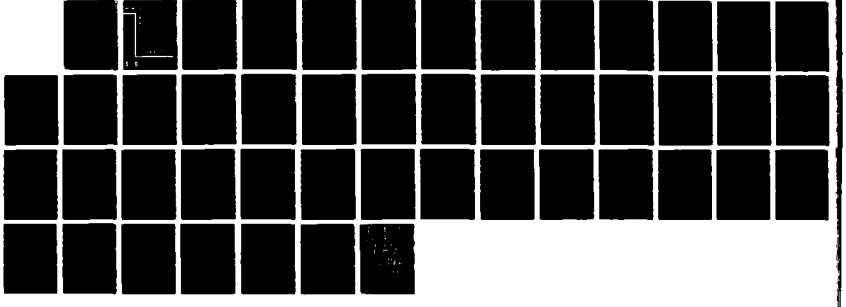
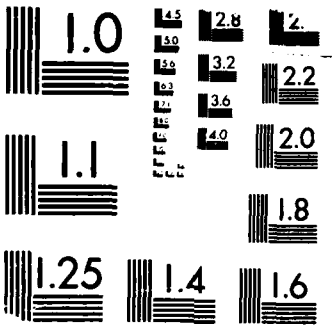
AD-A193 693 ARMED SERVICES VOCATIONAL APTITUDE BATTERY: DIFFERENTIAL ITEM FUNCTIONING..(U) UNIVERSAL ENERGY SYSTEMS INC DAYTON OH R L LINN ET AL. APR 88 1/1

UNCLASSIFIED AFHRL-TR-87-45 F41689-84-D-0002 F/G 5/9 NL





AFHRL-TR-87-45

(2)

AIR FORCE HUMAN RESOURCES LABORATORY

ARMED SERVICES VOCATIONAL APTITUDE BATTERY: DIFFERENTIAL ITEM FUNCTIONING ON THE HIGH SCHOOL FORM

Robert L. Linn
C. Nicholas Hastings
Pei-Hua Gillian Hu
Katherine E. Ryan

Universal Energy Systems, Inc.
4401 Dayton-Xenia Road
Dayton, Ohio 45432

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

April 1988
Final Technical Paper for Period October 1985 - May 1987

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This report has been reviewed and is approved for publication.

LINDA T. CURRAN
Contract Monitor

WILLIAM E. ALLEY, Technical Director
Manpower and Personnel Division

HAROLD G. JENSEN, Colonel, USAF
Commander

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5 MONITORING ORGANIZATION REPORT NUMBER(S) AFHRL-TR-87-45
6a. NAME OF PERFORMING ORGANIZATION Universal Energy Systems, Inc. 6b OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION Manpower and Personnel Division
6c. ADDRESS (City, State, and ZIP Code) 4401 Dayton-Xenia Road Dayton, Ohio 45432	7b. ADDRESS (City, State, and ZIP Code) Air Force Human Resources Laboratory Brooks Air Force Base, Texas 78235-5601
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Air Force Human Resources Laboratory 8b OFFICE SYMBOL (If applicable) HQ AFHRL	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER F41689-84-0-0002 Subcontract 744-014-001
8c. ADDRESS (City, State, and ZIP Code) Brooks Air Force Base, Texas 78235-5601	10 SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO	WORK UNIT ACCESSION NO
62703F	7719	18	40

11. TITLE (Include Security Classification)
Armed Services Vocational Aptitude Battery: Differential Item Functioning on the High School Form

12. PERSONAL AUTHOR(S)
Linn, R.L.; Hastings, C.N.; Hu, P.G.; Ryan, K.E.

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Final	FROM Oct 85 TO May 87	April 1988	46

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	aptitude testing, item analysis
05	09		Armed Services Vocational Aptitude Battery item bias
12	03		differential item functioning

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

This report documents results of a study of differential item functioning (also known as item bias) for the eight nonspeeded subtests of the form of the Armed Services Vocational Aptitude Battery currently used in the Department of Defense Student Testing Program. Twenty-seven different indices were calculated for each of the 200 items on the eight subtests for comparisons of item functioning for White and Black examinees, White and Hispanic examinees, and male and female examinees. Based on theoretical considerations and empirical results from this and previous research, three indices were emphasized. Items that were consistently classified as favoring one group in comparison to another group after controlling for overall performance were reviewed. None of the items on the two quantitative subtests was identified as functioning differently. On the other subtests, some items were identified as favoring one group while other items favored the second group in each of the gender and racial/ethnic group comparisons. Limited generalizations concerning the content characteristics of the identified items were suggested.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE (Include Area Code)	22c. OFFICE SYMBOL
Nancy J. Allin, Chief, STINFO Office	(512) 536-3877	AFHRL/TSR

DD Form 1473, JUN 86 *Previous editions are obsolete.* SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

SUMMARY

The results of analyses of responses of over 40,000 high school students to test items on the form of the Armed Services Vocational Aptitude Battery (ASVAB) that is currently used for the Department of Defense Student Testing Program are reported. The primary focus of the analyses was to identify items that function differently for men and women or for White, Black, and Hispanic examinees. A total of 27 different indices of differential item functioning for each of the 200 items on the eight non-speeded ASVAB subtests were obtained and compared for purposes of identifying items that may provide an advantage or a disadvantage to examinees on the basis of racial/ethnic group or gender.

Relatively few items were consistently identified as functioning differently for White examinees in comparison to Black examinees, White examinees in comparison to Hispanic examinees, or men in comparison to women. Correlations between pairs of indices and the agreement in the categorization of items as having high or low degrees of differential functioning were quite variable. Some pairs of indices, especially ones that are based on similar rationales, showed a high degree of consistency, with correlations of .90 or higher. Other pairs of indices showed little, if any, consistency.

The Shepard, Camilli, and Williams (1984) modified sum of squared differences between item response curves and the Mantel-Haenszel odds ratio (Holland & Thayer, 1986) were emphasized for purposes of interpretation. Both indices have strong theoretical and empirical justifications from previous work and were found to yield relatively consistent results in the present study. Since they are based on different analytical approaches, the agreement cannot be attributed to analytical artifacts.

There was essentially no indication of differential item functioning for either of the quantitative subtests (Arithmetic Reasoning and Mathematical Knowledge). However, a few items were consistently identified as functioning differently for White and Black examinees, White and Hispanic examinees, or men and women on the other six non-speeded subtests. Relatively few generalizations regarding the substantive characteristics of items could be used to guide future test development, however.

The results for the General Science subtest suggest that differential item functioning is associated with the distinction between physical and life science content. This result underscores the importance of maintaining a consistent balance among the content domains that are covered by the General Science subtest. On the Word Knowledge subtest, and to a lesser extent the Mechanical Comprehension and Electronics Information subtests, there was an indication that differential item functioning most often was associated with vocabulary that students are likely to encounter in science, shop, or math textbooks.

As has been found in previous studies, the direction of the differences between groups for items showing differential item functioning varied from item to item in each group comparison. It was just as common for an item to be easier for Black, Hispanic, or female examinees than for White or male examinees after controlling for overall performance as it was for the converse to be true. Thus, for the total subtest scores there is a tendency for the differential item functioning effects to be partially balanced.

PREFACE

This technical report was completed as part of the High School Equity: Item Level Approaches Study (Task 14 under Contract F41689-84-D-002). This contract is documented under Air Force Human Resources Laboratory (AFHRL) Work Unit 77191840. This study represents the continuing effort of the AFHRL to fulfill its research and development (R&D) responsibilities by examining internal test equity using state-of-the-art methodologies for the continued improvement of the Armed Services Vocational Aptitude Battery (ASVAB).

TABLE OF CONTENTS

LIST OF TABLES

List of Tables (Concluded)

Armed Services Vocational Aptitude Battery: Differential Item Functioning on the High School Form

I. INTRODUCTION

Questions of possible bias are of central concern for any testing program. The term "bias," however, is both ambiguous and controversial. It has a wide variety of technical and nontechnical meanings (see, for example, Cole, 1981; Flaugher, 1978). As a consequence, debates about "test bias" and the biased use of tests often suffer from the lack of a common use of terms and the confusion of technical and social policy issues. Because of this ambiguity, we have chosen to follow the recent lead of Holland and Thayer (1986) and use the more descriptive term "differential item functioning," rather than the more traditional "item bias," to describe the analyses reported below. Differential item functioning more accurately reflects the limited, albeit important, scope of the study and, hopefully, will have less surplus meaning than the more emotionally laden item bias label.

Item analysis has long been a familiar and useful tool in the test development process. When properly used, item analysis results can help identify items that do not function as intended, enhance reliability, and contribute to the accumulation of content and construct related evidence of validity. Although traditional item analysis procedures were not originally designed to deal with questions of group differences in performance, the inclusion of such information is a logical extension of the concept of item analysis. Studies of differential item functioning for groups that are known to differ in terms of their average performance on the test, their educational or cultural backgrounds, or other defining characteristics that are expected to have differential impact on test performance can contribute to the construct related evidence of validity. Construct validity can be enhanced by identifying and minimizing sources of difficulty on test items that are irrelevant to the construct that the test is designed to measure.

This study has three major purposes: (1) to identify items on the form of the Armed Services Vocational Aptitude Battery (ASVAB) that is currently used for the Department of Defense (DoD) Student Testing Program that function differently for men and women or for Black, Hispanic, and White students; (2) to evaluate the utility of a wide variety of procedures that have been proposed for the investigation of differential item functioning; and (3) to suggest principles and analytical procedures that could be used in the development of future versions of the ASVAB.

II. BACKGROUND

The DoD has offered the ASVAB for administration in high schools and postsecondary schools since 1966. Form 14 of the ASVAB has been administered annually to some 1.3 million students at approximately 14,000 schools throughout the nation over the past few years (DoD, 1984). The DoD Student Testing Program is intended to serve two purposes: to provide test results that are useful for educational and career counseling and to provide the military services with results that can be used "to identify students who potentially qualify for entry into the military and for assignment to military occupational training programs" (DoD, 1984, p. 2). Given the

potential importance of both of these intended uses of the ASVAB and its widespread use, it is essential that the test and associated interpretive materials meet high standards of technical quality.

The ASVAB consists of the 10 subtests listed in Table 1. Brief descriptions of the test contents, the time limits, and the number of items on each subtest are also provided in Table 1. As can be seen in Table 1, two of the subtests, Numerical Operations and Coding Speed, are highly speeded. The remaining eight subtests allow enough time for almost all students to complete the subtests and are reasonably classified as power tests.

The subtests form a variety of composite measures for use by the different services and for use in the Student Testing Program. The definitions of the Verbal (VE) composite, the Armed Forces Qualification Test (AFQT) composite, the three academic composites used in the Student Testing Program, and the four occupational composites are provided in Table 2. The VE and AFQT composites are sums of raw scores, while the academic and occupational composites are sums of standard scores.

III. METHOD

Data Set

The data for this study were provided by the Air Force Human Resources Laboratory (AFHRL) from operational administrations of Form 14 of the ASVAB to students in grades 10, 11, and 12. The data tape provided by AFHRL contained slightly over 42,000 records. In addition to responses to the items on the 10 subtests of the ASVAB, each record included the examinee's self reported grade level (10, 11, or 12), gender, and ethnicity (Black, White, or Hispanic). The total number of examinees and number in each group are listed in Table 3 for the General Science subtest of the ASVAB. Due to missing data on some subtests, the precise number of examinees available for analysis varied from a low of 41,341 for EI to a high of 42,341 for AR. The proportion of examinees in each group was relatively constant from one subtest to another, however. Thus, the number of examinees for a within-group analysis was never less than 2,000.

Descriptive Statistics

Frequency destributions, means, standard deviations, indices of skew, and indices of kurtosis were computed for each of the 10 ASVAB subtests and the ASVAB composites listed in Table 2. Matrices of intercorrelations among subtests and composites were computed. Coefficient alpha estimates of internal consistency reliability were also computed for each of the eight non-speeded subtests. All of the descriptive statistics were obtained for each of the eight groups listed in Table 3 and for the total sample. The following item statistics were also computed for each of the eight groups and the total sample: proportion correct, item deltas (i. e., a normal deviate transformation of proportion correct with a mean of 13 and a standard deviation of 4, which is used by Educational Testing Service), point-biserial correlations of items with subtest total score, and biserial correlations of items with subtest total score.

Table 1. Subtests of the Armed Services Vocational Aptitude Battery

Subtest	Code	Number of items	Time in minutes	Contents
General Science	GS	25	11	High school level physical, life, and earth sciences
Arithmetic Reasoning	AR	30	36	Arithmetic word problems
Word Knowledge	WK	35	11	Identification of synonyms and the best meaning of words in context
Paragraph Comprehension	PC	15	13	Questions regarding the information in written passages
Numerical Operations	NO	50	3	Speeded numerical calculations
Coding Speed	CS	84	7	Speeded use of a key assigning code numbers to words
Auto and Shop Information	AS	25	11	Automobile, tools, and shop terminology and practices
Mathematics Knowledge	MK	25	24	High school mathematics, including algebra and geometry
Mechanical Comprehension	MC	25	19	Use of mechanical and physical principles to visualize how illustrated objects work
Electronics Information	EI	20	9	Electricity and electronics, including circuits, inductance, capacitance, and devices such as batteries, generators, amplifiers, and test instruments

Table 2. Definitions of Selected ASVAB Composites

Composite	Subtest composition
Verbal (VE)	WK + PC
Armed Forces Qualification Test (AFQT)	AR + WK + PC + .5NO
Academic Composites	
Verbal	WK + PC + GS
Math	AR + MK
Academic Ability	WK + PC + AR
Occupational Composites	
Mechanical and Crafts	AR + AS + MC + EI
Business and Clerical	VE + CS + MK
Electronics and Electrical	GS + AR + MK + EI
Health/Social/Technology	AR + VE + MC

Table 3. Number and Percent of Examinees for the General Science Subtest by Group

Group	Number	Percent
1. Black	5,140	12.23
2. Hispanic	2,215	5.27
3. White	34,677	82.50
Total Groups 1, 2, & 3	42,032	100.00
4. Male	22,202	52.45
5. Female	20,127	47.55
Total Groups 4 & 5	42,329	100.00
6. Grade 10	8,047	19.01
7. Grade 11	21,017	49.65
8. Grade 12	13,266	31.34
Total Groups 6, 7, & 8	42,330	100.00
Total Respondents to GS	42,334	

Indices of Differential Item Functioning

A total of 27 indices of differential item functioning were computed and compared for each of the eight non-speeded subtests of the ASVAB. Each of these indices was computed for a total of five pairs of groups (Black and White, Hispanic and White, male and female, grades 10 and 11, and grades 11 and 12). Thus, in all, 27,000 (27 indices times 200 items on the 8 subtests times 5 pairs of groups) differential item functioning indices were computed.

Table 4 provides a brief description of the 27 indices, along with a reference to previous work with a given index. The first 10 indices rely on traditional test theory approaches. That is, they involve classical item statistics and conventional number right scores. Indices 11 through 27, on the other hand, are all based on item response theory (IRT). For the latter group of indices, item parameter estimates for the three-parameter logistic model were obtained separately for each of the eight groups listed in Table 3, using the BILOG computer program (Mislevy & Bock, 1984).

DELTA. Delta is an item statistic that has been used by Educational Testing Service for some time. It is simply a transformation of p, the proportion correct for an item, based on the normal distribution. A standard score, z, corresponding to a given p is obtained by finding the standard normal deviate that divides the area under the normal distribution such that the proportion of the curve above z equals p. Delta is obtained by a linear transformation of z; in particular, delta = 13 + 4z.

The delta differential item functioning index is based on the plot of deltas for one group against those of another. The magnitude of the deviation of the point corresponding to the pair of group deltas for an item from the major axis of the scatterplot of item deltas is used as the index of differential item performance (Angoff, 1982; Angoff & Ford, 1973). Delta is an improvement over indices based on a comparison of p values (e.g., Cardall & Coffman, 1964; Cleary & Hilton, 1968) because the transformation spreads out differences at the extremes (e.g., $p < .1$ or $p > .9$). However, the delta index of differential item performance can be faulted on both theoretical grounds (e.g., Hunter, 1975) and empirical grounds (e.g., Shepard, Camilli, & Williams, 1985). It was included in the present analyses primarily for comparative purposes since it has been used in a number of previous studies.

RESID. A major limitation of the delta index is that the magnitude of the between-group difference in delta is confounded with the discrimination power of an item. The residual delta index was developed and investigated by Shepard, Camilli, and Williams (1985) in response to this limitation of the delta index. The RESID index is obtained by regressing item deltas on the combined-group point-biserial correlations and computing the residual between the observed and predicted delta. Shepard, Camilli, and Williams, (1985) found the RESID index to work almost as well as Camilli's (1979) full chi-square or Linn and Harnisch's (1981) z-score approach for small samples.

PTBIS, ZPTBIS, RRPTB, and ZRRPTB. The difference between the point-biserial correlations of an item and the total subtest score for two groups has been used in several investigations of differential item performance (e.g., Rudner, Getson, & Knight, 1980). This index and three variations of it were included in the present study. PTBIS is the simple difference in the point-biserial correlations for two groups. ZPTBIS is the difference between

Table 4. Indices of Differential Item Functioning

Designation		Brief description	References
1.	DELTA	Transformed item difficulties based on normal distribution	Angoff & Ford, 1973; Angoff, 1982
2.	RESID	Residual delta based on regression of delta on combined-group point-biserial	Shepard, Camilli, & Williams, 1985
3.	PTBIS	Point-biserial correlation of item with subtest total by group	Rudner, Getson, & Knight, 1980; Green & Draper, 1972
4.	ZPTBIS	Z-tranformation of point-biserials	
5.	RRPTB	Point-biserials corrected for range restriction	
6.	ZRRPTB	Z-transformation of corrected point-biserial correlations	
7.	FCHISQ	Full chi-square based on five score intervals	Camilli, 1979
8.	SFCHISQ	Signed full chi-square based on five score intervals	Shepard, Camilli, & Averill, 1981
9.	MHCHISQ	Mantel-Haenszel chi-square statistic	Holland & Thayer, 1986; Holland, 1985
10.	MHODDS	Mantel-Haenszel common odds ratio	Holland & Thayer, 1986; Holland, 1985
11.	LCHISQ	Chi-square statistic for difference in IRT parameters a and b	Lord, 1980
12.	BHA	Base-high area - area where base group item response curve is above comparison group item response curve	Linn, Levine, Hastings, & Wardrop, 1981
13.	WBHA	Weighted base-high area	Linn, et al., 1981
14.	BLA	Base-low area	Linn, et al., 1981
15.	WBLA	Weighted base-low area	Linn, et al., 1981
16.	AREA	Total area between item response curves	Linn, et al., 1981
17.	WAREA	Weighted total area	Linn, et al., 1981

Table 4 (concluded)

Designation		Brief description	Reference
18.	SOS	Sum of squared differences between item response curves	Linn, et al., 1981
19.	WSOS	Weighted sum of squares	Linn, et al., 1981
20.	MSOS	Modified sum of squares	Shepard, Camilli, & Williams, 1984
21.	WMSOS	Weighted modified sum of squares	Shepard, et al., 1984
22.	SSOS	Signed modified sum of squares	Shepard, et al., 1984
23.	WSSOS	Weighted signed modified sum of squares	Shepard, et al., 1984
24.	TSOS	Target group sum of squares	
25.	WTSOS	Weighted target group sum of squares	
26.	STSOS	Signed target group sum of squares	
27.	WSTSOS	Weighted signed target group sum of squares	

the Fisher's z-transformations of the two correlations. To allow for differences in the within-group standard deviations, the item-test correlations were also adjusted for range restriction. That is, the correlation in group 2 was adjusted by setting the subgroup standard deviation on the total subtest score equal to the standard deviation for group 1. RRPTB is the difference between the adjusted point-biserial correlations, and ZRRPTB is the difference between the Fisher z-transformations of the adjusted correlations.

FCHISQ and SFCHISQ. Several techniques for computing chi-square-like statistics based on a comparison of groups divided into discrete intervals on the basis of total test scores have been proposed (Camilli, 1979; Ironson, 1982; Scheuneman, 1979). Scheuneman's chi-square method focuses only on a comparison of proportion correct for two groups within selected total score intervals. Due to limitations of this approach (see, for example, Baker, 1981; Camilli, 1979; Shepard, Camilli, & Averill, 1981), the "full" chi-square approach described by Camilli was used in the present study.

For the full chi-square statistic, the total subtest score range was divided into five categories. The use of five categories, rather than some other number, is consistent with most previous work with the chi-square statistic. Within each score category, the expected values were computed for the 2 X 2 contingency table formed by group (e.g., male-female) and item response (correct-incorrect), and the conventional chi-square statistic for a 2 X 2 table was computed. The full chi-square (FCHISQ) was then calculated by summing across the five score categories.

Shepard, Camilli, and Averill (1981) referred to the above chi-square as the "unsigned" chi-square. They also used a "signed" value which is obtained by multiplying signed differences times the corresponding absolute values, rather than squaring differences as in the usual chi-square. The latter value was also computed using five score categories in the present study and is denoted by SFCHISQ.

MHCHISQ and MHODDS. The last two indices based on traditional item and test statistics were recently proposed by Holland and Thayer (1986) and have been the subject of extensive study by researchers at Educational Testing Service (ETS) during the past year and a half (e.g., McPeek & Wild, 1986). The Mantel-Haenszel differential item functioning statistics suggested by Holland and Thayer are based on the formation of K 2 X 2 contingency tables for each item, where K is the number of observed total correct scores on the test. The 2 X 2 tables are defined in the same way as in Camilli's full chi-square, but rather than using a fixed number of score intervals, a 2 X 2 table is defined for each total test score that is obtained by one or more test-takers in either group being compared. A "Reference" and a "Focal" group are identified, and items are scored 1 for correct and 0 for incorrect.

For the ith item and a given total correct score of j on the subtest, a 2 X 2 contingency table with the following entries is constructed:

Score on the ith item

Group	1	0	Total
Reference	A_j	B_j	N_{rj}
Focal	C_j	D_j	N_{fj}
Total	M_{1j}	M_{0j}	T_j

The Mantel-Haenszel chi-square (MHCHISQ), which has one degree of freedom, and the Mantel-Haenszel common odds ratio (MHODDS) are defined by the following equations.

$$\text{MHCHISQ} = \frac{\left(\sum_{j=1}^{k} |A_j - E(A_j)| - .5\right)^2}{\sum_{j=1}^{k} var\,(A_j)}, \text{ and}$$

$$\text{MHODDS} = \frac{\sum_{j=1}^{k} A_j D_j / T_j}{\sum_{j=1}^{k} B_j C_j / T_j}, \text{ where}$$

$$E(A_j) = \frac{N_{rj} M_{1j}}{T_j}, \text{ and}$$

$$var(A_j) = \frac{N_{rj} N_{fa} M_{1j} M_{0j}}{(T_j)^2 (T_j - 1)}.$$

The MHODDS values can range from zero to infinity. A value of 1.0 indicates that there is no differential item performance for the two groups being compared. Values less than 1.0 indicate that the item is relatively easier for the focal group (e.g., women) than the reference group (e.g., men) after controlling for total score. The converse is true for MHODDS values greater than 1.0. As was noted by Holland and Thayer (1986), the MHODDS can be converted to the ETS delta metric by the following transformation:

Delta difference = (-2.35)ln(MHODDS).

Thus, MHODDS values of approximately .6534 or 1.5304 correspond to group differences in delta of 1.0 after controlling for total test score.

As was previously indicated, indices 11 through 27 in Table 4 are all based on item response theory. The first step in the computation of all of the IRT indices was the estimation of the item parameters for each of the eight groups of test-takers (Black, White, Hispanic, male, female, grade 10, grade 11, and grade 12). Estimates of item discrimination, a, item difficulty, b, and lower asymptote, c, parameters for the three-parameter logistic model were obtained separately for each of the eight groups on each of the eight ASVAB subtests using the BILOG computer program (Mislevy & Bock, 1984). The expectation a posteriori (EAP - Bayes) estimation procedure was used starting with the default normal priors for person ability, threshold, and log (slope). The default beta prior was used for the c parameter. The quadrature points and posterior weights were generated in phase 2 during item parameter estimation.

Before computing indices of differential item performance for a given pair of groups, it was necessary to equate the IRT scales for the two groups. This was done by transforming the estimates of person ability, item difficulty, and item discrimination of one group, labeled the comparison group, to the scale of the other group, labeled the base group. Following Linn, Levine, Hastings, and Wardrop (1981), the transformation was obtained by finding the linear equating constants such that after conversion, the weighted mean and variance of the b's were the same for both groups being compared (see Linn et al., 1981, for a detailed description of the procedures used to transform the parameters for the comparison group to the scale of the base group). Once the parameter estimates for each pair of groups were placed on a common scale, the calculation of the differential item functioning indices proceeded as described below.

LCHISQ. The chi-square statistic described by Lord (1980), which provides a simultaneous test of the between-group differences in the a and b parameters, was computed for each item. The chi-square statistics may be compared to a chi-square distribution with two degrees of freedom.

Areas Between Item Response Curves. Linn, Levine, Hastings, and Wardrop (1981) investigated a total of six area measures as indices of differential item functioning. These six indices all involve the calculation of item response curves for the two groups being compared and the calculation of areas or weighted areas between the curves over a range of theta from -3.0 to +3.0. The three unweighted areas are defined as follows:

BHA = the area between the curves where the probability of a correct response is higher for the base group (e.g., men) than the corresponding probability for the comparison group (e.g., women),

BLA = the area between the curves where the probability of a correct response is lower for the base group than the corresponding probability for the comparison group, and

AREA = BHA + BLA.

The three weighted area indices (WBHA, WBLA, and WAREA) have parallel definitions, but the differences between the curves for the two groups are weighted by the inverse of the estimated sampling variance of the difference between the probability of a correct response for the two groups at each theta value. Formulas for computing the weights are given by Linn et al. (1981).

SOS and WSOS. Indices 18 and 19 are the sum of squared differences between the item response curves and the weighted sum of squared differences. Like the area indices, these indices were first proposed by Linn, Levine, Hastings, and Wardrop (1981) and were computed over a range of theta values from -3.0 to +3.0.

Modified Sum of Squares. Shepard, Camilli, and Williams (1984) proposed and used modifications of the Linn et al. sum of squares indices. The modified sums of squares are computed for estimated values of theta for the sample, rather than over a prespecified range of -3.0 to +3.0. Specifically, the four indices that were computed are defined as follows:

$$\text{MSOS} = \frac{1}{N_1+N_2}\sum_{j=1}^{N_1+N_2} D_j^2$$

$$\text{WMSOS} = \frac{1}{N_1+N_2}\sum_{j=1}^{N_1+N_2} W_j D_j^2$$

$$\text{SSOS} = \frac{1}{N_1+N_2}\sum_{j=1}^{N_1+N_2} D_j |D_j|$$

$$\text{WSSOS} = \frac{1}{N_1+N_2}\sum_{j=1}^{N_1+N_2} W_j D_j |D_j|$$

where N_1 is the number of cases in group 1, Wj is the estimated sampling variance of the difference between the estimated item response curves at a given theta, and Dj is the difference between the probability of a correct response based on the item response curves for the two groups at a given theta.

Target Group Sum of Squares. One rationale for the Shepard et al. (1984) modified sum of squares indices is that they are "self-weighting." That is, unlike the procedure originally proposed by Linn et al., the focus is only on values of theta that are estimated for individuals in the two samples. This logic can be taken one step further by noting that the major concern is often for differences in the item response curves at values of theta that are estimated for a particular group. For example, in a comparison of the performance of Black and White examinees on an item, the primary concern may be with the difference in the probability of a correct response based on the two sets of item parameters for the theta values observed for the Black test-takers. Therefore, four additional indices were computed parallel to the four Shepard et al. indices, by summing only over estimated theta values for the "target" or comparison group. For example,

$$TSOS = \frac{1}{N_2} \sum_{j=1}^{N} D_j^2.$$

The remaining three indices (WTSOS, STSOS, and WSTSOS) are defined as the analogous counterparts of WMSOS, SSOS, and WSSOS, respectively.

Summaries and Comparisons of Indices

Distributions of indices were inspected to aid interpretation. The distributions of indices for the grade 10 versus grade 11 comparisons and the grade 11 versus grade 12 comparisons provided a means of defining unusually large values of each index. Some empirical basis for defining values of the indices that are large enough to be considered of practical importance was necessary for those indices lacking known distributional characteristics, but was also considered useful for those such as the Mantel-Haenszel chi-square statistics due to the large number of significant differences that could be anticipated even for small differences because of the sample sizes and the number of statistical tests.

Correlations among the various indices were computed for each pair of groups on each of the eight subtests. 2 X 2 contingency tables with items classified as having large or small indices according to two different methods were also obtained for selected pairs of unsigned indices. 3 X 3 contingency tables were used in a similar manner for selected pairs of signed indices with items classified as having a large index in favor of the target group, a large index in favor of the comparison group, or a small index.

IV. RESULTS

Descriptive Statistics

Descriptive statistics for the total sample based on the number correct scores on each of the 10 ASVAB subtests are presented in Table 5. As can be seen, the distributions are negatively skewed for five of the subtests (GS, WK, PC, NO, and CS) and positively skewed for the other five subtests. The degree of skewness is most extreme for NO. Twenty-eight percent of the total sample answered 90% or more of the NO items correctly. On the other two most negatively skewed subtests, WK and PC, 13.1% and 10.6%, respectively, of the test-takers answered 90% or more of the items correctly. In comparison to a normal distribution, all of the non-speeded tests are somewhat platykurtic, whereas the two speeded subtests are slightly leptokurtic. Complete frequency distributions and descriptive statistics for each group and the total sample have been provided to AFHRL on computer tape and will not be presented here due to their volume.

Subgroup means and standard deviations on the 10 ASVAB subtests are presented in Table 6. As can be seen, there are substantial differences in average test performance among the three racial/ethnic groups. It should be emphasized, however, that the score distributions for the three groups overlap substantially. For example, 12.3% of the Black test-takers scored above the median for White test-takers on AR. Females had higher mean scores than did males on PC and the two speeded subtests (NO and CS), while the converse was true on the other subtests. Mean differences by grade were relatively small, especially between grades 11 and 12. The means for grade 10 students, however, were lower than the means for grade 11 or grade 12 on all subtests. The patterns of differences in means for racial/ethnic groups and for gender were generally consistent with expectations based on previously published results for the ASVAB and other multiple aptitude test batteries (see, for example, Profile of American Youth: 1980 Nationwide Administration of the Armed Services Vocational Aptitude Battery, Office of the Assistant Secretary of Defense, 1982).

Internal consistency estimates of reliability (coefficient alpha) are listed in Table 7 for each of the eight non-speeded subtests by group and for the total sample. For the total sample, the alphas range from .67 for EI to .88 for WK. The alphas for grade 10 are slightly lower than the corresponding values for grades 11 and 12, but there are no marked discrepancies between grades. Alphas for women are generally lower than the corresponding values for men, substantially so for three subtests (AS, .56 vs .81; MC, .61 vs .81; and EI, .40 vs .72). On each subtest, the alpha for White examinees is higher than the alpha for Hispanic examinees, which, in turn, is higher than the one for Black examinees. For the five subtests that are most "academic" in nature (GS, AR, WK, PC, and MK), the Black-White differences in alpha coefficients were smallest -- ranging from .03 to .13. Larger differences, between .16 and .20, were obtained on the three other subtests. Thus, it is the three subtests (AS, MC, and EI) with content that is least similiar to the content of a traditional academic high school program that have the largest gender and race/ethnicity group differences in internal consistency.

Table 5. ASVAB Subtest Descriptive Statistics Based on Number Correct Scores for the Total Sample

Subtest[a]	N	Mean	SD	Skewness	Kurtosis
GS	42,334	14.68	4.75	-.16	-.59
AR	42,341	17.13	6.50	.11	-.95
WK	42,309	23.72	6.72	-.45	-.35
PC	42,214	9.53	3.29	-.40	-.68
NO	42,234	37.40	9.04	-.77	.09
CS	42,169	45.45	13.69	-.04	.26
AS	42,121	12.90	5.04	.25	-.75
MK	42,074	12.53	5.17	.37	-.61
MC	41,738	12.74	4.84	.24	-.67
EI	41,341	9.66	3.47	.27	-.26

[a] See Table 1 for subtest titles and definitions.

Table 6. Group Means and Standard Deviations (in parentheses) on the ASVAB Subtests

	Grade			Gender		Race/Ethnicity			Total
Subtest[a]	10	11	12	M	F	W	B	H	sample
GS	14.17 (4.56)	15.00 (4.68)	14.69 (4.84)	15.33 (5.06)	13.98 (4.26)	15.52 (4.42)	10.69 (4.06)	10.66 (4.35)	14.68 (4.75)
AR	15.98 (6.13)	17.41 (6.47)	17.40 (6.60)	17.71 (6.72)	16.49 (6.18)	18.11 (6.33)	12.47 (4.90)	12.55 (5.60)	17.13 (6.50)
WK	22.64 (6.33)	24.04 (6.53)	23.94 (6.94)	23.82 (7.05)	23.62 (6.35)	24.92 (6.12)	18.34 (6.29)	17.20 (6.93)	23.72 (6.72)
PC	9.04 (3.23)	9.71 (3.23)	9.61 (3.32)	9.47 (3.44)	9.60 (3.10)	10.04 (3.11)	7.15 (2.90)	6.97 (3.19)	9.53 (3.29)
NO	35.87 (9.20)	37.67 (8.81)	37.81 (9.06)	36.00 (9.35)	38.94 (8.41)	37.96 (8.75)	34.52 (9.91)	34.99 (9.79)	37.40 (9.04)
CS	42.20 (12.9)	45.59 (13.4)	46.61 (14.0)	41.87 (13.3)	49.40 (13.0)	46.47 (13.3)	39.84 (14.4)	42.14 (14.6)	45.45 (13.69)
AS	11.87 (4.52)	13.04 (4.93)	13.20 (5.25)	15.53 (4.85)	10.00 (3.40)	13.68 (4.93)	9.01 (3.57)	9.66 (4.28)	12.90 (5.04)
MK	11.75 (4.58)	12.87 (5.17)	12.61 (5.36)	12.61 (5.35)	12.44 (4.97)	13.13 (5.18)	9.66 (3.98)	9.59 (4.09)	12.53 (5.17)
MC	12.05 (4.42)	13.01 (4.82)	12.82 (4.98)	14.36 (5.01)	10.95 (3.93)	13.47 (4.73)	9.01 (3.46)	9.70 (4.07)	12.74 (4.84)
EI	9.05 (3.22)	9.69 (3.40)	9.87 (3.57)	10.94 (3.66)	8.25 (2.60)	10.09 (3.42)	7.60 (2.87)	7.70 (3.04)	9.66 (3.47)

[a] See Table 1 for subtest titles and descriptions.

Table 7. Non-Speeded ASVAB Subtest Coefficient Alpha Internal Consistency Estimates of Reliability by Group and Total Sample

Subtest[a]	Grade			Gender		Race/Ethnicity			Total sample
	10	11	12	M	F	W	B	H	
GS	.78	.81	.80	.83	.76	.78	.71	.75	.80
AR	.85	.88	.87	.89	.85	.87	.76	.82	.87
WK	.86	.89	.88	.90	.87	.87	.84	.86	.88
PC	.74	.77	.76	.78	.74	.75	.65	.71	.76
AS	.76	.83	.80	.81	.56	.81	.62	.73	.81
MK	.77	.84	.82	.83	.81	.82	.69	.71	.82
MC	.73	.80	.78	.81	.61	.77	.57	.69	.78
EI	.61	.69	.66	.72	.40	.67	.51	.55	.67

[a] See Table 1 for subtest titles and descriptions.

The subtest intercorrelations for the total sample are presented below the main diagonal of the correlation matrix in Table 8. For comparative purposes, correlations based on the 1980 national sample of 9,173 youths between the ages of 18 and 23 (Office of the Assistant Secretary of Defense, 1982) are provided above the diagonal of the correlation matrix in the same table. As would be expected, the correlations are, with one exception (MC correlated with PC), smaller for the more homogeneous, younger, in-school sample in the present study than for the nationally representative sample of youth, in and out of school, from the Profile of American Youth study.

The patterns of correlations for the high school and Profile samples are generally similar. For example, GS has correlations of .79 with WK and .68 with PC in the high school sample, and the corresponding Profile figures are .80 and .69, respectively. However, the correlations of three of the subtests (NO, CS, and EI) with the other seven subtests are noticeably lower for the high school sample than they are for the Profile sample.

The subtest intercorrelation matrices for grades 10, 11, and 12 are presented in Tables 9 and 10; those for men and women, in Table 11; and those for Black, White, and Hispanic test-takers, in Tables 12 and 13. The patterns of correlations are similar for all these groups. Differences between groups in the magnitude of the correlations are generally consistent with expectations based on the group's variability (see Table 6 for subtest standard deviations) and the reliability estimates of the subtests for the different groups (see Table 7). However, the within-sex correlations of the two speeded subtests with the non-speeded subtests tend to be higher than the corresponding correlations for the combined sex groups. The latter result reflects the fact that women have higher means than men on the two speeded subtests, while the converse is true for all but one of the non-speeded subtests.

The intercorrelations of the nine ASVAB composites (VE, AFQT, the three high school composites, and the four occupational composites) are presented in Table 14 for the total sample. As can be seen, some pairs of composites are virtually indistinguishable. For example, VE correlates .98 with the

high school Verbal composite and the Health/Social/Technology composite correlates .98 with the high school Academic Ability composite. These high correlations are hardly surprising, however, given the overlap of subtests that define these composites (e.g., VE = WK + PC and the high school Verbal composite = WK + PC + GS).

Composite score intercorrelation matrices were also computed for each of the eight groups. As would be expected, given the subtest intercorrelations and the overlapping definitions of the composites, similar patterns of correlations, albeit lower in some cases due to the smaller within-group variability, were obtained for all of the groups.

Table 8. ASVAB Subtest Intercorrelations for the Total High School Sample (below the diagonal) and for the Profile of American Youth Sample (above the diagonal)[a]

	ASVAB Subtest[b]									
	GS	AR	WK	PC	NO	CS	AS	MK	MC	EI
GS	--	.72	.80	.69	.52	.45	.64	.69	.70	.76
AR	.65	--	.71	.67	.63	.51	.53	.83	.68	.66
WK	.79	.67	--	.80	.60	.55	.52	.67	.59	.68
PC	.68	.67	.76	--	.60	.56	.42	.64	.52	.57
NO	.34	.46	.40	.41	--	.70	.29	.62	.40	.41
CS	.29	.40	.37	.39	.63	--	.22	.52	.33	.34
AS	.55	.45	.48	.42	.10	.03	--	.41	.74	.75
MK	.60	.76	.62	.62	.45	.40	.34	--	.60	.58
MC	.61	.61	.57	.55	.23	.20	.65	.56	--	.74
EI	.58	.48	.53	.48	.17	.14	.66	.42	.62	--

[a] From the Profile of American Youth, Office of the Assistant Secretary of Defense, 1982, p. 65, (n = 9,173).

[b] See Table 1 for subtest titles and descriptions.

Table 9. ASVAB Subtest Intercorrelations for the Grade 10 Examinees (below the diagonal) and for the Grade 11 Examinees (above the diagonal)

	ASVAB Subtest									
	GS	AR	WK	PC	NO	CS	AS	MK	MC	EI
GS	--	.64	.78	.68	.32	.27	.54	.60	.61	.57
AR	.60	--	.66	.66	.45	.39	.43	.77	.62	.48
WK	.73	.63	--	.75	.39	.35	.46	.62	.60	.52
PC	.64	.65	.73	--	.40	.36	.40	.63	.55	.48
NO	.32	.45	.40	.41	--	.62	.08	.46	.22	.16
CS	.28	.39	.37	.41	.65	--	.00	.39	.19	.11
AS	.49	.39	.43	.36	.06	.00	--	.32	.64	.65
MK	.55	.72	.59	.60	.44	.41	.30	--	.56	.41
MC	.55	.57	.53	.52	.22	.19	.59	.53	--	.62
EI	.53	.43	.50	.44	.16	.14	.59	.39	.55	--

Table 10. ASVAB Subtest Intercorrelations for Grade 12 Examinees

	ASVAB Subtest									
	GS	AR	WK	PC	NO	CS	AS	MK	MC	EI
GS	--									
AR	.67	--								
WK	.80	.69	--							
PC	.70	.68	.77	--						
NO	.35	.47	.40	.41	--					
CS	.30	.40	.37	.39	.63	--				
AS	.57	.47	.50	.44	.12	.03	--			
MK	.61	.77	.62	.63	.45	.40	.35	--		
MC	.62	.62	.57	.55	.24	.20	.68	.57	--	
EI	.60	.49	.53	.49	.18	.13	.68	.43	.64	--

Table 11. ASVAB Subtest Intercorrelations for Men (below the diagonal) and for Women (above the diagonal)

	ASVAB Subtest									
	GS	AR	WK	PC	NO	CS	AS	MK	MC	EI
GS	--	.62	.77	.67	.35	.30	.53	.59	.53	.47
AR	.66	--	.66	.66	.44	.40	.46	.76	.58	.41
WK	.81	.68	--	.75	.38	.34	.53	.62	.54	.47
PC	.71	.69	.77	--	.37	.35	.49	.61	.54	.44
NO	.39	.52	.43	.44	--	.60	.21	.43	.26	.21
CS	.38	.49	.43	.44	.64	--	.20	.37	.27	.21
AS	.59	.49	.58	.53	.25	.25	--	.42	.47	.42
MK	.61	.77	.62	.63	.49	.47	.39	--	.57	.39
MC	.65	.66	.64	.63	.35	.37	.65	.62	--	.40
EI	.63	.54	.62	.59	.30	.32	.65	.49	.65	--

Table 12. ASVAB Subtest Intercorrelations for White Examinees (below the diagonal) and Black Examinees (above the diagonal)

	ASVAB Subtest									
	GS	AR	WK	PC	NO	CS	AS	MK	MC	EI
GS	--	.50	.72	.60	.29	.21	.43	.48	.42	.43
AR	.61	--	.56	.56	.36	.29	.35	.62	.42	.34
WK	.75	.63	--	.68	.34	.29	.42	.52	.42	.45
PC	.64	.64	.73	--	.33	.30	.35	.51	.40	.39
NO	.32	.46	.39	.40	--	.60	.08	.36	.13	.14
CS	.25	.39	.35	.37	.63	--	.00	.27	.11	.12
AS	.49	.38	.39	.33	.05	-.04	--	.26	.47	.47
MK	.57	.76	.60	.60	.46	.40	.28	--	.41	.32
MC	.57	.58	.51	.50	.21	.16	.62	.53	--	.44
EI	.55	.44	.48	.43	.14	.08	.64	.38	.60	--

Table 13. ASVAB Subtest Intercorrelations for Hispanic Examinees

	ASVAB Subtest									
	GS	AR	WK	PC	NO	CS	AS	MK	MC	EI
GS	--									
AR	.63	--								
WK	.75	.67	--							
PC	.66	.64	.75	--						
NO	.32	.39	.37	.33	--					
CS	.28	.30	.32	.31	.56	--				
AS	.53	.50	.55	.49	.16	.09	--			
MK	.54	.69	.56	.53	.38	.29	.38	--		
MC	.57	.61	.58	.54	.23	.19	.61	.51	--	
EI	.53	.48	.52	.47	.21	.19	.59	.40	.55	--

Table 14. Intercorrelations of the ASVAB Composites for the Total Sample

	Composite								
	VE	AFQT	V	M	AA	MC	BC	EE	HST
VE	--								
AFQT	.92	--							
V	.98	.90	--						
M	.73	.88	.74	--					
AA	.95	.97	.94	.88	--				
MC	.71	.77	.75	.79	.81	--			
BC	.80	.86	.78	.76	.81	.61	--		
EE	.81	.90	.86	.94	.92	.89	.76	--	
HST	.92	.95	.93	.88	.98	.89	.79	.93	--

Note. Correlations not corrected for spurious overlap.

VE = Verbal Subtest.
Composites:
AFQT = Armed Forces Qualification Test.
V = Verbal.
M = Math.
AA = Academic Ability.
MC = Mechanical and Crafts.
BC = Business and Clerical.
EE = Electronics and Electrical.
HST = Health/Social/Technology.

Differential Item Functioning

Since a total of 27 differential item functioning indices were computed for each of five pairs of groups on each of the 200 items on the eight non-speeded subtests, a listing of all 27,000 indices is not practical. Complete files of the indices, as well as the intermediate statistics such as

the IRT parameter estimates and the constants computed for converting parameter estimates for one group to the scale of the group to which it was compared, have been provided to AFHRL on computer tapes, but only summary information is included in this report.

Considerable effort was put into reviewing distributions of indices and their interrelationships. The distributions of indices for the grade 10 vs grade 11 and the grade 11 vs grade 12 comparisons provided a benchmark for identifying items with large indices in the other group comparisons. Theoretical considerations and the results of previous empirical and simulation studies were used--together with the Pearson product-moment correlations, Spearman rank-order correlations, and agreement of the classification of the items having large or small differential functioning indices for the gender and racial/ethnic group comparisons--to select a subset of indices that are emphasized in this report. These steps are illustrated in detail for the GS subtest; then, summary results for all subtests are presented.

The means and standard deviations of the 27 differential item functioning indices for the GS subtest are listed in Table 15 for each of the pairs of groups that were compared. With the exception of the four chi-square indices, the means and standard deviations have all been multiplied by 100. As can be seen, the means and standard deviations for the comparisons of grades 10 and 11 are very similar to those for the comparisons of grades 11 and 12. The standard deviations for the gender and racial/ethnic group comparisons, on the other hand, are generally substantially larger than the corresponding figures for the grade comparisons. Furthermore, for indices that are unsigned (e.g., the chi-square indices and the sum of squares indices), the means are also substantially larger for the gender and racial/ethnic group comparisons than they are for the grade comparisons.

An inspection of the intercorrelations for the five pairs of groups revealed several clusters of indices. Some of these clusters of indices consist primarily of those indices with similar definitions. For example, the four indices based on the point-biserial correlations (PTBIS, ZPTBIS, RRPTB, and ZRRPTB) were highly interrelated (r's $>.95$) with each other for all eight subtests and all five pairs of groups, but showed very little relationship to the other indices. The various sum of squares indices were also highly interrelated, but they also had generally high correlations with some of the other indices, particularly the full chi-square (FCHISQ) and the Mantel-Haenszel chi-square (MHCHISQ).

Table 16 lists all of the pairs of indices that had Spearman rank-order correlations of .50 or higher for all three comparisons involving gender or racial/ethnic groups for the GS subtest. From an inspection of Table 16, it can be seen that, in most cases, high correlations occur where indices are based on similar definitions. For example, the correlations of FCHISQ and MHCHISQ range from .79 for the White-Black comparison to .97 for the male-female comparison. The three correlations between MSOS and TSOS are all .95 or higher, and the three between AREA and SOS are all .94 or higher. Close correspondence between similarly defined indices is to be expected and was found.

Table 15. Means and Standard Deviations (in parentheses) of the Differential Item Functioning Indices for the General Science Subtest

		Groups Compared[a]									
	Index[b]	11-10		11-12		M-F		W-B		W-H	
1.	DELTA	0	(26)	0	(23)	0	(90)	0	(74)	0	(104)
2.	RESID	0	(24)	0	(23)	0	(89)	0	(54)	0	(97)
3.	PTBIS	2	(2)	-1	(2)	1	(7)	5	(8)	3	(12)
4.	ZPTBIS	2	(2)	-2	(2)	8	(7)	3	(13)	5	(9)
5.	RRPTB	1	(2)	0	(2)	0	(8)	2	(8)	2	(12)
6.	ZRRPTB	1	(2)	0	(2)	6	(6)	2	(10)	2	(14)
7.	FCHISQ	22	(19)	27	(35)	378	(431)	150	(174)	88	(118)
8.	SFCHISQ	0.7	(7.4)	0.8	(9.4)	24	(68)	20	(108)	23	(101)
9.	MHCHISQ	16	(18)	20	(30)	269	(342)	88	(163)	63	(106)
10.	MHDDS	101	(14)	101	(12)	108	(45)	108	(41)	110	(55)
11.	LCHISQ	4.6	(6)	3.7	(5.3)	11	(13)	7.4	(11)	9.7	(8.5)
12.	BHA	10	(10)	7	(8)	20	(19)	29	(33)	20	(25)
13.	WBHA	4	(7)	4	(7)	13	(17)	13	(20)	10	(15)
14.	BLA	8	(10)	10	(11)	15	(19)	9	(21)	18	(21)
15.	WBLA	5	(8)	7	(10)	4	(7)	4	(10)	12	(19)
16.	AREA	18	(11)	17	(11)	35	(17)	38	(32)	38	(25)
17.	WAREA	10	(10)	11	(12)	17	(17)	17	(21)	22	(21)
18.	SOS	1	(2)	1	(1)	5	(5)	7	(13)	6	(9)
19.	WSOS	1	(1)	1	(1)	1	(1)	2	(4)	2	(4)
20.	MSOS	0	(0.5)	0	(0.3)	1	(1)	1	(2)	1	(2)
21.	WMSOS	0	(0.4)	0	(0.2)	0	(0.4)	1	(1)	1	(1)
22.	SSOS	0	(0.5)	0	(0.3)	1	(1)	1	(2)	1	(2)
23.	WSSOS	0	(0.4)	0	(0.3)	1	(0.4)	1	(1)	1	(1)
24.	TSOS	0	(0.5)	0	(0.3)	1	(1)	1	(3)	2	(2)
25.	WTSOS	0	(0.5)	0	(0.2)	1	(0.6)	1	(2)	1	(1)
26.	STSOS	0	(0.5)	0	(0.3)	1	(1)	1	(3)	2	(2)
27.	WSTSOS	0	(0.4)	0	(0.2)	0	(0.6)	1	(2)	1	(1)

[a] The first group listed is the base or reference group and the second group is the comparison or focal group.

[b] All indices, with the exception of the chi-square indices (i. e., FCHISQ, SFCHISQ, MHCHISQ, and LCHISQ), have been multiplied by 100.

Of greater interest are the cases where indices based on distinctly different methodologies have consistently high relationships. In this regard, the relationships between FCHISQ and MHCHISQ and the IRT-based sum of squares indices are especially noteworthy. By the .50 criterion used for the entries in Table 16, FCHISQ and MHCHISQ are related to AREA, SOS, MSOS, SSOS, TSOS, and STSOS (i.e., all of the unweighted sum of squares indices). It should also be noted that the weighted IRT-based indices generally show less agreement with other indices than do the unweighted indices.

Table 16. Pairs of Indices with Spearman Rank-Order Correlations Greater than .50 for the Three Gender or Racial/Ethnic Group Comparisons on General Science Subtest

Index	Corrected indices	Rank-Order correlations W-B	W-H	M-F
1. DELTA	2. RESID	.59	.85	.97
	10. MHODDS	.73	.85	.78
	12. BHA	.52	.59	.69
2. RESID	1. DELTA	.59	.85	.97
	10. MHODDS	.85	.95	.81
	12. BHA	.80	.61	.75
	14. BLA	-.63	-.76	-.69
3. PTBIS 4. ZPTBIS 5. RRPTB 6. ZRRPTB	None other than the other indices based on the point-biserial correlations, which by definition must have perfect or nearly perfect rank-order correlations with each other.			
7. FCHISQ	9. MHCHISQ	.79	.86	.97
	16. AREA	.60	.60	.75
	18. SOS	.56	.67	.73
	20. MSOS	.84	.71	.84
	22. SSOS	.82	.73	.86
	24. TSOS	.84	.69	.76
	26. STSOS	.82	.69	.79
8. SFCHISQ	None			
9. MHCHISQ	7. FCHISQ	.79	.86	.97
	16. AREA	.61	.57	.72
	18. SOS	.54	.64	.70
	20. MSOS	.74	.60	.79
	22. SSOS	.71	.66	.83
	24. TSOS	.72	.60	.72
	26. STSOS	.68	.67	.76
10. MHODDS	1. DELTA	.73	.85	.78
	2. RESID	.85	.95	.81
	12. BHA	.84	.72	.93
	14. BLA	-.77	-.86	-.83
	15. WBLA	-.78	-.69	-.59
11. LCHISQ	16. AREA	.55	.68	.63
	20. MSOS	.66	.79	.75
	22. SSOS	.63	.71	.63
	24. TSOS	.68	.73	.65
	26. STSOS	.65	.64	.61
12. BHA	1. DELTA	.52	.59	.69
	2. RESID	.80	.61	.75
	10. MHODDS	.84	.72	.93

Table 16 (Continued)

Index	Correlated indices	Rank-Order correlations W-B	W-H	M-F
12. BHA	13. WBHA	.59	.80	.66
	14. BLA	-.75	-.72	-.76
13. WBHA	12. BHA	.59	.80	.66
	19. WSOS	.63	.56	.60
14. BLA	2. RESID	-.63	-.76	-.69
	10. MHODDS	-.77	-.86	-.83
	12. BHA	-.75	-.72	-.76
	15. WBLA	.91.	.66	.78
15. WBLA	10. MHODDS	-.78	-.69	-.59
	14. BLA	.91	.66	.78
16. AREA	7. FCHISQ	.60	.60	.75
	9. MHCHISQ	.61	.57	.72
	11. LCHISQ	.55	.68	.63
	18. SOS	.95	.96	.94
	20. MSOS	.85	.90	.89
	21. WMSOS	.79	.79	.73
	22. SSOS	.87	.86	.73
	23. WSSOS	.83	.74	.75
	24. TSOS	.80	.82	.79
	25. WTSOS	.81	.74	.73
	26. STSOS	.80	.78	.72
	27. WSTSOS	.83	.68	.75
17. WAREA	19. WSOS	.91	.91	.94
	21. WMSOS	.76	.54	.67
	23. WSSOS	.71	.56	.55
	27. WSTSOS	.71	.56	.55
18. SOS	7. FCHISQ	.56	.67	.73
	9. MHCHISQ	.54	.64	.70
	16. AREA	.95	.96	.94
	20. MSOS	.80	.95	.92
	21. WMSOS	.61	.71	.53
	22. SSOS	.83	.94	.79
	23. WSSOS	.69	.68	.58
	24. TSOS	.76	.88	.82
	25. WTSOS	.69	.72	.60
	26. STSOS	.77	.84	.77
	27. WSTSOS	.73	.67	.64
19. WSOS	13. WBHA	.63	.56	.60
	17. WAREA	.91	.91	.94
	21. WMSOS	.93	.73	.82
	23. WSSOS	.88	.72	.71

Table 16 (Continued)

Index	Correlated indices		Rank-Order correlations W-B	W-H	M-F
20. MSOS	7.	FCHISQ	.84	.71	.84
	9.	MHCHISQ	.74	.60	.79
	11.	LCHISQ	.66	.79	.75
	16.	AREA	.85	.90	.89
	18.	SOS	.80	.95	.92
	21.	WMSOS	.79	.70	.66
	22.	SSOS	.99	.96	.90
	23.	WSSOS	.83	.64	.69
	24.	TSOS	.97	.95	.95
	25.	WTSOS	.88	.77	.77
	26.	STSOS	.96	.90	.91
	27.	WSTSOS	.86	.71	.79
21. WMSOS	16.	AREA	.79	.79	.73
	17.	WAREA	.76	.54	.67
	18.	SOS	.61	.71	.53
	19.	WSOS	.93	.73	.82
	20.	MSOS	.79	.70	.66
	22.	SSOS	.76	.64	.51
	23.	WSSOS	.94	.94	.94
	24.	TSOS	.80	.72	.67
	25.	WTSOS	.93	.89	.95
	26.	STSOS	.77	.67	.58
	27.	WSTSOS	.87	.85	.91
22. SSOS	7.	FCHISQ	.82	.73	.86
	9.	MHCHISQ	.71	.66	.83
	11.	LCHISQ	.63	.71	.63
	16.	AREA	.87	.86	.73
	18.	SOS	.83	.94	.79
	20.	MSOS	.99	.96	.90
	21.	WMSOS	.76	.64	.51
	23.	WSSOS	.84	.58	.61
	24.	TSOS	.94	.91	.88
	25.	WTSOS	.85	.68	.64
	26.	STSOS	.95	.91	.96
	27.	WSTSOS	.86	.63	.71
23. WSSOS	16.	AREA	.83	.74	.75
	17.	WAREA	.71	.56	.55
	18.	SOS	.69	.68	.58
	19.	WSOS	.88	.72	.71
	20.	MSOS	.83	.64	.69
	21.	WMSOS	.94	.94	.94
	22.	SSOS	.84	.58	.61
	24.	TSOS	.82	.66	.70
	25.	WTSOS	.91	.86	.91
	26.	STSOS	.83	.64	.64
	27.	WSTSOS	.93	.87	.95

Table 16 (Concluded)

Index	Correlated indices		Rank-Order correlations W-B	W-H	M-F
24. TSOS	7.	FCHISQ	.84	.69	.76
	9.	MHCHISQ	.72	.60	.72
	11.	LCHISQ	.68	.73	.65
	16.	AREA	.80	.82	.79
	18.	SOS	.76	.88	.82
	20.	MSOS	.97	.95	.95
	21.	WMSOS	.80	.72	.67
	22.	SSOS	.94	.91	.88
	23.	WSSOS	.82	.66	.70
	25.	WTSOS	.92	.88	.81
	26.	STSOS	.99	.95	.96
	27.	WSTSOS	.90	.83	.84
25. WTSOS	13.	WBHA	.63	.56	.60
	16.	AREA	.81	.74	.73
	18.	SOS	.69	.72	.60
	20.	MSOS	.88	.77	.77
	21.	WMSOS	.93	.89	.95
	22.	SSOS	.85	.68	.64
	23.	WSSOS	.91	.86	.91
	24.	TSOS	.92	.88	.81
	26.	STSOS	.90	.80	.72
	27.	WSTSOS	.93	.97	.97
26. STSOS	7.	FCHISQ	.82	.69	.79
	9.	MHCHISQ	.68	.67	.76
	11.	LCHISQ	.65	.64	.61
	16.	AREA	.80	.78	.72
	18.	SOS	.77	.84	.77
	20.	MSOS	.96	.90	.91
	21.	WMSOS	.77	.67	.58
	22.	SSOS	.95	.91	.96
	23.	WSSOS	.83	.64	.64
	24.	TSOS	.99	.95	.96
	25.	WTSOS	.90	.80	.72
	27.	WSTSOS	.91	.78	.77
27. WSTSOS	16.	AREA	.83	.68	.75
	17.	WAREA	.71	.56	.55
	18.	SOS	.73	.67	.64
	20.	MSOS	.86	.71	.79
	21.	WMSOS	.87	.85	.91
	22.	SSOS	.86	.63	.71
	23.	WSSOS	.93	.87	.95
	24.	TSOS	.90	.83	.84
	25.	WTSOS	.93	.97	.97
	26.	STSOS	.91	.78	.77

Given the redundancy of the IRT-based indices and earlier results (Shepard, Camilli, and Williams, 1984; 1985) suggesting that MSOS or its signed counterpart, SSOS, are among the best indices, we have chosen to focus heavily on these two indices. Among the non-IRT indices, emphasis will be given to FCHISQ and MHCHISQ. The latter indices were selected based, in part, on the previous findings of Shepard *et al.*, and inspection of the interrelationships among the indices such as those shown in Table 16.

Based on the above considerations, five unsigned indices and four signed, or directional, indices were selected for primary focus. Two of the selected unsigned indices, FCHISQ and MHCHISQ, are based on observed scores, whereas the other three, LCHISQ, AREA, and MSOS, are based on IRT results. The selected directional indices consist of RESID, SFCHISQ, MHODDS, and BHA.

The Spearman rank-order correlations among the selected signed indices are presented in Table 17 for each of the three gender and racial/ethnic group comparisons on each of the eight non-speeded subtests. Median correlations for each pair of indices are reported separately for each group comparison, and the median correlation between each pair of indices over all 24 coefficients is provided at the bottom of the table. As might be expected, the correlations between the FCHISQ and MHCHISQ indices and those between AREA and MSOS are generally above .80. Correlations of IRT-based indices with non-IRT-based indices, while typically positive, are considerably lower. The median of the 24 correlations between MSOS and MHCHISQ, for example, is only .42.

For most of the subtests, the correlations between the IRT- and non-IRT based indices are higher for the male-female and White-Black comparisons than for the White-Hispanic comparison. The poor agreement between these two categories of indices for the White-Hispanic comparison is most notable for AS and EI, where several of the correlations are negative.

Table 18 lists the correlations and median Spearman rank-order correlations among the selected signed, or directional, indices. Given its simplicity, the RESID index correlates well with the other three directional indices, a result that is in agreement with earlier findings for smaller samples by Shepard, Camilli, and Williams (1985). All of the correlations in Table 18 are positive, with the overall median correlations between pairs of indices ranging from a low of .58 to a high of .92.

Distributions of indices for the grade-to-grade comparisons were inspected and used as the basis for defining high and low values of the indices. As before, this is first illustrated for the GS subtest; then, summary results are presented for all subtests.

Stem-and-leaf plots of the MHCHISQ indices for the grade comparisons on GS are shown in Table 19. With one degree of freedom, chi-square values greater than 3.84 are significant at the .05 level, while those greater than 10.83 are significant at the .001 level. For grade 10 vs grade 11, 18 of the 25 items have MHCHISQ values that are significant at the .05 level and 11 are significant at the .001 level. The corresponding figures for the grade 11 vs grade 12 comparison are 14 at the .05 level and 10 at the .001 level.

Table 17. Spearman Rank-Order Correlations Among Selected Unsigned Differential Item Functioning Indices by Group Comparison and Subtest

Groups	Test	FCHISQ with MHCHISQ	FCHISQ with LCHISQ	FCHISQ with AREA	FCHISQ with MSOS	MHCHISQ with LCHISQ	MHCHISQ with AREA	MHCHISQ with MSOS	LCHISQ with AREA	LCHISQ with MSOS	AREA with MSOS
W-B	GS	.79	.51	.60	.84	.73	.61	.74	.55	.66	.85
	AR	.92	.56	.38	.50	.57	.42	.55	.51	.73	.90
	WK	.74	.52	.65	.69	.48	.61	.65	.88	.87	.93
	PC	.95	.54	.80	.70	.57	.76	.73	.81	.82	.93
	AS	.92	.59	.31	.47	.51	.17	.40	.56	.71	.83
	MK	.81	.41	.32	.60	.37	.22	.41	.71	.71	.90
	MC	.67	.37	.37	.43	.41	.30	.35	.78	.87	.90
	EI	.93	.66	.51	.79	.55	.42	.65	.58	.77	.75
W-B Median		.86	.53	.44	.64	.53	.42	.60	.64	.75	.90
W-H	GS	.86	.37	.60	.71	.27	.57	.60	.68	.79	.90
	AR	.66	.27	.12	.09	.37	.21	.20	.68	.77	.91
	WK	.41	.32	.23	.25	-.01	.27	.27	.76	.66	.88
	PC	.72	.30	.45	.40	.11	.46	.42	.63	.59	.98
	AS	.90	-.02	-.21	-.16	.06	-.08	-.02	.63	.66	.88
	MK	.60	.41	.26	.43	.03	.05	.29	.82	.83	.90
	MC	.90	.25	.01	.00	.07	.05	.13	.40	.20	.77
	EI	.95	.12	-.34	-.49	.20	-.21	-.34	-.02	-.27	.82
W-H Median		.79	.28	.17	.15	.09	.13	.23	.65	.66	.89
M-F	GS	.97	.57	.75	.84	.49	.72	.79	.63	.75	.89
	AR	.88	.60	.69	.84	.43	.54	.72	.78	.84	.90
	WK	.89	.59	.41	.65	.60	.44	.69	.64	.85	.74
	PC	.98	.36	.51	.64	.36	.52	.65	.71	.75	.91
	AS	.93	.27	.20	.49	.20	.06	.34	.47	.60	.81
	MK	.84	.58	.37	.49	.43	.19	.41	.85	.87	.87
	MC	.84	.59	.50	.66	.51	.22	.47	.69	.83	.84
	EI	.84	.43	.06	.61	.24	-.21	.42	.62	.80	.68
M-F Median		.88	.57	.45	.64	.43	.34	.56	.63	.81	.85
Overall Median		.87	.42	.39	.55	.39	.28	.42	.63	.75	.88

It is obvious that with large sample sizes, even small differences can be statistically significant. Thus, it is mandatory to consider the magnitude of the differences as well as the statistical significance of the results for a given item. Practical importance cannot be sacrificed to statistical significance. The MHODDS statistic and the transformation of the odds ratio to differences in the delta scale metric are useful for this purpose. The five largest MHCHISQ values from Table 19 and the corresponding values of MHODDS and the transformation of MHODDS to a Delta difference, DD = (-2.35) ln (MHODDS), are as follows: (a) MHCHISQ = 123.0, MHODDS = .75, DD = .68; (b) MHCHISQ = 76.0, MHODDS = .80, DD = -.52; (c) MHCHISQ = 69.3, MHODDS = .76, DD = .64; (d) MHCHISQ = 56.2, MHODDS = 1.29, DD = -.60; and (e) MHCHISQ = 52.3, MHODDS = .84, DD = .41. None of the other items has more extreme

Table 18. Spearman Rank-Order Correlations Among Selected Directional Differential Item Functioning Indices by Group Comparison and Subtest

Groups	Subtest	RESID with SFCHISQ	RESID with MHODDS	RESID with BHA	SFCHISQ with MHODDS	SFCHISQ with BHA	MHODDS with BHA
W-B	GS	.83	.85	.80	.96	.81	.84
	AR	.70	.74	.34	.97	.64	.57
	WK	.61	.57	.55	.94	.61	.67
	PC	.89	.84	.71	.90	.64	.71
	AS	.84	.85	.74	.99	.65	.61
	MK	.77	.80	.60	.91	.63	.49
	MC	.79	.80	.57	.97	.58	.53
	EI	.33	.35	.61	.96	.37	.30
W-B Median		.78	.80	.66	.96	.62	.59
W-H	GS	.91	.95	.61	.98	.70	.72
	AR	.84	.91	.31	.89	.30	.15
	WK	.66	.84	.72	.86	.54	.60
	PC	.66	.70	.53	.93	.52	.46
	AS	.86	.87	.69	.99	.59	.60
	MK	.87	.88	.50	.95	.35	.23
	MC	.91	.89	.68	.98	.61	.60
	EI	.67	.70	.49	.98	.38	.44
W-H Median		.85	.87	.57	.97	.53	.53
M-F	GS	.04	.81	.75	.13	.20	.93
	AR	.50	.72	.61	.37	.51	.66
	WK	.58	.82	.53	.85	.77	.74
	PC	.47	.82	.59	.78	.70	.66
	AS	.32	.87	.74	.25	.28	.78
	MK	.82	.94	.63	.91	.71	.63
	MC	.31	.50	.78	.31	.42	.63
	EI	.24	.95	.78	.29	.02	.70
M-F Median		.39	.82	.68	.34	.46	.68
Overall Median		.69	.83	.61	.92	.58	.62

odds ratios or delta differences than those listed. ETS uses a DD of 1.0 or larger to flag items that show group differences large enough to be of practical importance (Holland, personnal communication, 1986). Thus, even the highly significant MHCHISQ values for the grade-to-grade comparisons are associated with between-group differences that are less than the 1.0 criterion of practical importance on the delta scale.

Based on these results, the MHCHISQ values were divided into four ranges: (a) low (less than 10), (b) moderate (between 10 and 100), (c) high (between 100 and 200), and (d) extreme (greater than 200). The distributions of the MHCHISQ values in these ranges are shown in Table 20 for all eight subtests and each of the five pairs of groups that were compared. As can be

seen, none of the 200 items on the eight subtests had MHCHISQ values in the "extreme" range for either of the grade-to-grade comparisons, and only 4 items (1 for the grades 10 vs 11, and 3 for the grades 11 vs 12) had MHCHISQ values in the "high" range.

The number of items in the high or extreme ranges was substantially higher for each of the racial/ethnic group comparisons and for the gender comparison than for the two grade-to-grade comparisons. The difference is most notable for the gender comparison, where 52 of the 200 items had MHCHISQ values in the extreme range and another 29 items were in the high range. The corresponding figures for the White-Black comparison were 14 extreme and 22 high; and for the White-Hispanic comparison, there were 6 extreme and 9 high.

It should be noted that the large number of high and extreme values in the male-female comparison is, at least partially, a function of the great power of the chi-square test. Recall that there are slightly over 20,000 individuals in each group for the male-female comparison. The grade-to-grade comparisons are also quite powerful, however. There are approximately 8,000, 21,000, and 13,000 individuals in grades 10, 11, and 12, respectively. It should also be noted that the racial/ethnic group comparisons, which have a substantial number of items with indices in the high or extreme ranges, are less powerful than the grade-to-grade comparisons because there are fewer Black (approximately 5,000) and Hispanic (slightly over 2,000) examinees in the total sample than there are examinees in any of the three grade-level samples.

Table 19. Stem-and-Leaf Plots of MHCHISQ Statistics for the Grade Comparisons on the GS Subtest

Grade 10 vs Grade 11 Leaf (units digit)	Stem (tens)	Grade 11 vs Grade 12 Leaf (units digit)
	12	3
	*	
.	7	6
	7	
9	6	
	6	
	5	6
1	5	2
	4	
2	4	
9	3	5
2	3	2
76	2	
3	2	0
55	1	8
30	1	224
877775	0	89
3221100	0	000011222334

Table 20. Distributions of MHCHISQ Values by Subtest and Total Across Subtests

Range[a]	GS	AR	WK	PC	AS	MK	MC	EI	Total
Grade 10 and Grade 11									
Low	14	28	31	12	14	18	22	17	156
Moderate	11	2	3	3	11	7	3	3	43
High	-	-	1	-	-	-	-	-	1
Extreme	-	-	-	-	-	-	-	-	0
Grade 11 and Grade 12									
Low	14	26	23	14	12	14	24	17	144
Moderate	10	3	11	1	13	11	1	3	53
High	1	1	1	-	-	-	-	-	3
Extreme	-	-	-	-	-	-	-	-	0
White-Black									
Low	8	11	14	6	7	12	15	7	80
Moderate	13	14	14	7	6	12	10	8	84
High	1	4	2	2	8	1	-	4	22
Extreme	3	1	5	-	4	-	-	1	14
White-Hispanic									
Low	9	23	11	9	9	18	11	5	95
Moderate	11	7	21	5	13	7	14	12	90
High	2	-	1	1	3	-	-	2	9
Extreme	3	-	2	-	-	-	-	1	6
Male -Female									
Low	3	4	7	-	2	6	8	2	32
Moderate	7	15	17	7	7	15	9	10	87
High	4	6	5	3	6	1	2	2	29
Extreme	11	5	6	5	10	3	6	6	52

[a] Low: MHCHISQ $<$ 10.
Moderate: 10 $\leq$ MHCHISQ $<$ 100.
High: 100 $\leq$ MHCHISQ $<$ 200.
Extreme: 200 $\leq$ MHCHISQ.

Similar comparisons among distributions were made for the other differential item functioning indices. For reasons of space, however, distributions for only two other indices are presented here. Table 21 reports the distributions of the MSOS indices, and Table 22 presents the distributions of the MHODDS indices. The latter index takes the direction of the difference into account and thereby provides additional descriptive information to go along with the unsigned MHCHISQ and MSOS indices. Unlike

MHCHISQ (Table 20), the MSOS and MHODDS indices are descriptive statistics that do not depend directly on the number of individuals involved in a comparison.

As can be seen in Table 21, none of the 200 items on the eight subtests has an MSOS value greater than .02 for the grade 11 vs grade 12 comparison, and only 1 item has an MSOS value greater than .02 in the grade 10 vs grade 11 comparison. On the other hand, 16, 18, and 23 items have MSOS values greater than .02 in the White-Black, the White-Hispanic, and the male-female comparisons, respectively. Furthermore, there are 7 items in the White-Black comparison, 3 in the White-Hispanic comparison, and 7 in the male-female comparison that have extreme MSOS values of .04 or higher.

There were two subtests, Arithmetic Reasoning and Mathematics Knowledge, that did not have any items with MSOS values greater than .02 for either of the racial/ethnic group comparisons. General Science, on the other hand, had three items with extreme MSOS values for both the White-Black and the White-Hispanic comparisons.

The MHODDS results in Table 22 divide the items into three categories: items that favor the reference group (i.e., grade 11, White, or male examinees), items for which the likelihood of getting an item right is similar for the two groups after controlling for total score, and items that favor the focal group (i.e., grade 10, grade 12, Black, Hispanic, or female examinees). The cutting points used correspond to delta scale differences of roughly 1.0. Thus, when an item is classified as "Ref. Foc.," the performance of examinees in the reference group is approximately one delta unit or more higher than that of examinees in the focal group, after matching on total test score.

In terms of number of items classified as "approximately equal" using the MHODDS values, there is a close correspondence with the numbers that would be considered essentially equivalent using an MSOS cutoff of .01. That is, 98.0% of the grade 10 vs grade 11 MSOS values are less than .01, and 99.5% of the items are classified as having approximately equal MHODDS for these two groups. The corresponding percentages for the other four comparisons are: grades 11 and 12, 99.5% and 99.5%; White-Black, 82.0% and 83.0%; White-Hispanic, 79.0% and 84.5%; and male-female, 77.5% and 76.5%.

The agreement between the MSOS and MHODDS for classifying specific items is far from perfect. However, for the White-Black and the male-female comparisons, items with MSOS values of .02 or higher also have low (less than .7) or high (greater than 1.5) odds ratios. Fourteen of the 16 items (87.5%) with MSOS values greater than .02 in the White-Black comparison have low or high MHODDS values, and 19 of the 23 items (82.6%) of the MSOS values greater than .02 in the male-female comparison have low or high MHODDS values according to the above criteria. These figures can be compared to base rates of 17% and 23.5% for the White-Black and male-female comparisons, respectively. On the other hand, only 7 of the 18 items (38.9%) with MSOS values greater than .02 in the White-Hispanic comparisons have MHODDS values less than .7 or greater than 1.5, which is only slightly better than the base rate of 31 out of 200 items (15.5%) that have low or high MHODDS values.

Table 21. Distributions of the MSOS Values by Subtest and Total Across Subtests

Range[a]	GS	AR	WK	PC	AS	MK	MC	EI	Total
				Grade 11 - Grade 10					
Low	23	30	34	15	25	25	25	19	196
Moderate	1	-	1	-	-	-	-	1	3
High	1	-	-	-	-	-	-	-	1
Extreme	-	-	-	-	-	-	-	-	0
				Grade 11 - Grade 12					
Low	25	30	35	15	24	25	25	20	199
Moderate	-	-	-	-	1	-	-	-	1
High	-	-	-	-	-	-	-	-	0
Extreme	-	-	-	-	-	-	-	-	0
				White-Black					
Low	20	27	29	13	16	23	22	14	164
Moderate	1	3	2	1	3	2	2	6	20
High	1	-	2	-	5	-	1	-	9
Extreme	3	-	2	1	1	-	-	-	7
				White-Hispanic					
Low	18	30	31	14	21	24	9	11	158
Moderate	2	-	1	-	2	1	12	6	24
High	2	-	3	1	2	-	4	3	15
Extreme	3	-	-	-	-	-	-	-	3
				Male-Female					
Low	17	23	29	11	17	24	19	15	155
Moderate	4	5	4	3	1	-	3	2	22
High	3	1	2	1	5	1	2	1	16
Extreme	1	1	-	-	2	-	1	2	7

[a] Low: MSOS $<$ .01.
Moderate: .01 $\leq$ MSOS $<$ .02.
High: .02 $\leq$ MSOS $<$ .04.
Extreme: .04 $\leq$ MSOS.

Interpretation of MSOS and MHODDS by Subtest

General Science. Four GS items had MSOS values greater than .02 in the Black-White comparison. The direction and magnitude of the difference were evaluated by using the MHODDS ratios. An item that is flagged by an MSOS value greater than .02 and has an MHODDS ratio greater than 1.5 favors examinees in the base or reference group in the sense that their odds of

Table 22. Distributions of MHODDS by Subtest and Total Across Subtests

Range[a]	GS	AR	WK	PC	AS	MK	MC	EI	Total
Grade 11 - Grade 10									
Ref. > Foc.	-	-	1	-	-	-	-	-	1
Approx Eq.	25	30	34	15	25	25	25	20	199
Foc. > Ref.	-	-	-	-	-	-	-	-	0
Grade 11-Grade 12									
Ref. > Foc.	-	-	-	1	-	-	-	-	1
Approx Eq.	25	30	35	14	25	25	25	20	199
Foc. > Ref.	-	-	-	-	-	-	-	-	0
White-Black									
Ref. > Foc.	3	3	6	1	4	-	-	2	19
Approx Eq.	21	27	24	14	15	24	25	16	166
Foc. > Ref.	1	-	5	-	6	1	-	2	15
White-Hispanic									
Ref. > Foc.	3	-	6	1	2	-	1	4	17
Approx Eq.	18	30	26	14	19	25	24	13	169
Foc. > Ref.	4	-	3	-	4	-	-	3	14
Male-Female									
Ref. > Foc.	4	2	4	2	6	1	2	2	23
Approx Eq.	16	26	28	10	15	23	20	15	153
Foc. > Ref.	5	2	3	3	4	1	3	3	24

[a] Reference > Focal: MHODDS $\geq$ 1.5.
Approximately Equal: .7 < MHODDS < 1.5.
Focal > Reference: MHODDS $\leq$ .7

getting the item right are at least 1.5 times as great as their matched counterparts in the comparison or focal group. On the other hand, items with MHODDS values less than .7 favor examinees in the comparison or focal group. Three of the four GS items with MSOS values greater than .02 in the Black-White comparison were found to favor White examinees according to their MHODDS values (greater than 1.5), while the fourth item favored Black test-takers (MHODDS ratio less than .7). For reasons of test security, the specific items cannot be presented here. However, these items were reviewed and compared to the test specifications for GS.

Recall that the GS subtest covers three general content domains, physical science, life science, and earth science. All three items favoring Whites according to the above criteria were in the physical science category, while the item favoring Blacks was a life science item. The means and standard

deviations of the MHODDS for all GS items within each content category are as follows: physical science, mean = 1.31, SD = .49; life science, mean = .89, SD = .31; and earth science, mean = .99, SD = .13.

Similar results were obtained for the GS subtest in the White-Hispanic comparison. Of the five items with MSOS values greater than .02, three items had MHODDS values greater than 1.5, and the other two had MHODDS values less than .7. The means and standard deviations of the MHODDS values for all items within the three content areas are: physical science, mean = 1.36, SD = .71; life science, mean = .82, SD = .27; and earth science, mean = 1.12, SD = .34.

On average, the odds of a White examinee giving the correct answer on a physical science item are about 1.3 times as great as for a Black or Hispanic examinee with the same overall performance on the test. On the other hand, the odds that a Black or Hispanic examinee will answer a life science item correctly are about 1.1 to 1.2 (the reciprocals of .89 and .82) times as great as for a matched White examinee. As indicated by the standard deviations, however, there is substantial variability of the MHODDS values within each content category.

In the male-female comparison, the four items with MSOS values greater than .02 are again all associated with items that have either high or low MHODDS values. The distinction of items favoring men and those favoring women by content category, however, is less clear than in the racial/ethnic comparisons. Two of the items with MSOS values greater than .02 favor men (MHODDS greater than 1.5) and two favor women (MHODDS less than .7). One of the two items favoring men is from the physical science area, and the other from the earth science area. Of the two favoring women, one is a life science item, but the other is a physical science item. The means and standard deviations of the MHODDS values by content area are: physical science, mean = 1.20, SD = .41; life science, mean = .98, SD = .41; and earth science, mean = 1.05, SD = .61. Thus, though there may be some indication that men are more likely to do better than women on physical science items after matching on total score, the tendency is, at best, weak; and there is considerable variability from item to item within a content area.

Arithmetic Reasoning. None of the 30 items on AR had MSOS values greater than .02 for either of the racial/ethnic group comparisons, and only two items had MSOS values greater than .02 for the male-female comparison. Both of the items so identified in the male-female comparison had low MHODDS values (less than .7). However, no obvious interpretations were forthcoming from a review of the items. One of the items involved the calculation of the area of a rectangle, and the other item was a time/rate problem. Since the MHODDS values for the time/rate problems ranged from a low of .62 (the identified item) to a high of 1.44 and had a mean of .95, it seems unwise to attempt any generalizations about the characteristics of AR items that may favor men or women relative to their matched counterparts.

Word Knowledge. Four WK items had MSOS values greater than .02 in the White-Black comparison. Three of these favored Whites (MHODDS greater than 1.5) and the other item favored Blacks (MHODDS less than .7). Of the three items with MSOS values greater than .02 in the White-Hispanic comparison, two

favored Whites; and the third had an intermediate MHODDS value of .78, which is only slightly above the lower cutoff of .7 for an item to be classified as favoring Hispanics.

Two item types are used to assess knowledge of synonyms on the WK subtest. The stem of the item is either of the form "_____ most nearly means ...." or the stem presents a complete sentence with one word underlined (DoD, 1985, p. 60). There was no distinction, however, among items favoring Whites and those favoring Blacks or Hispanics in terms of item type.

A review of the words that were identified as functioning differently for Whites and Blacks or for Whites and Hispanics did not lead to any clear rules that would help identify such items in advance. However, two of the three words that favored Whites over Blacks, one of which was also one of the two words that favored Whites over Hispanics, are words that are most likely to be encountered in science classes. Support for this conclusion comes from the Carroll, Davies, and Richman (1971) word frequency study, in which 5,088,721 words of running text from 1,045 publications were analyzed. Carroll et al. broke the frequency of occurrence counts down by subject matter of the source publication. One of the two words mentioned above occurred 21 times in the Carroll et al. (1971) study: 14 times in science texts, 6 times in library reference materials, and once in art texts. It never occurred in any of the other 14 categories of text. The second word occurred a total of 6 times: 3 times in science texts and 3 times in library reference materials. None of the other words in the WK subtest was so clearly associated with a particular subject matter.

In the male-female comparison, both words with MSOS values greater than .02 had MHODDS values greater than 1.5 (i.e., they favored men). One of those items involves the more frequent of the science-related words described above. The other item involves a sentence concerning the use of a shop tool, though the target word may be used in a variety of other contexts.

Paragraph Comprehension. A single item was identified on the PC subtest in all three racial/ethnic group and gender comparisons. The MHODDS values indicated that the item favored Blacks and Hispanics over Whites and men over women after controlling for overall performance on the PC subtest. The item consists of a 25-word stem which contains one complete sentence and a second partial sentence that is to be completed by the appropriate choice of a phrase from among the four options. Since the other items involving this "complete-the-sentence" format did not show consistent between-group differences, there is no basis for concluding that this format produced the differences observed for the single PC item.

Auto and Shop Information. Six items on the AS subtest had MSOS values greater than .02 for the White-Black comparison: Two of these favored Whites; three favored Blacks; and the other item was in the intermediate category according to the MHODDS criteria, though the MHODDS value of 1.46 for the item fell just below the cutting point. All three items favoring Whites (including the one with MHODDS = 1.46) are in the automotive information domain. Two of the three items favoring Blacks, on the other hand, are concerned with the proper use of hand tools for particular purposes in the shop. This division by content area is violated, however, by the remaining item favoring Black examinees. Furthermore, a comparison of the

MHODDS ratios for all automotive items with those for shop items provides no support for an hypothesis that differential item functioning is explained by this content categorization. The mean MHODDS for the automotive items is 1.11, while the mean for shop items is 1.02.

Only two AS items had MSOS values greater than .02 for the White-Hispanic comparison. Neither of these items had high or low MHODDS values according to the criteria used. Furthermore, there were no apparent consistent content differences for the items with high or low MHODDS values. Hence, no interpretation of the differential functioning indices for the White-Hispanic comparison is suggested.

Five of the seven AS items with high MSOS values in the male-female comparison also had MHODDS values above 1.5 (i.e., favored men). The remaining two items identified by the MSOS criterion had MHODDS values of 1.45 and 1.18. Thus, men are more likely to answer all seven identified items correctly than are women with equal overall performance on the AS subtest. Five of the seven identified items deal with the proper use of shop tools. The remaining two items deal with lubrication, albeit in the automotive context.

Mathematical Knowledge. None of the MK items had MSOS values greater than .02 for either of the racial/ethnic group comparisons, and only one item met this criterion in the male-female comparison. The MHODDS indicated that the latter item favored women. A single item provides no basis for generalization, and none will be attempted.

Mechanical Comprehension. The single item on the MC subtest with an MSOS value greater than .02 for the White-Black comparison had an intermediate MHODDS value of 1.41. Similarly, all four of the items that satisfied the MSOS criterion for the White-Hispanic comparison had intermediate MHODDS values. However, all four of the latter had similar MHODDS values (.76, .76, .79 and .84), suggesting that the identified items all were somewhat easier for Hispanic examinees (i.e., they favored Hispanics) than for White examinees with comparable overall performance on the MC subtest. Since the four items varied considerably in terms of content (spring, valves, levers, and gears), no interpretation of the apparent difference is evident.

In the male-female comparison, three MC items exceeded the MSOS criterion. One of these items had an extremely large MHODDS value (2.70), while the other two had MHODDS values of .7 or less. The item that men were 2.7 times as likely as women with comparable overall performance on the MC to answer correctly contained two technical words, one of which did not occur in over 5,000,000 words analyzed by Carroll, Davies, and Richman (1971); and the other one occurred only in the science, shop, and magazine categories. The two items favoring women dealt with a piston and a pump, but no basis for generalization was apparent.

Electronics Information. None of the EI items had MSOS values greater than .02 in the comparison of White and Black examinees. Although there were three items that satisfied this criterion in the White-Hispanic comparison, none of the three was categorized as high or low according to the MHODDS criteria.

Two of the three items on the EI subtest with MSOS values greater than .02 for the male-female comparison had MHODDS values greater than 1.5. Both of these items involved the definition of a word. According to the Carroll, Davies, and Richman (1971) study, one of these words occurred only in science (7 times), math (2 times), and library reference (1 time) texts. The other word occurred most frequently in shop (7 times), followed by science (5 times), reading (5 times), and library reference (1 time) texts. Neither of these words was found in any of the other content area texts.

V. DISCUSSION AND CONCLUSIONS

A total of 27 differential item functioning indices were computed for each of the 200 items on the eight non-speeded tests of the ASVAB for each of 5 group comparisons (grade 10 with grade 11 students, grade 11 with grade 12 students, White with Black students, White with Hispanic students, and male with female students). Greater variability in the magnitude of all 27 indices was observed when the comparison involved gender or racial/ethnic groups than for either of the grade-to-grade comparisons. Higher means were also found in the gender and racial/ethnic group comparisons for the unsigned (i.e., nondirectional indices such as the chi-square or sum of squares indices). In other words, regardless of the index used, some items are apt to be flagged as functioning differently for Black than White, for Hispanic than White, or for women than men.

The relationships among the indices--whether judged in terms of Pearson product-moment correlations, Spearman rank-order correlations, or agreement on items flagged for differential functioning--are quite variable. A few indices, such as the four that are based on the differences between groups in the point-biserials of an item with the total subtest score or some transformation of the point-biserials, were found to be related to each other but had little relationship with any of the other indices. In general, as would be expected, similarly defined indices, such as the several types of IRT-based sum of squared difference indices or the "full chi-square" and the Mantel-Haenszel chi-square statistics, were found to be highly related, and would, in most cases, lead to the identification of the same items in either of the racial/ethnic group comparisons or the male-female comparison.

The Mantel-Haenszel chi-square and the Shepard, Camilli, and Williams (1984) modified sum of squared differences between item response curves had moderately high Spearman rank-order correlations for the White-Black (median $r = .60$) and the male-female (median $r = .56$) comparisons. The median correlation across the eight subtests for the White-Hispanic comparison, however, was only .23.

Emphasis for interpretive purposes was given to two indices, one based on observed scores and one based on IRT. By selecting indices based on different rationales, but for which there are strong theoretical and empirical justifications, it was reasoned that the likelihood would be reduced that an item would be identified as one that functions differently for different groups simply as the result of analytical artifacts. Based on theoretical justifications, empirical results from previous research studies, and the results of the present analyses, the MSOS and MHODDS indices were selected as the combination that is most likely to lead to unambiguous identification of items (Holland & Thayer, 1986; Shepard, Camilli, & Williams, 1984).

Two of the subtests, Arithmetic Reasoning and Mathematical Knowledge, were found to have few, if any, items that function differently for White, Black, or Hispanic examinees. There was also little, if any, indication that items on these two tests function differently for men and women. On the remaining six non-speeded subtests, however, there were some items on which the performance of one group of examinees was clearly better than that of another group, even after adjusting for differences in overall performance on the subtest total score or the estimated latent ability from the item response theory analyses.

It should be noted, however, that the direction of the difference on items with large indices varied from item to item in each of the group comparisons. In other words, some items with large indices favored one group, whereas other items favored the other group. It is not the case that such items always favored White examinees over Black or Hispanic examinees or men over women. Consequently, there is, at least, some tendency for items favoring one group to be balanced by items favoring the other group.

In trying to relate this finding to the results of the differential item functioning analyses, relatively few general conclusions about the characteristics of items that are apt to function differently for White and Black test-takers, for White and Hispanic examinees, or for male and female examinees appeared to be justified. Nonetheless, a few generalizations that may be useful either as hypotheses for future analyses of the ASVAB (and other multiple aptitude test batteries) or as considerations in future item development and test construction seem worthy of consideration.

1. After control for overall performance on the General Science subtest, White examinees tend to do better than Black or Hispanic examinees on physical science items, while the converse is true for life science. Although the reasons for this difference are unclear, it is possible that it is the result of different course-taking patterns in high school. In any event, the result underscores the importance of maintaining balance among the content domains covered by the General Science subtest, which, according to the Technical Supplement to the Counselor's Manual for the ASVAB, Form 14 (DoD, 1985, p. 59), include "approximate weightings ... [of] 45%, 45%, and 10%" for physical science, life science, and earth science, respectively.

2. On the Word Knowledge subtest, there was some indication that words found in science textbooks--but not in other texts that students are likely to encounter in school--favor White examinees over Black or Hispanic examinees with comparable overall performance on the subtest. To a lesser extent, such words may also favor men over women. Consideration of the distributional characteristics provided by Carroll, Davies, and Richman (1971) in the selection of vocabulary words for future editions of the Word Knowledge subtest may be desirable.

3. On both the Mechanical Comprehension and the Electronics Information subtests, the items that most clearly favored men over women with comparable overall performance on the subtests required knowledge of vocabulary that is most likely found in science, shop, or math texts. The extent to which specialized vocabulary is an essential part of the constructs that these subtests are intended to measure deserves consideration.

Although generalizations from this and previous investigations of differential item functioning are limited and provide relatively little practical guidance for test construction, the importance of the issue demands continued attention. The incorporation of results of differential item functioning analyses as part of the routine item analysis performed in test construction appears desirable. In this way, it may be possible to accumulate a large enough collection of flagged items to make generalizations possible. It could also provide a means of avoiding the use of items with unusually large differences and instead using available alternative items that better satisfy the operational test specifications.

REFERENCES

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore, MD: Johns Hopkins University Press.

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

Baker, F. B. (1981). A criticism of Scheuneman's item bias technique. Journal of Educational Measurement, 18, 59-62.

Camilli, G. (1979). A critique of the chi-square method for assessing item bias. Unpublished paper, Laboratory of Educational Research, University of Colorado, Boulder.

Cardall, C., & Coffman, W. E. (1964). A method for comparing the performance of different groups on the items in a test (Reseach Bulletin 64-61). Princeton, NJ: Educational Testing Service.

Carroll, J. B., Davies, P., & Richman, B. (1971). The American Heritage Word Frequency Book. Boston: Houghton Mifflin.

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

Cole, N. S. (1981). Bias in testing. American Psychologist, 36, 1067-1077.

Department of Defense. (1984). Counselor's manual for the Armed Services Vocational Aptitude Battery Form 14. Chicago, IL: U. S. Military Entrance Processing Command.

Department of Defense. (1985). Technical supplement to the counselor's manual for the Armed Services Vocational Aptitude Battery, Form 14. Chicago, IL: U. S. Military Entrance Processing Command. (See p. 45.)

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

Green, D. R., & Draper, J. F. (1972). Exploratory studies of bias in achievement tests. Paper presented at the annual meeting of the American Psychological Association, Honolulu.

Holland, P. W. (1985). On the study of differential item performance without IRT. Proceedings of the Military Association, pp. 282 - 287.

Holland, P. W., & Thayer, D. T. (1986). Differential item functioning and the Mantel-Haenszel procedure. (Research Report No. 86-31). Princeton, NJ: Educational Testing Service.

Hunter, J. E. (1975). A critical analysis of the use of item means and item-test correlations to determine the presense or absense of content bias in achievement test items. Paper presented at the National Institute of Education Conference on Test Bias, Annapolis, MD.

Ironson, G. H. (1982). Use of chi-square and latent trait approaches for detecting item bias. In R. A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore, MD: Johns Hopkins University Press.

Linn, R. L., & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test item. Journal of Educational Measurement, 18, 109-118.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. Applied Psychological Measurement, 5, 159-173.

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

McPeek, W. M., & Wild, C. J. (1986). Performance of the Mantel-Haenszel statistic in a variety of situations. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Mislevy, R. J., & Bock, R. D. (1984). BILOG: Item analysis and test scoring with binary logistic models. Mooresville, IN: Scientific Software, Inc.

Office of the Assistant Secretary of Defense. (1982). Profile of American Youth: 1980 nationwide administration of the Armed Services Vocational Aptitude Battery. Washington, DC: Department of Defense.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). A Monte Carlo comparison of seven biased item detection techniques. Journal of Educational Measurement, 17, 1-10.

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

Shepard, L. A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. Journal of Educational Statistics, 6, 317-375.

Shepard, L. A., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. Journal of Educational Statistics, 9, 93-128.

Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22, 77-105.

* GOVERNMENT PRINTING OFFICE: 1988 - 561048 / 80003

END
DATE
FILMED
7-88
DTIC