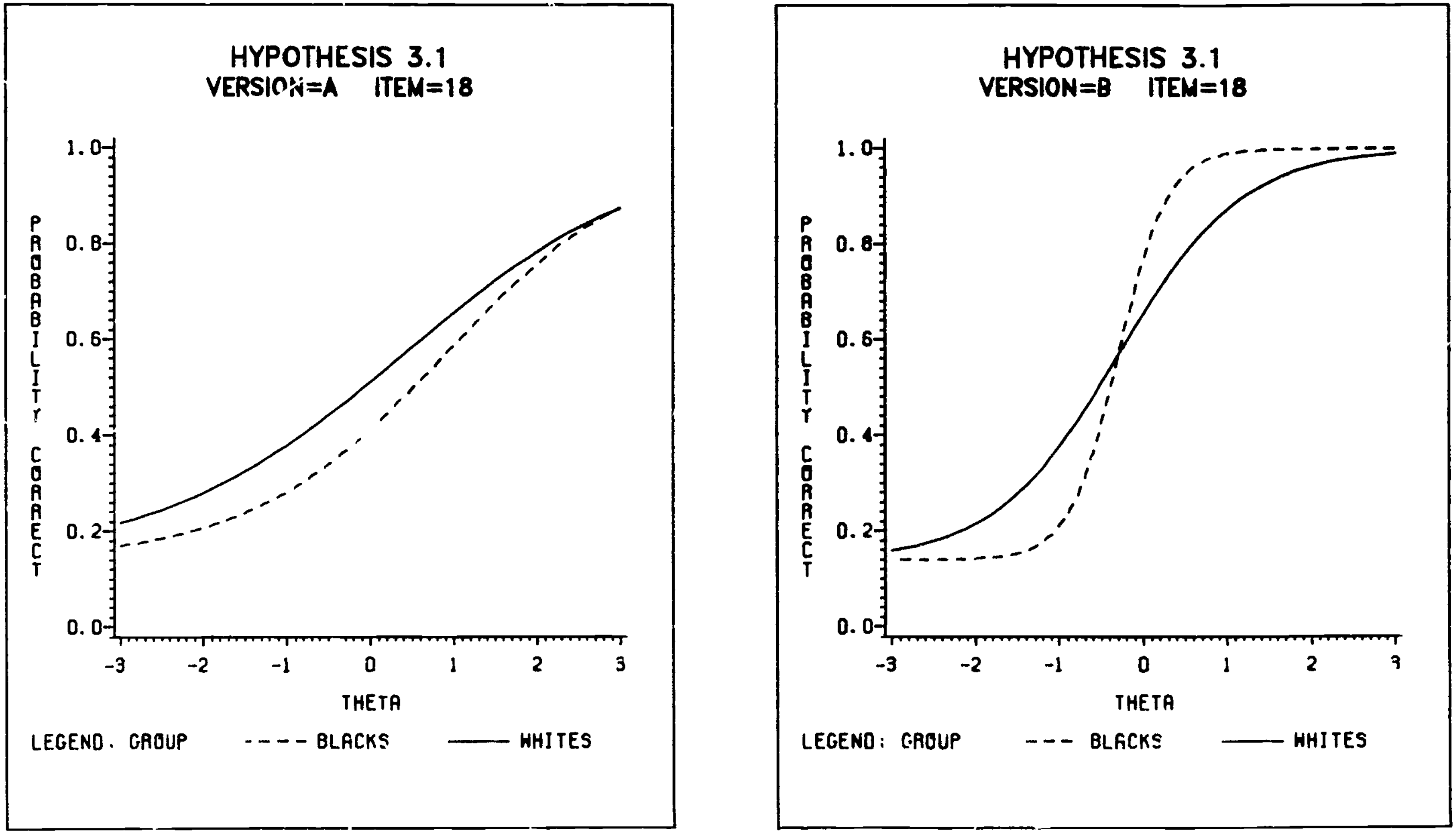


ference in slope, so that again the group favored changed with the level of ability. The item characteristic curves for the two versions of this item are shown in Figure 2.

For Hypothesis 5.1, none of the items showed significant differences in the *a* and *b* parameters. However, the lower asymptotes were different for both forms of the test. In this case, only item 25 was nearly difficult enough to be producing this result. In one version (test-wiseness cues absent), a small but significant difference existed that favored the Black group; in the other, a slightly larger difference favored the Whites. The change from one item to the other was larger for Blacks. It should be noted, however, that the significance of differences between versions was not tested, and that the .23 value of *c* for Blacks in Version A was equal to the obtained percent correct for that group.

For quantitative hypothesis 6.2, *c*'s were not constrained to be equal for all items of a given version but were estimated separately for each item. Only two items showed evidence of bias, Version B (key not C) of item 13 and both versions of item 28. Item 13 appeared to be both more difficult and more discriminating and had a higher asymptote for Whites. The *c* parameter for Whites was .63, however, an extremely high value for which no reasonable explanation is apparent in light of the other data. The three parameters for the Black group in Version B were very similar to those for both groups in the unbiased Version A, where *a* was .6, *b* was -.5, and *c*, .27, suggesting that the change in option placement largely affected the performance of the White examinees. Parameter values for the Version B item are shown in Table 10 in the body of the paper. Both versions of item 28 were both more difficult and more discriminating for Blacks. Version B had an asymptote slightly higher for Blacks. Again the obtained parameter values are given in Table 10. Item characteristic curves for these items are shown in Figure 3.

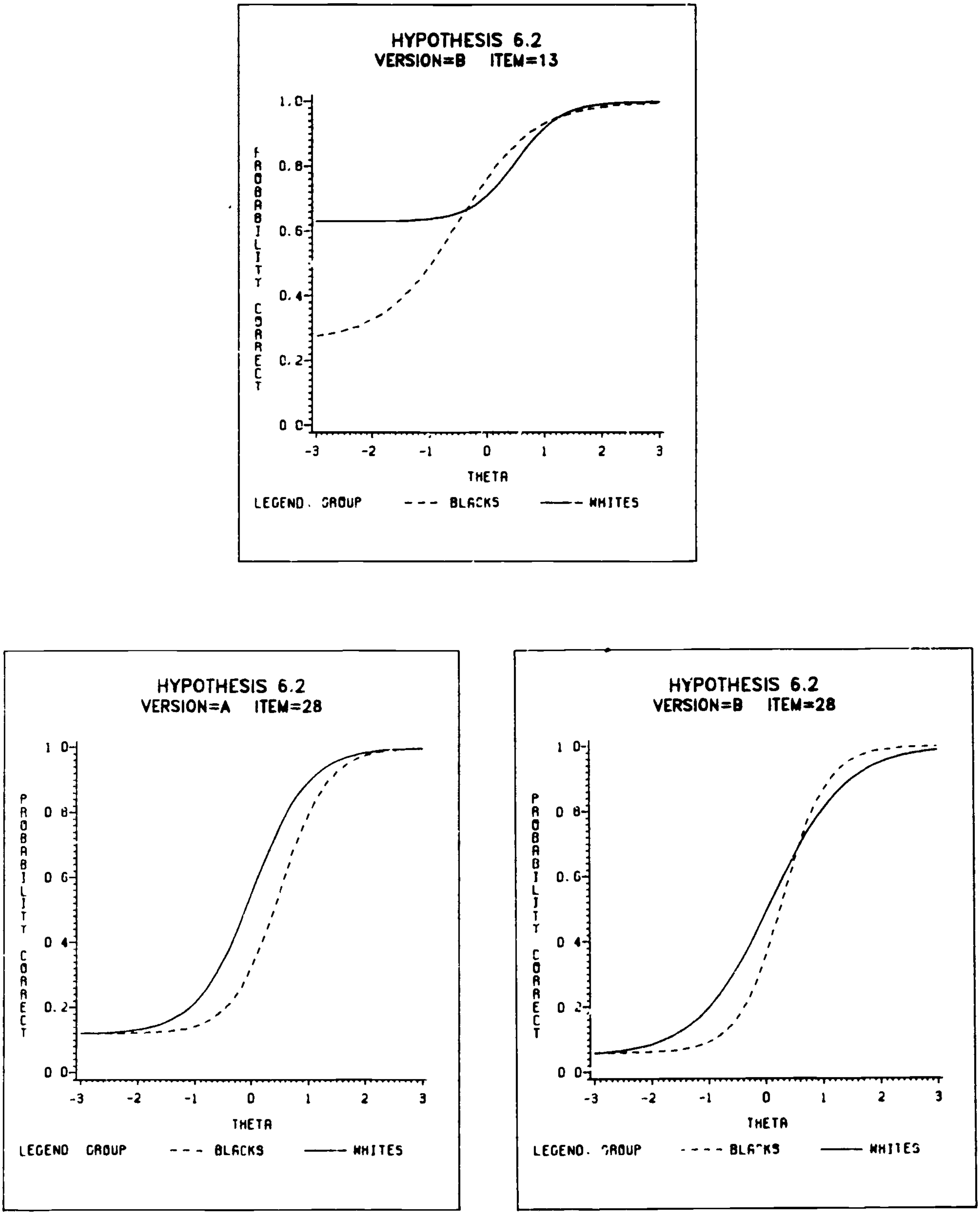
Figure 2
Item Characteristic Curves
Hypothesis 3.1



HYPOTHESIS 3.1
VERSION=A ITEM=18

HYPOTHESIS 3.1
VERSION=B ITEM=18

Figure 3
Item Characteristic Curves
Hypothesis 6.2

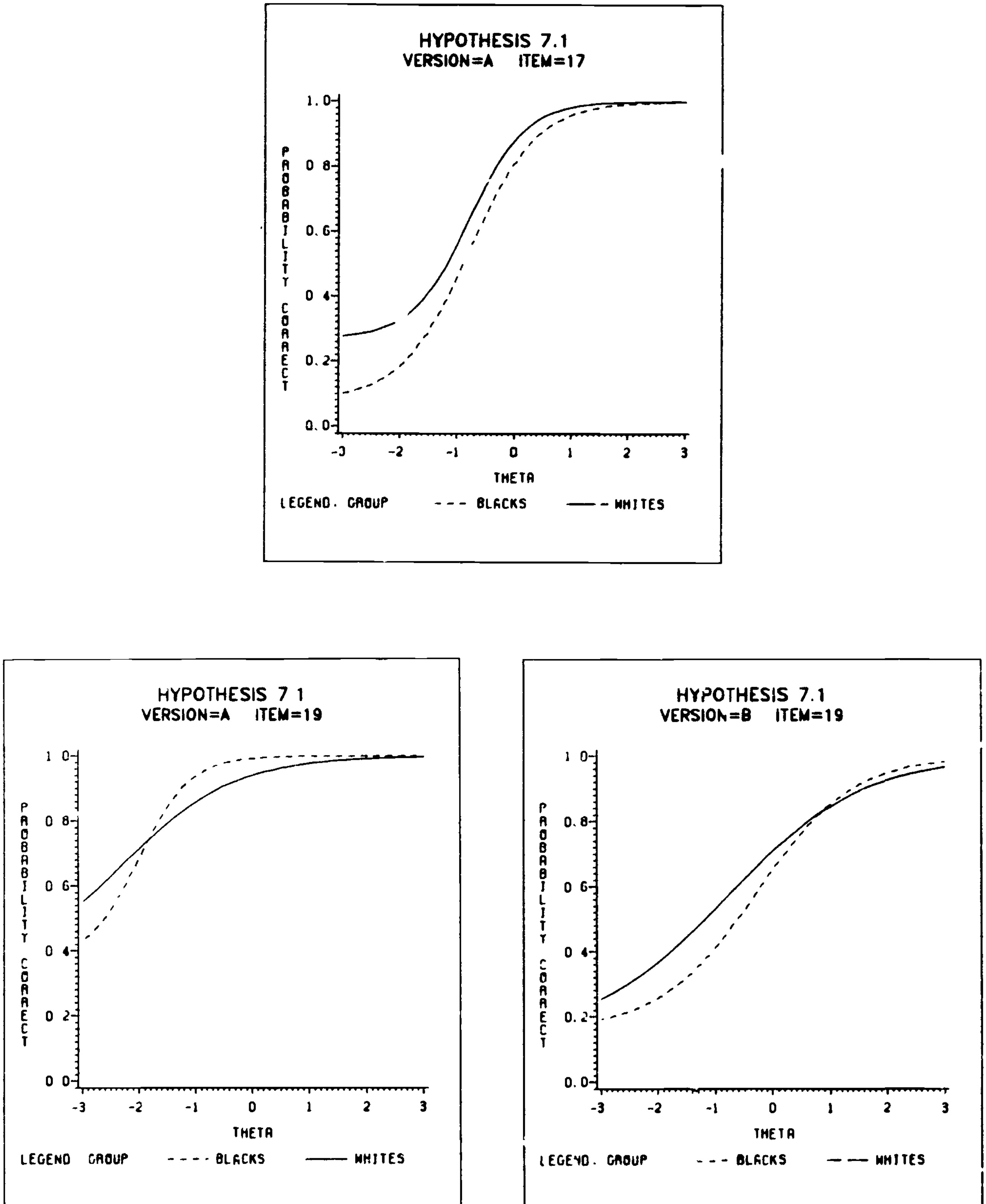


The two quantitative comparison items for Hypothesis 7.1 (items 1 and 4) were analyzed separately from the other four. No evidence of bias was seen here. For the other four items, the hypothesis that the *c* parameters were the same for both groups was rejected within both item versions. Further, item 17 in Version A (numbers) could not be shown to have the same *c* parameters as the other items in that form. The asymptotes were lower for both groups than for the other items of the same version. though higher for Whites and lower for Blacks than in the alternate item 17. It also showed small differences in discrimination.

The other items showed a higher asymptote for Blacks than for Whites, with much higher *c*'s in Version A (with numbers) than in Version B (with symbols). It is not entirely clear which items produced sufficient data to reject the hypothesis of equal *c*'s, but the most likely candidates appear to be item 27 in Version A and item 17 in Version B, neither of which showed bias in the other parameters. Item 19 showed differences in both the *a* and *b* parameters in both versions with the item both more difficult and more discriminating for Blacks. Parameter values for the biased items are given in Table 10. Item characteristic curves for item 17 Version A and item 19 Versions A and B are shown in Figure 4.

Four of the six items for Hypothesis 7.2 (items 2, 3, 5, and 6) were quantitative comparisons and were analyzed together. For Version A (with diagrams), all four items were found to be unbiased. For Version B (with verbal description), the *c* parameters were estimated for the four items together and were found to be higher for Blacks (.14) than for Whites (.01), in contrast with a of .16 for both groups on Version A. The items responsible for this result seem most likely to have been item 2 and/or item 6. Item 2 did not show differences on the other parameters, but item 6 was again both more difficult and more discriminating for Blacks. This result was

Figure 4
Item Characteristic Curves
Hypothesis 7.1



also found for item 3. Item characteristic curves for all four Version B items are shown in Figure 5.

For the two remaining items in Hypothesis 7.2, the c parameters were estimated separately within each version. The Version B (description) items used a common c estimate for Whites and Blacks and appeared to be otherwise unbiased. In Version A, item 15 was both more difficult and more discriminating for Whites and also had a much higher asymptote for Whites, .40, in contrast to .16 for Blacks and .21 for the B version. Item 23 was more difficult and discriminating for Blacks with a somewhat higher asymptote. The item characteristic curves for these items are shown in Figure 6. The parameter values for all biased items for Hypothesis 7.2 are given in Table 10.

Analyses for the analytic Hypothesis 3.2 were based on ability estimates using all items from the operational analytic sections, since item types used were quite similar for all items. No significant differences between parameters for these items were found.

Figure 5
Item Characteristic Curves
Hypothesis 7.2
Quantitative Comparisons

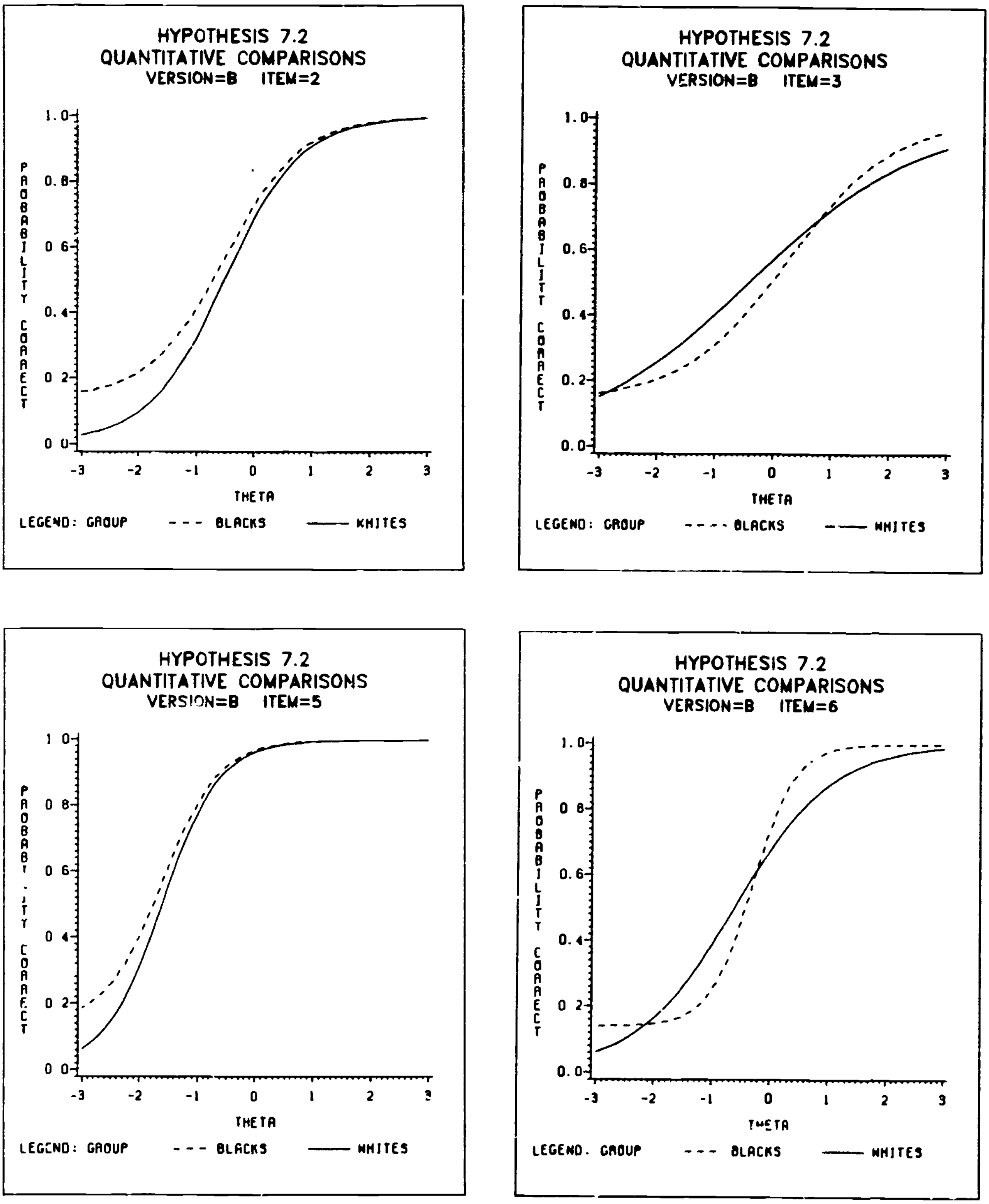
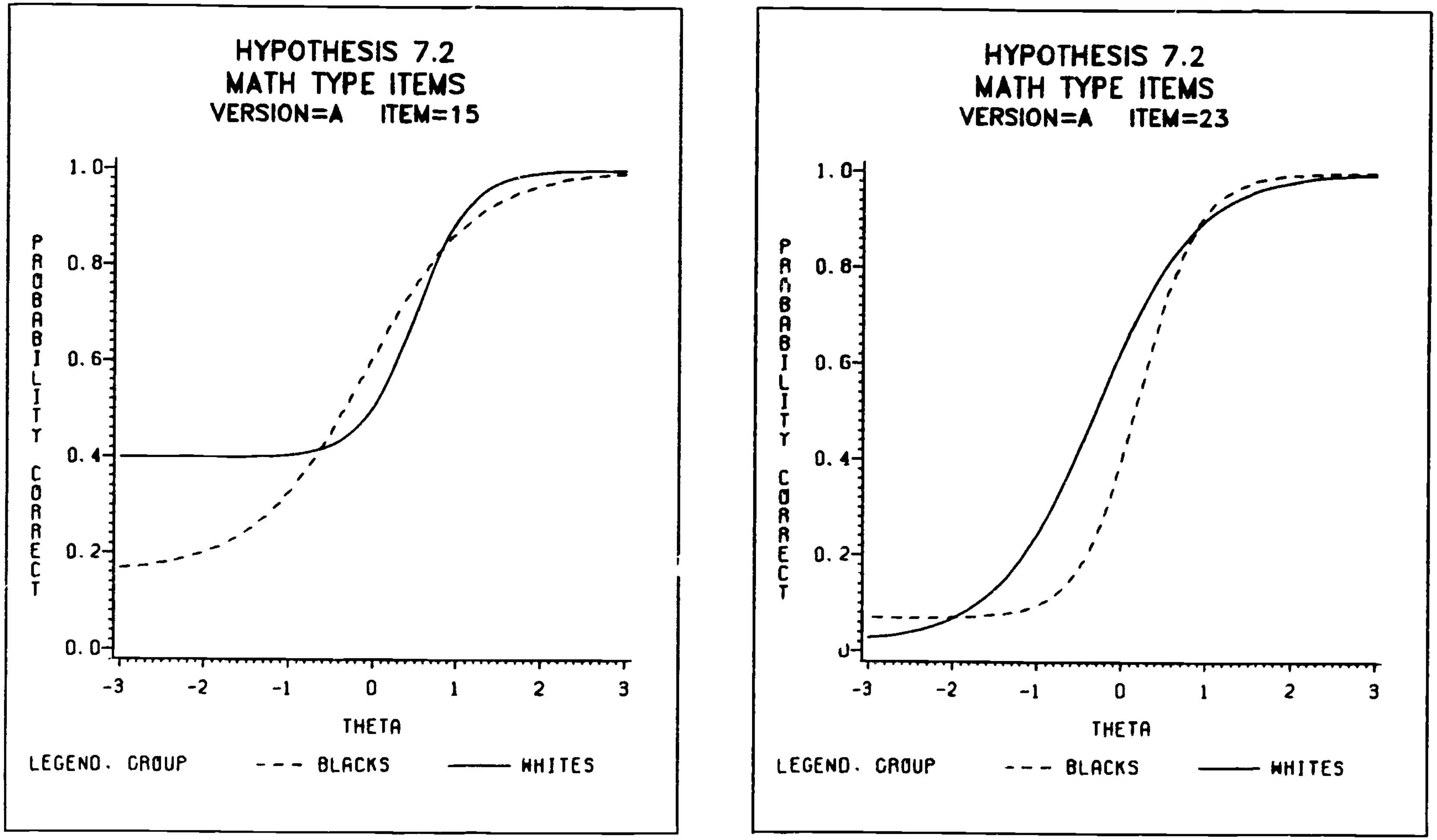


Figure 6
Item Characteristic Curves
Hypothesis 7.2
Standard Math Items


HYPOTHESIS 7.2
MATH TYPE ITEMS
VERSION=A ITEM=15
PROBABILITY CORRECT
THETA
LEGEND. GROUP
--- BLACKS
—— WHITES
HYPOTHESIS 7.2
MATH TYPE ITEMS
VERSION=A ITEM=23
PROBABILITY CORRECT
THETA
LEGEND. GROUP
--- BLACKS
—— WHITES