NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER SAN DIEGO, CALIFORNIA 92152

NPRDC TR 76-6

JULY 1975

COMPARATIVE RACIAL ANALYSIS OF ENLISTED ADVANCEMENT EXAMS: ITEM-DIFFICULTY

D. W. Robertson
M. H. Royle

APPROVED FOR PUBLIC RELEASE;
DISTRIBUTION UNLIMITED.

NPRDC TR 76-6 July 1975

COMPARATIVE RACIAL ANALYSIS OF ENLISTED ADVANCEMENT EXAMS: ITEM-DIFFICULTY

David W. Robertson
Marjorie H. Royle

Reviewed by
Robert P. Thorpe

Approved by
James J. Regan
Technical Director

Navy Personnel Research and Development Center
San Diego, California 92152

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NPRDC TR 76-6	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) COMPARATIVE RACIAL ANALYSIS OF ENLISTED ADVANCEMENT EXAMS: ITEM-DIFFICULTY		5. TYPE OF REPORT & PERIOD COVERED Final Report (May 1974 - July 1975
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) David W. Robertson Marjorie H. Royle		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Navy Personnel Research and Development Center San Diego, California 92152		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 62763N PF55.521.032.01.01
11. CONTROLLING OFFICE NAME AND ADDRESS		12. REPORT DATE July 1975
		13. NUMBER OF PAGES 49
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Item-analysis
Promotion
Racial comparison
Equal opportunity

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

An item-analysis of 24 Navy Enlisted Advancement Exams was conducted to determine which test characteristics might account for the higher promotion rate of White than Black racial groups. Specific questions addressed included (1) whether it is feasible to construct exams containing only items which are similar in difficulty for both Blacks and Whites, (2) what types of items are similar in difficulty, and (3) whether the same items are relatively easy or difficult for Blacks and Whites.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. ABSTRACT (continued)

The proportion of items identified as similar in difficulty for both Blacks and Whites varied from about one-half to six-sevenths of the 150 items in each test. The similar-type items were concentrated in the difficult range, and presented applied (as distinguished from conceptual) content. Relative item-difficulty was low on some exams.

The development of advancement exams of items similar in difficulty for Blacks and Whites could not be recommended, because the concentration of similar-difficulty items in the difficult range would degrade test quality, and items largely limited to factual content might not cover all necessary content for a particular occupational specialty.

FOREWORD

This study was initiated in response to a request from the Chief of Naval Personnel (Pers-6) to determine the feasibility of developing Enlisted Advancement Exams from items similar in difficulty for both Black and White racial groups, as an approach to improving equal opportunity in career growth for minority groups. The study of exam item-difficulty levels is the first of a series of technical reports on comparative racial analyses of Enlisted Advancement Factors. (Other studies nearing completion include two reports on item-differentiation and relative item-difficulty.)

The substantial and valuable assistance of the following persons is gratefully acknowledged: Ms. Penny L. Allison, Mr. David J. Morena, and DP2 Suzanne Olson, for data processing and computation; and Ms. Hazel F. Schwab, for clerical support.

This study was performed under Exploratory Development Task Area PF55.521.032 (Contemporary Social Issues).

J. J. CLARKIN
Commanding Officer

SUMMARY

Purpose

A number of studies have found that a greater proportion of Caucasian than non-Caucasian candidates are advanced in the military services. The present study, an item-analysis of Navy Enlisted Advancement Exams, investigated various test characteristics which might account for differences between Black and White racial groups. Specific questions addressed included: (1) whether it is feasible to construct exams containing only items which are similar in difficulty for both Blacks and Whites, (2) what types of items are similar in difficulty, and (3) whether the same items are relatively easy or difficult for both Blacks and Whites.

Approach

The exams of six occupational specialties across the four Pay Grades E4 through E7 of each specialty (i.e., 24 different exams) were analyzed. Item-difficulties (P values) were calculated separately for the two racial groups. Two types of simulated tests were constructed, one of items similar in difficulty, and one without excessively difficult items. Changes in racial differences between the existing and simulated tests were identified.

Findings and Conclusions

The proportion of items identified as similar in difficulty for both Blacks and Whites varied from about one-half to six-sevenths of the 150 items in each test (page 7). The similar-type items were concentrated in the difficult range, and presented applied (as distinguished from conceptual) content (page 16). Tests constructed of similar-type items would reduce, but not eliminate, differences in exam scores between Blacks and Whites (pages 16, 21, 34). The relative item-difficulty between Blacks and Whites is low in some rate groups (as indicated by a low correlation, Rho value, between the two rank-orders of the Black and White item-difficulty levels), which suggests a possibility of racial bias in those exams. However, no conclusion was reached, pending completion of another on-going study (pages 27, 35).

Recommendation

The development of advancement exams with items similar in difficulty for Blacks and Whites cannot be recommended, because the concentration of similar-difficulty items in the difficult range would degrade test quality (page 33), and items largely limited to factual content may not cover all necessary content for a particular occupational specialty (pages 35, 36).

CONTENTS

LIST OF TABLES

LIST OF TABLES (CONT.)

LIST OF TABLES (CONT.)

INTRODUCTION

Background

Many aspects of personnel selection and management in the Navy are being studied to identify and alleviate conditions which might be detrimental to equal opportunity for all individuals and groups. Since promotion is a major factor in career opportunity and growth in an organization, the Enlisted Advancement System was one of the personnel systems selected for study regarding comparative racial opportunity for career growth.

Advancement competition. Different procedures are employed for each of three pay grade groups.

1. Advancement to Pay Grades 2 and 3 is noncompetitive, and simply requires demonstrated performance and knowledge in a technical school or on the job after serving a minimum period of time.

2. Advancement to Pay Grades 4, 5 and 6 is competitive, and is based on several differentially weighted factors, including technical knowledge, on-job performance, time in pay grade, total time in the Navy, and commendations. (Until recently, competition to Pay Grade 7 was also within this procedure. The data used in this analysis, and described below, include Pay Grade 7 competitions. The new Pay Grade 7 procedures are described in the next paragraph below.)

3. Advancement to Pay Grades 7, 8 and 9 is also competitive. Twenty five to fifty percent of the candidates are eliminated by preliminary screening on technical knowledge and performance factors. Selection Boards base their final selections on the factors which they consider most relevant.

Advancement Exams. The Technical Knowledge Factor is measured by a 150-item multiple choice test with four alternative answers per item. A separate test is developed for each of approximately 80 Navy ratings (i.e., the Navy term for occupational specialties), and also for each pay grade within each rating. Each test comprises approximately 6 to 10 content areas or sections.

Problem

A number of studies have found that, generally, a greater proportion of Caucasian than of non-Caucasian candidates are advanced (Flyer, 1971a, 1971b, 1971c, 1972). This situation could be a result of one or both of two conditions--differences in several factor scores, or differences in one factor which carries most of the weight. This study concerns an analysis of items in one of the factors, the Technical Knowledge Exam. (Other studies are addressed to the problem of the appropriate weights and weighting procedures for the factors, e.g., Robertson et al., 1972. The exam has been found to represent all or most of the effective weight in some advancement series.)

Purpose

Since it also has been found that Blacks score lower than Whites on the exam factor, this Center was directed by the Chief of Naval Personnel to investigate the feasibility of constructing exams which would reduce the differences in scores. One approach, identified by the Chief of Naval Personnel as of particular interest, was whether exams could be constructed by drawing from a pool of items only those which were similar in difficulty for both Blacks and Whites. The questions specifically addressed are:

1. What the levels of item-difficulty are for Blacks and Whites.

2. What proportion of the total set of items comprises items similar in difficulty for both Blacks and Whites.

3. Whether the similar items are generally the difficult or easy items.

4. To what extent a test constructed of similar items would reduce Black-White score differences.

5. Whether there are any differences in "guessing" behavior between Blacks and Whites.

6. Whether there are any differences between Blacks and Whites in completing the entire test.

7. Whether the same items are relatively easy or difficult for both Blacks and Whites.

METHOD

Data

Item response data from the Technical Knowledge Exams of the Series 61 (August 1972) advancement competitions were provided by the Naval Examining Center (now the Naval Education and Training Program Development Center, NETPDC). The ratings selected for analysis were those in which minority group representation was relatively high. The six ratings selected, in competition to Pay Grades 4 through 7, were:

Aviation Machinist's Mate (Jet Engine Mechanic)(ADJ)
Boatswain's Mate (BM)
Boiler Technician (BT)
Commissaryman (CS)
Hospital Corpsman (HM)
Machinist's Mate (MM)

Thus, data for the 24 separate competing groups were analyzed.

P Value Concepts

Guidelines of good test construction emphasize the importance of the _P_ value, which is the level of difficulty of an item, since it influences the distribution of scores and test reliability.

1. The _P_ value is expressed as the percentage of examinees who answered an item correctly. The possible range of _P_ values is from 00 to 100. Thus, high _P_ values indicate easy items; and low values, difficult items. If the average _P_ value differs much from 50, either higher or lower, the distribution of scores will tend to be skewed. An average of low (i.e., difficult) _P_ values would tend to distribute total scores with a pile-up towards the low end of the scoring range. This, in turn, tends to reduce the spread (i.e., variance) of scores, thereby reducing the ability of the test to differentiate between good and poor examinees.

2. Items with _P_ values in extreme ends of the possible range (e.g., greater than 80 or less than 20) tend to have relatively low interitem correlations which in turn tend to reduce a test's internal consistency (i.e., reliability). Thus, _P_ values should cluster around 50, and ideally, be within the range of 40 to 60 (except as qualified below).

3. Consideration of a _P_ value of 50 as the ideal difficulty level for test items applies only in the case of free-response type tests in which the logical range of probabilities (of the proportion answering the item correctly) varies from 00 to 100. (Answers are not structured in free-response tests, e.g., essay or completion type. Each answer must be composed and written in by the examinee.) In tests of _structured_ items (e.g., multiple choice), the lower limit of the _P_ value

range is attenuated by the probability of guessing (i.e., selecting an alternative answer without regard to its content). For example, for an item with four alternatives, there is a .25 probability of answering the question correctly by guessing. Thus, the ideal *P* value would be greater than 50--perhaps nearer to 62.5, the median of the 25 to 100 range.

4. Distractors, which are the incorrect alternative answers in multiple-choice tests, also affect *P* values.

a. Narrowing is the examinee's mental process of first eliminating from consideration the clearly and obviously incorrect alternatives; then selecting an answer from the remaining plausible alternatives.

b. A mislead is a highly plausible or appealing, but incorrect, alternative.

If there is an appreciable opportunity for narrowing, the probability for guessing is an underestimate. If there are particularly effective misleads, the probability for guessing is an overestimate. Generally, tests tend to have more opportunity for narrowing than for responding to misleads (Nunnally, 1967).

Analysis

Two racial groups, Blacks and Whites, were identified for analysis. The representation of other racial minority groups was too small to permit any analysis which would yield stable (i.e., statistically significant) findings. All statistics were computed separately for Blacks and Whites, and then compared for racial differences.

1. Percentage endorsement of each item alternative was calculated. The percentage endorsement of the correct alternative was extracted as the *P* value.

2. Means and standard deviations of total test scores were computed.

3. Percentage differences between Black and White *P* values were tested for statistical significance (Walker and Lev, 1969, p. 188, Formula 11-13) to identify items of *similar* difficulty. Since mean total test scores of Whites are generally greater than those of Blacks, it was assumed that Black *P* values were less than White *P* values. A one-tail test ($p \leq .05$) was performed on each item. An item was identified as similar if the null hypothesis that the White *P* value was no greater than the Black *P* value could not be rejected. The frequency of similar items was tallied by both total test and by sections.

4. To determine whether similar *P* values were concentrated in any particular segment of the item-difficulty range, the items of each test were rank-ordered on the White *P* values, and then trichotomized into

50-item categories of *easy* (i.e., high *P* values), *medium*, and *difficult* (i.e., low *P* values) items. Frequencies of similar items within each category were tallied.

5. The frequency of Black and White *P* values proximate to the guessing range (i.e., near a proportionate correct endorsement of 25) were also tallied.

6. The following two types of *simulated* tests were constructed for 5 of the 24 Rate groups analyzed:

a. Similar *P* Values (SIM-P), containing only the items in which White *P* values were not significantly greater than Black *P* values (i.e., as determined by the one-tailed test described above), and

b. Upgraded Average *P* Value (UPA-P), containing only items in which the Black *P* value was greater than 25.

To compare the means of the simulated tests with those of the original 150-item tests, it was assumed that SIM-P and UPA-P type tests of 150 items would yield proportionately the same values. Thus, the SIM-P and UPA-P total test mean raw scores were adjusted by multiplying the obtained score by $\frac{\text{N items in original test}}{\text{N items in simulated test}}$. Intercorrelations among the three types of tests--original (all 150 items), SIM-P and UPA-P--and racial group membership were computed. (The dichotomous race variable was coded, 0 = White and 1 = Black, which, of course, yielded point biserial coefficients.)

7. Independent of the above one-tail test which identified the SIM-P type items, a two-tail test was applied, a posteriori, to the items in which the Black *P* value was greater than the White *P* value to identify which items were significantly different (however, items so identified are still included as SIM-P type items).

8. Medians of *P* values and of proportions of blank responses were compared between the first and last 10-item sets of Pay Grades 4 and 6 exams to analyze guessing and test completion behaviors.

9. *Rho* values--the extent to which a total test provides items which are similar in *relative* difficulty for different groups (e.g., the items which are most difficult--with the lowest *P* values--for one group are the same items which are most difficult for other groups)--were calculated by the following procedure:

a. *P* values were first calculated and then rank-ordered separately for each group.

b. In order to provide for ties in ranks, the Pearson, rather than the Spearman, correlation coefficient was applied to the item ranks.

The relationship of the two rank-orders is expressed by the correlation coefficient, Rho. The possible range of Rho values is thus from +1.00 to -1.00, although negative and low positive values are rare. The measure is sometimes employed as an indication of cultural or racial bias [i.e., if there is not a strong positive correlation (e.g., .90 to 1.00), the same items are not relatively easy and difficult for different groups].

RESULTS

The average total test scores of Whites were found to be higher than Blacks. This finding is consistent with other studies in the military services (e.g., Flyer, 1971a). Table 1 presents the means, standard deviations, and sample sizes (which are also population sizes) for three racial groups. All racial groups other than Blacks and Whites are combined in the "Other" group. The only Rate group in which the average Black score is higher than the White score is HMC, and the difference is slight.

Comparison of P Values

Since average scores of Whites are higher than Blacks, it was logical to expect that White *P* values would also be higher than Black *P* values. This is generally the case with the medians of the *P* values presented in Table 2. There are two exceptions--CSC and HMC--in which the differences are slight. The median *P* values of the Blacks are generally in the lower 40's; and of Whites, in the upper 40's.

Identification of Similar P Values (SIM-P)

Table 3 presents results of significance tests performed on 20 selected items of the ADJ2 Exam. Nineteen are SIM-P type items (i.e., not significantly greater than those of the Blacks). The exception is Item 31. Also, Item 33, while it is a SIM-P type item, is one of the very atypical cases in which the *P* value of Blacks is significantly greater than that of Whites.

The proportion of SIM-P type items identified in each of the 24 groups analyzed varied from about one-half to six-sevenths of the total items in each test. In Table 4 the three rate groups with the highest and lowest number of SIM-P items respectively are as shown below.

Rate	*N* of SIM-P Items (highest)	Rate	*N* of SIM-P Items (lowest)
HM1	133	ADJ3	74
HMC	129	MM1	85
BM2	127	BT1	90

The proportions of SIM-P items within *sections* were generally distributed normally about the proportion of SIM-P items in each *total* exam. Tables 5 through 8 present the proportion of such items in each section of the ADJ Exams for Pay Grades 4 through 7 respectively. For example, the proportion of SIM-P items within the ADJ3

TABLE 1

Advancement Exam Sample Sizes, Means And Standard Deviations by Race

Competition to		Race								
Pay Grade	Rate	Black			White			Other[a]		
		N	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD
4	ADJ3	47	52.38	12.60	644	69.96	14.75	21	56.86	12.99
	BM3	83	58.07	9.38	1033	64.15	11.86	10	58.30	12.52
	BT3	33	61.76	13.37	831	73.77	16.68	7	58.14	10.75
	CS3	27	67.59	10.15	447	76.12	11.76	17	79.12	6.95
	HM3	104	68.00	11.17	1429	73.45	15.53	22	73.50	14.00
	MM3	58	62.48	12.26	1259	72.44	16.56	49	59.27	13.86
5	ADJ2	30	58.27	14.39	565	63.55	15.01	74	62.27	14.12
	BM2	74	60.12	11.70	569	63.43	10.56	16	58.44	11.18
	BT2	28	60.11	10.25	511	73.61	16.57	12	57.42	14.65
	CS2	47	64.00	11.41	412	69.01	10.66	97	75.71	13.11
	HM2	111	63.60	9.43	1391	70.27	13.40	35	71.83	9.72
	MM2	30	56.37	13.69	984	74.09	15.95	29	59.62	13.30
6	ADJ1	50	67.78	15.56	400	72.31	15.19	14	63.29	14.14
	BM1	115	66.33	11.18	502	69.66	11.49	17	60.00	11.79
	BT1	79	70.44	13.57	495	80.70	17.18	7	69.43	17.61
	CS1	127	68.27	12.22	661	72.04	11.78	66	78.26	14.59
	HM1	26	68.58	6.87	546	71.32	11.08	32	69.63	12.09
	MM1	62	62.44	11.26	774	75.39	14.04	20	68.50	14.71
7	ADJC	88	66.77	14.23	1014	70.07	14.50	10	66.10	16.89
	BMC	193	63.60	12.42	1103	65.75	10.87	28	62.54	10.11
	BTC	138	77.91	17.61	956	80.57	15.59	20	84.05	15.53
	CSC	165	63.01	14.24	771	65.58	13.92	68	68.41	15.65
	HMC	157	71.24	13.73	1817	70.75	13.02	50	71.28	13.28
	MMC	110	75.35	13.81	1547	78.73	13.63	52	80.64	16.21

[a] Other - All other minority groups except Black combined.

TABLE 2

Range and Median of Black and White _P_ Values

	_P_ Value				
Rate	Black		White		Black minus White
	Range	Median	Range	Median	Median Difference
ADJ3	8.5 - 78.7	34.0	10.1 - 83.2	45.8	-11.8
BM3	8.4 - 89.2	38.7	8.3 - 89.7	43.6	-4.9
BT3	6.1 - 90.9	42.4	7.0 - 94.5	48.6	6.2
CS3	3.7 - 96.3	44.4	2.9 - 96.4	49.1	-4.7
HM3	5.8 - 95.2	45.2	7.4 - 96.4	49.6	-4.4
MM3	6.9 - 82.8	39.7	9.1 - 88.6	48.3	-8.6
ADJ2	10.0 - 73.3	36.7	10.9 - 74.9	42.5	-5.8
BM2	6.8 - 81.1	39.3	5.1 - 83.8	43.2	-3.9
BT2	3.6 - 85.7	39.3	10.8 - 84.9	50.2	-10.9
CS2	2.1 - 91.5	42.6	2.9 - 95.4	45.6	-3.0
HM2	2.7 - 97.3	41.4	4.0 - 96.1	46.3	-4.9
MM2	3.3 - 83.3	36.7	7.3 - 87.5	49.5	-12.8
ADJ1	12.0 - 92.0	44.0	9.0 - 93.0	48.1	-4.1
BM1	8.7 - 90.4	42.6	5.8 - 92.4	45.3	-2.7
BT1	8.9 - 89.9	45.6	9.1 - 95.2	55.3	-9.7
CS1	4.7 - 83.5	45.8	7.7 - 85.6	50.5	-4.7
HM1	0.0 -100.0	42.3	2.6 - 98.0	45.1	-2.8
MM1	4.8 - 93.6	40.3	11.1 - 97.0	48.5	-8.2
ADJC	6.8 - 85.2	44.3	2.2 - 91.2	48.7	-4.4
BMC	6.7 - 92.8	41.5	5.7 - 95.7	43.7	-2.2
BTC	16.7 - 92.0	51.4	22.6 - 96.0	52.4	-1.0
CSC	7.9 - 88.5	43.0	6.5 - 89.8	42.9	+0.1
HMC	3.2 - 96.2	46.6	1.7 - 96.5	46.5	+0.1
MMC	10.9 - 95.5	50.9	11.6 - 94.6	53.0	-2.1

TABLE 3

Significance Test of Black-White P Value Differences
For 20 Selected Items of the ADJ2 Exam

Item No.	P Value			t[a]	W > B[b]	B > W[c]	Similar Item
	Black	White	Black Minus White				
21	36.7	50.4	-13.7	-1.523			✓
22	36.7	48.9	-12.2	-1.347			✓
23	30.0	33.6	-3.6	-.422			✓
24	30.0	44.1	-14.1	-1.632			✓
25	20.0	26.0	-6.0	-.799			✓
26	40.0	30.8	9.2	1.006			✓
27	56.7	48.1	8.6	.918			✓
28	33.3	28.9	4.4	.509			✓
29	40.0	42.1	-2.1	-.231			✓
30	46.7	52.4	-5.7	-.612			✓
31	23.3	47.3	-24.0	-2.990	*		
32	56.7	41.4	15.3	1.643			✓
33	36.7	17.4	19.3	2.161		*	✓
34	36.7	22.8	13.9	1.542			✓
35	63.3	54.3	9.0	.995			✓
36	26.7	25.8	0.9	.100			✓
37	40.0	46.6	-6.6	-.713			✓
38	50.0	52.4	-2.4	-.255			✓
39	63.3	74.9	-11.6	-1.284			✓
40	60.0	63.5	-3.5	-.386			✓

Note. N = 30 Black, N = 565 White.

[a] Walker and Lev, 1969; p. 188, formula 11-13. Negative t indicates Black P Value is the lower.

[b] * - One-tail test, $P \leq .05$.

[c] * - Two-tail test, $P \leq .05$.

TABLE 4

Frequency of Similar and Different
P Values by Race

Rate	Frequency of P Values in Which			
	Black Similar to White		White > Black[a]	Black > White[b]
	N[c]	%	N	N
ADJ3	74	49.33	76	0
BM3	109	72.66	41	5
BT3	107	71.33	43	1
CS3	117	78.52	32	0
HM3	115	77.18	34	1
MM3	95	63.33	55	2
ADJ2	122	81.33	28	3
BM2	126	84.00	24	3
BT2	106	70.66	44	2
CS2	119	79.86	30	5
HM2	109	73.15	40	2
MM2	91	60.66	59	1
ADJ1	121	80.66	29	3
BM1	120	80.00	30	6
BT1	90	60.00	60	0
CS1	110	74.32	38	6
HM1	133	88.66	17	6
MM1	85	57.04	64	7
ADJC	117	79.05	31	10
BMC	119	79.33	31	12
BTC	117	78.00	33	13
CSC	118	80.82	28	11
HMC	129	88.35	17	13
MMC	114	76.00	36	12

[a] One-tail test, $P \leq .05$.

[b] Two-tail test, $P \leq .05$.

[c] Includes the N of items, Black > White, of right column.

TABLE 5

Frequency of *P* Values Which are Similar for Blacks And Whites Within Sections of the ADJ3 Exam

Subtest	Frequency of Items which are			
Section	Total	Similar		Not Similar (White > Black)
	f	*f*	%	*f*
1	18	8	44.4	10
2	13	7	53.9	6
3	22	8	36.4	14
4	14	7	50.0	7
5	17	10	58.8	7
6	13	8	61.5	5
7	14	5	35.7	9
8	12	8	66.7	4
9	15	7	46.7	8
10	12	6	50.0	6

TABLE 6

Frequency of P Values Which are Similar for Blacks And Whites Within Sections of the ADJ2 Exam

Subtest Section	Frequency of Items which are			
	Total	Similar		Not Similar (White > Black)
	f	f	%	f
1	18	15	83.3	3
2	19	18	94.7	1
3	25	16	64.0	9
4	17	13	76.5	4
5	16	14	87.5	2
6	19	14	73.7	5
7	18	17	94.4	1
8	18	15	83.3	3

TABLE 7

Frequency of P Values Which are Similar for Blacks And Whites Within Sections of the ADJ1 Exam

Subtest Section	Frequency of Items which are			
	Total	Similar		Not Similar (White > Black)
	f	f	%	f
1	18	13	72.2	5
2	15	15	100.0	0
3	12	11	91.7	1
4	13	11	84.6	2
5	23	21	91.3	2
6	13	10	76.9	3
7	15	11	73.3	4
8	16	10	62.5	6
9	25	19	76.0	6

TABLE 8

Frequency of _P_ Values Which are Similar for Blacks And Whites Within Sections of the ADJC Exam

Subtest Section	Frequency of Items which are			
	Total	Similar		Not Similar (White > Black)
	_f_	_f_	%	_f_
1	25	23	92.0	2
2	14	9	64.3	5
3	14	14	100.0	0
4	12	10	83.3	2
5	12	9	75.0	3
6	17	13	76.5	4
7	15	9	60.0	6
8	14	12	85.7	2
9	25	18	72.0	7

Exam sections (Table 5) ranges from 35.7 to 66.7 percent around the proportion for the total test of 49.3 percent (Table 4); and in the sections of the ADJ1 Exam (Table 7), from 62.5 to 100.0 percent, around 80.7 percent (Table 4). With one exception there was no consistent concentration of similar or nonsimilar *P* values in any section number or sequence. The one exception was in the case of the Military section (items 126 through 150), in the Pay Grade 7 exams, in which Black *P* values were consistently lower than the White *P* values.

The SIM-P items were concentrated in the category of "most difficult" (i.e., lowest *P* value) items. Tables 9 and 10 present the proportions of SIM-P items in the easy, medium and difficult item categories for the Pay Grades 4 and 5 exams respectively. In Table 9, in five of the six Pay Grade 4 exams (the HM3 Exam is the exception), the highest proportion of SIM-P items is in the *difficult* category (e.g., for ADJ3, 70 percent are categorized as difficult, compared with 60 percent medium and 18 percent easy). In Table 10, all of the six Pay Grade 5 exams indicate a similar concentration.

The frequency of *P* values *in* the guessing range (00-25) and *near* (25-30) is presented in Table 11. The three rate groups with the highest frequencies *in* the guessing range in each racial group are as shown below.

	N Items	
Rate	Black	White
ADJ3	36	20
BM3	38	30
BM2		27
BT2	32	

Eighteen of the 24 exams have at least 10 percent of the items (i.e., ≥ 15 items) in the guessing range for Black *P* values; and 12, for White *P* values.

Item *content* of SIM-P items, in some exams, tended to represent a greater proportion of *applied* or *factual* content, than of theoretical or conceptual content. Although content analysis was not planned as part of the design of the present study, a brief, preliminary analysis was performed on five of the Pay Grade 4 exams to identify conditions which might contribute to generating hypotheses for subsequent studies. Each item was placed in one of the following two broad categories: (1) *Theoretical or Conceptual* - "why" or "because" items concerning the reason for doing a job a certain way, and (2) *Applied or Factual* - actual procedural steps, how something is done, or something necessary to know in order to do the job. For example, Table 12 indicates that 74 percent of the applied items in the BT3 Exam are similar, compared to 63 percent of the theoretical items. In the one exception, the HM3 Exam, 79 percent of the theoretical items are similar, compared to 75 of the applied items.

TABLE 9

Frequency of Similar Black and White *P* Values
Within Three Ranges of Item-Difficulty
For the Pay Grade 4 Exams

Rate	Difficulty	Frequency of Items			
		Similar		Dissimilar	Total
	Range	*f*	%	*f*	*f*
ADJ3	Easy	9	18.0	41	50
	Medium	30	60.0	20	50
	Difficult	35	70.0	15	50
	Total				150
BM3	Easy	34	68.0	16	50
	Medium	33	66.0	17	50
	Difficult	42	84.0	8	50
	Total				150
BT3	Easy	34	68.0	16	50
	Medium	34	68.0	16	50
	Difficult	39	78.0	11	50
	Total				150
CS3	Easy	34	69.4	15	49
	Medium	41	82.0	9	50
	Difficult	42	84.0	8	50
	Total				149
HM3	Easy	43	86.0	7	50
	Medium	33	66.0	17	50
	Difficult	38	77.6	11	49
	Total				149
MM3	Easy	25	50.0	25	50
	Medium	33	66.0	17	50
	Difficult	37	74.0	13	50
	Total				150

TABLE 10

Frequency of Similar Black and White P Values
Within Three Ranges of Item-Difficulty
For the Pay Grade 5 Exams

Rate	Difficulty	Frequency of Items			
		Similar		Dissimilar	Total
	Range	f	%	f	f
ADJ2	Easy	37	74.0	13	50
	Medium	39	78.0	11	50
	Difficult	46	92.0	4	50
	Total				150
BM2	Easy	39	79.6	10	49
	Medium	40	78.4	11	51
	Difficult	47	94.0	3	50
	Total				150
BT2	Easy	33	67.3	16	49
	Medium	36	70.6	15	51
	Difficult	37	74.0	13	50
	Total				150
CS2	Easy	33	66.0	17	50
	Medium	42	84.0	8	50
	Difficult	44	89.8	5	49
	Total				149
HM2	Easy	39	78.0	11	50
	Medium	27	54.0	23	50
	Difficult	43	87.8	6	49
	Total				149
MM2	Easy	23	46.0	27	50
	Medium	26	52.0	24	50
	Difficult	42	84.0	8	50
	Total				150

TABLE 11

Frequency of P Values Proximate to the
Guessing Range of Item-Difficulty

Rate	Frequency of Items within P Value range of			
	00.00-25.00		25.01-30.00	
	Black	White	Black	White
ADJ3	36	20	24	7
BM3	38	30	12	8
BT3	27	6	12	12
CS3	27	17	11	8
HM3	23	18	8	4
MM3	17	12	14	6
ADJ2	29	15	17	9
BM2	31	27	12	13
BT2	32	8	9	5
CS2	21	19	16	9
HM2	22	17	14	9
MM2	30	8	29	8
ADJ1	12	10	15	10
BM1	14	15	16	13
BT1	10	6	10	7
CS1	19	17	9	8
HM1	31	18	8	16
MM1	28	13	15	10
ADJC	17	11	10	12
BMC	16	17	17	13
BTC	4	3	9	3
CSC	18	13	16	12
HMC	11	11	7	3
MMC	11	6	8	7

TABLE 12

Type of Content in Items of Similar Difficulty

Rate	Item Difficulty[a]	Type of Item Content						
		Theoretical or Conceptual		Applied or Factual		Other[b]		Total
		N	%	*N*	%	*N*	%	*N*
	Sim	14	44	58	53	3	38	75
ADJ3	Not	18	56	52	47	5	62	75
	Tot	32	(100)	110	(100)	8	(100)	150
	Sim	6	55	84	69	14	78	104
BM3	Not	5	45	37	31	4	22	46
	Tot	11	(100)	121	(100)	18	(100)	150
	Sim	30	63	75	74	0	--	105
BT3	Not	18	37	27	26	0	--	45
	Tot	48	(100)	102	(100)	0	--	150
	Sim	25	51	56	64	10	77	91
MM3	Not	24	49	32	36	3	23	59
	Tot	49	(100)	88	(100)	13	(100)	150
	Sim	34	79	72	75	8	80	114
HM3	Not	9	21	24	25	2	20	35
	Tot	43	(100)	96	(100)	10	(100)	149

[a] Item Difficulty:

Sim - Items with *P* values similar in difficulty for both Blacks and Whites.

Not - Items *not* similar in difficulty.

[b] Other - Item content includes: Definitions, minor procedural details, indeterminant category, etc.

Comparison of Simulated and Original Tests

The racial effects from the two types of simulated tests, SIM-P and UPA-P, are presented in Tables 13 and 14. From the three Pay Grade 4 and the two Pay Grade 5 exams analyzed, the following observations are made:

1. Shifts in mean score differences from the original to the simulated tests correspond to shifts in median P values. Generally, the Black-White differences in both scores and P values are reduced substantially by the SIM-P tests, and are reduced slightly by the UPA-P tests (with the exception of the ADJ3 and MM3 P value differences which increased slightly on the UPA-P tests). For example, the HM2 P value differences decreased from -4.9 to -0.8 by the SIM-P test, and to -4.8 by the UPA-P test.

2. The correlations in Table 14 present similar results. The racial effect is reduced substantially by the SIM-P tests, and slightly by the UPA-P tests (e.g., for the ADJ3 Exam, from -.290 to -.071, and to -.272 respectively).

End-of-Test Guessing and Completion

Response distributions to alternatives for a sample of 20 items of the MM3 Exam are displayed in Table 15. Some of the distributions in this table are illustrative of certain types of item alternatives, including:

1. Misleads--Items 3, 6, 142, and 145 for Blacks; and 145, 146, 147, and 150 for Whites--in which highly plausible, but incorrect, alternatives were selected by a greater proportion of examinees than the proportion which selected the correct alternative.

2. Narrowing--Alternatives c and d of Item 1--which were apparently identified by nearly all Blacks and Whites as clearly incorrect.

To analyze guessing behavior at the end of the test it was assumed that the actual difficulty levels of items were distributed randomly throughout the test, and a drop in P values at the end of the test is indicative of the examinees' tendency to guess as time runs out, or as mental fatigue mounts. Tables 16 and 17 present the differences in median P values between the first and last 10-item sets for Pay Grades 4 and 6 respectively. For example, the median P values on the ADJ3 Exam drop 1.06 for the Blacks, and 10.63 for the Whites. The Pay Grade 4 exams (Table 16) strongly support the guessing hypothesis, since the P values for both groups decreased on five of the six exams. However, the data from the Pay Grade 6 exams (Table 17) contradict the hypothesis, since the P values increased on four of the six exams.

TABLE 13

Mean Total Score and Median P Values
By Race on Three Types of Tests

Rate Group		Original[a]			SIM-P[b, d]			UPA-P[c, d]		
		Black	White	Difference	Black	White	Difference	Black	White	Difference
ADJ3	$\bar{X}$	52.38	69.96	-17.58	55.16	58.51	-3.35	60.19	77.52	-17.33
	P	34.0	45.8	-11.8	36.2	40.1	-3.9	38.3	51.3	-13.0
HM3	$\bar{X}$	68.00	73.45	-5.45	72.35	74.39	-2.04	76.10	81.34	-5.24
	P	45.2	49.6	-4.4	48.1	50.3	-2.2	49.0	53.1	-4.1
MM3	$\bar{X}$	62.48	72.44	-9.96	66.59	68.99	-2.40	67.18	77.04	-9.86
	P	39.7	48.3	-8.6	41.4	44.6	-3.2	41.4	50.3	-8.9
BM2	$\bar{X}$	60.12	63.43	-3.31	61.03	61.34	-.31	69.27	72.24	-2.97
	P	39.3	43.2	-3.9	41.2	41.5	-0.3	46.0	48.0	-2.0
HM2	$\bar{X}$	63.60	70.27	-6.67	67.83	69.67	-1.84	70.84	76.52	-5.68
	P	41.4	46.3	-4.9	45.0	45.8	-0.8	45.1	49.9	-4.8

[a] Includes the complete set of 150 items.

[b] Includes only items in which the Black P value was not significantly less than the White P value.

[c] Includes only items in which the Black P value was greater than .25.

[d] Mean total scores are simulated by obtained SIM-P or UPA-P score times

$$\frac{N \text{ items in original test}}{N \text{ items in simulated test}}.$$

TABLE 14

Intercorrelations of Three Types
Of Exams and Racial Group

Rate	Type Test[a]	Correlation	
		Orig.	Race[b]
ADJ3	UPA	.979	-.272
	SIM	.838	-.071
	Orig.		-.290
HM3	UPA	.992	-.078
	SIM	.976	-.034
	Orig.		-.089
MM3	UPA	.995	-.116
	SIM	.950	-.032
	Orig.		-.124
BM2	UPA	.978	-.077
	SIM	.967	-.009
	Orig.		-.098
HM2	UPA	.987	-.115
	SIM	.954	-.040
	Orig.		-.132

[a] Type Test

Orig. - Original 150 items
UPA - Contains only P Values of which Black P Values > .25.
SIM - Contains only items for which P Value is similar for Blacks and Whites.

[b] Race coded: 0 = White, 1 = Black.

TABLE 15

Proportionate Endorsement by Race of Alternative Answers
For 20 Selected Items of the MM3 Exam

Item No.	Endorsement Percentage									
	Black (N=58)					White (N=1259)				
	Alternative[a]									
	A	B	C	D	Blank	A	B	C	D	Blank
1	29.31	63.79	5.17	1.72	0.0	28.75	68.55	1.51	1.11	0.08
2	51.72	12.07	27.59	8.62	0.0	50.75	12.63	21.37	14.93	0.32
3	34.48	39.66	10.34	13.79	1.72	57.51	26.61	8.82	6.59	0.48
4	20.69	18.97	50.00	10.34	0.0	28.12	11.44	50.68	9.53	0.24
5	15.52	48.28	18.97	17.24	0.0	11.28	63.78	16.60	8.10	0.24
6	10.34	43.10	29.31	17.24	0.0	3.97	20.33	65.37	10.33	0.0
7	6.90	50.00	10.34	32.76	0.0	2.30	54.65	7.78	35.19	0.08
8	10.34	6.90	3.45	77.59	1.72	5.56	4.45	5.48	84.35	0.16
9	0.0	37.93	60.34	1.72	0.0	3.26	24.62	68.55	3.42	0.16
10	10.34	13.79	53.45	22.41	0.0	7.86	10.01	65.61	16.20	0.32
141	55.17	17.24	22.41	5.17	0.0	58.38	14.54	20.49	6.12	0.48
142	36.21	12.07	39.66	12.07	0.0	41.70	7.94	37.09	12.79	0.48
143	41.38	3.45	6.90	46.55	1.72	33.68	3.65	3.65	58.86	0.16
144	12.07	12.07	51.72	24.14	0.0	6.43	22.48	51.87	18.67	0.56
145	62.07	12.07	15.52	10.34	0.0	45.91	21.92	20.89	10.41	0.87
146	12.07	31.03	8.62	48.28	0.0	9.21	38.13	7.55	44.72	0.40
147	39.66	13.79	15.52	29.31	1.72	32.09	9.29	12.55	45.83	0.24
148	8.62	62.07	18.97	10.34	0.0	5.48	66.64	10.64	16.92	0.32
149	25.86	6.90	17.24	50.00	0.0	10.88	11.36	16.68	60.60	0.48
150	25.86	34.48	27.59	12.07	0.0	33.44	26.85	30.42	8.66	0.64

[a] Underline indicates correct answer, and also the level of difficulty (P value) of the item.

TABLE 16

Differences in P Values Between First and Last 10-Item Sets for Pay Grade 4 Exams

		P Value				
Rate	10-Item	Black		White		Median
	Set[a]	Range	Median	Range	Median	Difference
						(B-W)
	First	17.02-57.45	39.36	20.50-82.61	59.00	-19.64
ADJ3	Last	17.02-61.70	38.30	28.73-83.23	48.37	-10.07
	Diff.		1.06		10.63	-9.57
	First	19.28-89.16	46.39	24.20-89.74	50.00	-3.61
BM3	Last	12.05-60.24	30.72	16.75-68.05	34.65	-3.93
	Diff.		15.67		15.35	.32
	First	6.06-69.70	25.75	6.98-73.41	50.00	-24.25
BT3	Last	30.30-87.88	50.00	41.28-79.42	55.23	-5.23
	Diff.		-24.25		-5.23	-19.02
	First	11.11-88.89	50.00	24.61-85.46	59.17	-9.17
CS3	Last	11.11-74.07	46.29	2.91-88.37	53.58	-7.29
	Diff.		3.71		5.59	-1.88
	First	17.31-86.54	42.31	16.52-89.71	46.68	-4.37
HM3	Last	8.65-59.62	36.54	8.19-65.78	44.33	-7.79
	Diff.		5.77		2.35	3.42
	First	29.31-77.59	50.86	50.68-84.35	64.58	-13.72
MM3	Last	12.07-62.07	43.11	21.92-66.64	46.79	-3.68
	Diff.		7.75		17.79	-10.04

[a] Diff. - Median difference of first minus last set.

TABLE 17

Differences in P Values Between First and Last 10-Item Sets for Pay Grade 6 Exams

Rate	10-Item	P Value				
		Black		White		Median
	Set[a]	Range	Median	Range	Median	Difference (B-W)
ADJ1	First	34.00-68.00	53.00	29.00-68.25	60.63	-7.63
	Last	28.00-92.00	51.00	28.00-93.00	54.25	-3.25
	Diff.		2.00		6.38	-4.38
BM1	First	31.30-71.30	43.91	38.25-80.08	53.88	-9.97
	Last	15.65-84.35	38.69	14.34-88.84	39.74	-1.05
	Diff		5.22		14.14	-8.92
BT1	First	21.52-65.82	44.30	26.46-69.90	53.53	-9.23
	Last	25.32-78.48	45.57	28.48-82.83	56.67	-11.10
	Diff.		-1.27		-3.14	1.87
CS1	First	18.11-74.02	46.06	15.28-67.62	52.87	-6.81
	Last	19.69-82.68	55.11	16.79-85.63	58.09	-2.98
	Diff.		-9.05		-5.22	-3.83
HM1	First	7.69-80.77	36.54	10.99-79.30	35.80	.74
	Last	26.92-84.62	46.15	30.04-89.19	45.88	.27
	Diff.		-9.61		-10.08	.47
MM1	First	14.52-69.35	35.48	26.74-72.87	43.21	-7.73
	Last	29.03-85.48	50.81	31.27-88.37	64.92	-14.11
	Diff.		-15.33		-21.71	6.38

[a] Diff. - Median difference of first minus last set.

The proportions of unanswered items in the first and last 10-item sets are compared in Tables 18 and 19 for Pay Grades 4 and 6 respectively. All median differences are zero or less than one percent. The negligible differences indicate that both Blacks and Whites complete the entire test, even though the tendency to guess appears to increase at the end of the Pay Grade 4 exams.

Relative Item Difficulty

Rho values (as defined on page 5) in Table 20 vary from .551 (ADJ2) to .932 (HM3). Some of the Rho values are fairly low. Taking the square of the Rho correlation coefficient as the proportion of variance accounted for in the relative difficulty of items between Blacks and Whites, the groups are distributed as follows.

Range of		
Rho	Rho^2 (x 100%) as proportion of accountable variance	Rate Groups
.551-.721	30-52%	ADJ2, MM2, BT2, ADJ3
.750-.818	56-67%	MM1, ADJ1, BT3, CS3, BTC
.832-.901	69-81%	MM3, HM1, ADJC, MMC, CS2, BT1, BMC, HM2, CSC
.910-.932	83-87%	CS1, BM1, BM3, BM2, HMC, HM3

Thus, in four of the groups (those in the 30-52 percent range), about one-half or more of the variance in the relative difficulty of items is unaccounted for. In five groups (those in the 56-67 percent range), one-third to one-half of the variance is unaccounted for.

Comparing the Rho values between the original and two simulated tests (Table 21), the Rho values increase on SIM-P tests, and decrease on the UPA-P tests.

TABLE 18

Differences in Proportion of Unanswered (Blank) Items Between First and Last 10-Item Sets for the Pay Grade 4 Exams

Rate	10-Item Set[a]	Blank Endorsement Percentage				
		Black		White		Median
		Range	Median	Range	Median	Difference (B-W)
ADJ3	First	0.00-2.13	0.00	0.00-0.78	0.08	-0.08
	Last	0.00-6.38	0.00	0.00-1.09	0.47	-0.47
	Diff.		0.00		-0.39	0.39
BM3	First	0.00-2.41	0.00	0.10-0.48	0.24	-0.24
	Last	0.00-2.41	0.00	0.19-0.58	0.34	-0.34
	Diff.		0.00		-0.10	0.10
BT3	First	0.00-6.06	0.00	0.00-0.48	0.12	-0.12
	Last	0.00-0.00	0.00	0.12-0.96	0.48	-0.48
	Diff.		0.00		-0.36	0.36
CS3	First	0.00-3.70	0.00	0.00-0.67	0.11	-0.11
	Last	0.00-3.70	0.00	0.00-0.67	0.45	-0.45
	Diff.		0.00		-0.34	0.34
HM3	First	0.00-1.92	0.00	0.00-0.49	0.21	-0.21
	Last	0.00-1.92	0.00	0.70-0.56	0.35	-0.35
	Diff.		0.00		-0.14	0.14
MM3	First	0.00-1.72	0.00	0.00-0.48	0.20	-0.20
	Last	0.00-1.72	0.00	0.16-0.87	0.48	-0.48
	Diff.		0.00		-0.28	0.28

[a] Diff. - Median Difference of first minus last set.

TABLE 19

Differences in Proportion of Unanswered (Blank) Items Between First and Last 10-Item Sets for the Pay Grade 6 Exams

		Blank Endorsement Percentage				
Rate	10-Item	Black		White		Median
	Set[a]	Range	Median	Range	Median	Difference (B-W)
ADJ1	First	0.00-4.00	0.00	0.00-0.50	0.00	0.00
	Last	0.00-4.00	0.00	0.00-1.00	0.25	-0.25
	Diff.		0.00		-0.25	0.25
BM1	First	0.00-1.74	0.00	0.00-0.40	0.10	-0.10
	Last	0.00-2.61	0.87	0.00-0.80	0.20	0.67
	Diff.		-0.87		-0.10	-0.77
BT1	First	0.00-1.27	0.00	0.00-0.20	0.00	0.00
	Last	0.00-1.27	0.00	0.00-0.61	0.30	-0.30
	Diff.		0.00		-0.30	0.30
CS1	First	0.00-0.79	0.00	0.00-0.61	0.15	-0.15
	Last	0.00-3.15	0.79	0.00-1.21	0.30	0.49
	Diff.		-0.79		-0.15	-0.64
HM1	First	0.00-0.00	0.00	0.00-0.37	0.09	-0.09
	Last	0.00-0.00	0.00	0.00-0.37	0.18	-0.18
	Diff.		0.00		-0.09	0.09
MM1	First	0.00-1.61	0.00	0.00-0.39	0.13	-0.13
	Last	0.00-1.61	0.00	0.00-1.03	0.26	-0.26
	Diff.		0.00		-0.13	0.13

[a] Diff. - Median Difference of first minus last set.

TABLE 20

Relative Order of Item-Difficulty (Rho Value) Between Blacks and Whites

Rate	Relative Difficulty		Sample Size of Black Group	
	Rho Value	Rank[a]	N	Rank[a]
ADJ3	.721	4	47	7.5
BM3	.916	21	83	14
BT3	.780	7	79	13
CS3	.805	8	27	2
HM3	.932	24	104	16
MM3	.840	11	58	10
ADJ2	.551	1	30	4.5
BM2	.919	22	74	12
BT2	.714	3	28	3
CS2	.846	13	47	7.5
HM2	.898	17	111	18
MM2	.565	2	30	4.5
ADJ1	.767	6	50	9
BM1	.913	20	115	19
BT1	.876	15	33	6
CS1	.910	19	127	20
HM1	.872	14	26	1
MM1	.750	5	62	11
ADJC	.832	10	88	15
BMC	.884	16	193	23
BTC	.818	9	138	21
CSC	.901	18	165	24
HMC	.927	23	157	22
MMC	.842	12	110	17

[a] Rank number 1 assigned to lowest Rho Value and lowest Black sample size.

TABLE 21

Rho Values and Number of Items
On Three Types of Tests

Rate Group		Original[a]	SIM-P[b]	UPA-P[c]
ADJ3	Rho	.721	.889	.594
	No. Items used	150	74	114
HM3	Rho	.932	.970	.924
	No. Items used	149	115	126
MM3	Rho	.840	.931	.794
	No. Items used	150	95	133
BM2	Rho	.919	.953	.876
	No. Items used	150	126	119
HM2	Rho	.898	.967	.874
	No. Items used	149	109	125

[a] Includes the complete set of 150 items.

[b] Includes only items in which the Black P value was not significantly less than the White P value.

[c] Includes only items in which the Black P value was greater than .25.

DISCUSSION

P Values Similar in Difficulty for Blacks and Whites

A large proportion of the items in each test (i.e., from one-half to six-sevenths) were identified as similar because the White P value was not significantly greater than the Black P value. Nonetheless, the Black P value of many of the SIM-P items is 10 to 14 percentage points less than the White P value (e.g., Items 21 and 22 of Table 3). The result is that mean scores on SIM-P tests are still slightly greater for Whites than for Blacks (e.g., the ADJ3 SIM-P scores in Table 13 are 55.16 for Black and 58.51 for White). Thus, the construction of SIM-P type tests reduces, but does not eliminate, Black-White score differences.

Concentration of Similar Items in the Difficult Range

This finding is important to the consideration of the quality of test development.

The more difficult an item, the greater the tendency is to guess. Guessing introduces measurement error by reducing the reliability of scores, especially scores near the mean-chance level (e.g., on the present 150-item advancement exams, scores near 37.5). Extremely easy items are less detrimental than extremely difficult items, because there is a lesser tendency to guess, and therefore less measurement error (Nunnally, 1967). Thus, developing advancement exams consisting only of SIM-P type items would necessarily eliminate many of the items which make the highest contribution to the quality of the test (i.e., those with P values of 40 or greater).

The reason that Blacks and Whites perform more similarly on low P value (i.e., difficult) items than on medium or high P value items is not because of any particular type of content. More likely, when questions are excessively difficult, Blacks are just as inclined to guess as are Whites. It appears that, even with the distribution of P values on existing (i.e., original) exams, there is an undesirably high incidence of excessively difficult items, since at least 10 percent of the items are in the guessing range (Table 11) in 18 of the 24 exams for Blacks, and in 12 for Whites.

Another effect of a concentration of difficult items is that it produces undesirable conditions (i.e., a pile-up of scores towards the low end of the range, and a corresponding reduction in score variance). Although the theoretically ideal P value may be located somewhere between 50.0 and 62.5, Table 13 indicates that the median P values for Whites are in the upper 40's, and for Blacks, in the lower 40's (for some exams, even in the 30's). (The effect of present P values on item differentiation, D value, i.e., the test's effectiveness in differentiating between good and poor examinees, is being analyzed in a parallel study.)

Effects of SIM-P and UPA-P Tests

It was generally found that, relative to the original tests:

1. SIM-P tests *substantially* (a) reduce Black-White differences in mean scores, median *P* values, and the (correlational) racial effect, and (b) increase the relationship in relative item difficulty. (This increase in *Rho* values was reasonably to be expected. Since SIM-P items are similar, none can be very far removed from the same rank order.) On the other hand,

2. UPA-P tests *slightly* reduce Black-White differences in mean scores, median *P* values, and the racial effect, but decrease the relationship in relative item difficulty. It would thus appear, since the SIM-P tests substantially reduce racial differences, where UPA-P tests do so only slightly (and even increase the differences in relative difficulty, i.e., reduce *Rho* values), that the development of SIM-P tests is indicated. However, it also appears that the advantages of the SIM-P type test are more than offset by the disadvantages discussed above. A heavy concentration of *P* values in the difficult range, characteristic of the SIM-P type test, degrades the quality of the test for both Blacks and Whites.

Increased Scores by UPA-P Tests

The increased scores which result from UPA-P type tests are advantageous for both Blacks and Whites, but particulary for Blacks. Since Blacks score lower on present exams that Whites, the proportion of Blacks who fail (by cut-scores which are applied to the Exam factor in advancement competition) is greater than that of Whites. The substantial increase in Black scores on UPA-P type tests (e.g., for the ADJ3 Exam in Table 13, from 52.38 to 60.19) would have a corresponding decrease in proportion of Blacks who fail. Thus, these Blacks, even though they may have relatively low values on the Exam factor, could continue to compete. These Blacks may have relatively high values on the other competitive factors (e.g., on Performance or Experience).

Apparent Low Strength of the Racial Factor

The relatively small size of the correlation coefficients for the original tests (e.g., ADJ3 *r* = -.290) in Table 14 might suggest a rather negligible racial effect in the tests. However, this apparent small size is largely attributable to the restriction in the maximum possible size of a point-biserial coefficient, even with an actual, perfect relationship, when the two proportions (*p* and *q*) on the dichotomous variable depart from 50-50. For example, the sample size (Table 1) for the ADJ3 racial groups is 47 Black and 644 White (i.e., a *p*-*q* split of 6.8-93.2) which enables a maximum possible point-biserial coefficient of about *r* = ± .47 (Nunnally, 1967:133, Figure 4-5). Thus, the obtained coefficient of -.290 represents a substantial proportion of the maximum possible relationship.

Guessing at the End of Test

The tendency to guess on Pay Grade 4 exams (Table 16), but not on Pay Grade 6 exams (Table 17), is perhaps attributable to differences in maturity and test-taking experience between the candidates of the two pay grades. The Pay Grade 6 candidates have, of course, been previously successful in advancing up through the lower pay grades, probably by minimizing a resort to guessing.

Rho Values Which are Not High

It was observed that the *Rho* values of *relative* item difficulty between Blacks and Whites did not account for a very sizable proportion of the variance in many rate groups (page 27). This finding raises the question of what might account for the remainder of the variance (e.g., racial bias) in the exam. The median *Rho* value is *r* = .844 which, again, using the square of the correlation coefficient times 100 percent as the proportion of variance accounted for, leaves about one-third of the variance unaccounted for. The sample size of the minority (Black) group was investigated as a possible influence by correlating the *rank* (in Table 20) of the *Rho* value with the *rank* of the Black sample size. The coefficient obtained was *r* = .598, the square of which suggests that sample size is one source of much of the otherwise unaccounted for variance.

A useful approach for addressing the question as to whether *Rho* values sizably less than 1.00 are attributable to racial bias is to match (by exam score) the minority Black group (B) with a "minority" White group (MW) sampled from the majority White group (JW), and then to compute separate *Rho* values for the MW and JW groups, as well as the B and JW group. If

$$\underline{r}_{MW\cdot JW} \leq \underline{r}_{B\cdot JW}$$

(i.e., if the relationship between the White matched-to-Black and White majority group is no greater than that between the Black and White majority group) then, whatever the source of the unaccounted for variance, the source is unlikely to be from racial bias. (A separate, follow-on study of this type is in progress.)

Factual Content of SIM-P Type Items

These items were found to comprise, in some exams, more applied or factual than theoretical or conceptual content, which suggests that consideration be given to increase the representation of applied-type items. Although the item-analysis procedures employed in the present study are useful to determine certain psychometric characteristics, first it is necessary to determine the *content* areas *necessary* for a test to cover. In the Enlisted Advancement System, it is essential that each test cover all of the important skill areas for a particular rating. Some ratings probably require a demonstration of a greater ratio of theoretical to factual knowledge (e.g., perhaps for electronic technicians) than other ratings. Although the SIM-P items appear to reflect more factual than conceptual or theoretical material, the questions as to whether the proportion of factual items should be increased, and conceptual matter reduced, and whether the tests thereafter would still cover all necessary content, are beyond the scope of this study.

Job Relevance of Exams

The fundamental question regarding the measuring instrument, or the weighting of a particular factor in a multifactor selection procedure, is in regard to its relevancy (i.e., validity) to the job for which the personnel selection decision is made. In the present case, the selection decision concerns performance in the next higher grade. Thus, the ultimate criterion for the appropriateness of item content must be determined by job-relevant criteria and performance in the next higher grade. Also, the fact that differences between two groups may exist on factor scores is not of itself sufficient grounds for concluding that the measure of the factor is racially biased against the minority group. Such a conclusion must be based on differences in relative item difficulty, or in validity on a job-relevant criterion of the type described above.

FINDINGS

The following findings are presented in the order of the questions addressed in the introduction.

1. Item-difficulty is generally easier for Whites than for Blacks (i.e., the proportion of examinees responding to the correct item alternative, the *P* value, is higher for Whites than Blacks). The median *P* values of Blacks are generally in the low 40's; and of Whites, in the upper 40's.

2. In the 24 exams analyzed, the proportion of items identified as similar in difficulty for both Blacks and Whites varied from about one-half to six-sevenths of the 150 items in each test. Although the similar items appear, in some exams, to reflect more factual than conceptual or theoretical material, the questions as to whether the representation of factual type items should be increased and conceptual matter correspondingly reduced, and whether the tests thereafter would still cover all necessary content, are beyond the scope of this study.

3. The similar items are generally concentrated in the *difficult* range of *P* values (i.e., the lowest *P* values, near the probability of guessing of .25). There appears to be an undesirably high incidence of excessively difficult items, since at least 10 percent of the items are in the guessing range in 18 of the 24 exams for Blacks, and in 12 for Whites. The reason that Blacks and Whites perform more similarly on low *P* value (i.e., difficult) items than on medium or high *P* value items is probably because, when the questions are extremely difficult, Blacks are just as inclined to guess as Whites.

4. Although items were identified as similar because the White *P* value was not *significantly* greater than the Black *P* value, the Black *P* value of many similar-type items is still 10 to 14 percentage points less than the White *P* value. Thus, simulated tests, containing only the items of similar difficulty for both Blacks and Whites, reduced substantially (compared with the present exams), but did not completely eliminate, Black-White differences in test scores and *P* values. However, a heavy concentration of P values in the difficult range, characteristic of the similar-item type test, degrades the quality of the test for both Blacks and Whites. By contrast, simulated tests containing items with no *P* values in the guessing range, would, of course, upgrade the quality of the tests.

5. Both Black and White examinees in the lower pay grades tend to guess at the end of the exams. Since the same tendency was not found in the higher pay grades, the difference is perhaps attributable to the superior test-taking experience and previous successes of the candidates in the upper pay grades.

6. Both Blacks and Whites tend to complete the entire test.

7. On some exams, the same items are not relatively easy or difficult for both Blacks and Whites, as indicated by a low correlation (i.e., Rho value) between the two rank-orders of the Black and White item-difficulty levels (i.e., P values). On the two simulated tests, the Rho values increased on tests of similar item-difficulty, and decreased on the tests of upgraded P values.

NPRDC - TR-76-6

CONCLUSIONS AND RECOMMENDATIONS

1. The development of advancement exams with items similar in difficulty level for both Blacks and Whites <u>cannot</u> be recommended for the following reasons:

a. Since the items of similar difficulty are concentrated in the difficult (i.e., low end of the *P* value) range, it would degrade test quality for both Blacks and Whites.

b. Since items of similar difficulty (other than low *P* values) also appear to reflect applied or factual content, as distinguished from theoretical or conceptual content, it is problematical (but doubtful) whether an exam comprising items largely limited to this type of content would cover all necessary content for an adequate demonstration of technical knowledge for any given rating.

2. It appears that the quality of the tests can be improved, for both Blacks and Whites, by upgrading item-difficulty levels, and particularly, by reducing the incidence of *P* values in the guessing range. (Specific recommendations will be made upon completion of an on-going analysis of item-differentiation.)

3. Although the relative item-difficulty between Blacks and Whites is low in some rate groups, suggesting a possibility of racial bias in those exams, no such conclusion is yet appropriate. (Relative item-difficulty, with matched Black and White groups, is being analyzed in a follow-on study.)

TR-76-6

REFERENCES

Flyer, E. S. Comparison between Negro and Caucasian enlisted personnel on tests used for promotion. Washington, D. C.: Directorate for Manpower Research, Office of the Assistant Secretary for Defense (M&RA), April 1971a (Manpower Research Note 71-2).

Flyer, E. S. Advancement rates of career enlisted personnel by race and years of service. Washington, D. C.: Directorate for Manpower Research, Office of the Assistant Secretary for Defense (M&RA), April 1971b (Manpower Research Note 71-3).

Flyer, E. S. Promotion opportunities of first-term enlisted personnel by race, aptitude, educational level, and military occupation. Washington, D. C.: Directorate for Manpower Research, Office of the Assistant Secretary for Defense (M&RA), April 1971c (Manpower Research Note 71-4).

Flyer, E. S. Pay grade means and distributions for male enlisted personnel by race and years of active service. Washington, D. C.: Directorate for Manpower Research, Office of the Assistant Secretary for Defense (M&RA), April 1972 (Manpower Research Note 72-11).

Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1967.

Robertson, D. W., James, J., & Royle, M. H. Comparison of alternative criteria and weighting methods for the Enlisted Advancement System. San Diego: Navy Personnel Research and Development Center, June 1972, Technical Bulletin STB 72-11.

Walker, H. M., & Lev, J. Elementary statistical methods. (3rd ed.) New York: Holt, Rinehart and Winston, 1969.

DISTRIBUTION LIST

Assistant Secretary of the Navy (Manpower and Reserve Affairs)
Office of Assistant Secretary of Defense (M&RA), Washington, D. C.
Chief of Naval Operations (OP-39) (OP-59) (OP-098T) (OP-098TL)(2)
 (OP-099)(2) (OP-103B) (OP-964) (OP-987F) (OP-987P10)
Chief of Naval Personnel (Pers-1) (Pers-10c) (Pers-2B) (Pers-21)
 (Pers-212) (Pers-2120) (Pers-5) (Pers-523) (Pers-6) (Pers-611)(2)
Chief of Naval Technical Training
Chief of Naval Technical Training (Code 016) (Code N45)
Chief of Naval Education and Training (N-5)
Chief of Naval Education and Training Support
Chief of Naval Material (NMAT 0344) (NMAT 035)
Chief of Naval Research (Code 450)(4) (Code 458)(2)
Commandant of the Marine Corps (Code MP)
Commander in Chief, U. S. Pacific Fleet
Commander in Chief, U. S. Atlantic Fleet
Commander Second Fleet
Commander Third Fleet
Commander Training Command, U. S. Pacific Fleet
Commander Training Command, U. S. Atlantic Fleet (Code N3A)
Commander Navy Recruiting Command
Commander, Naval Electronics Laboratory Center
Commanding Officer, Fleet Combat Direction Systems Training
 Center, Pacific (Code 03A)
Commanding Officer, Naval Education and Training Program
 Development Center
Commanding Officer, Naval Health Research Center
Commanding Officer, Naval Development and Training Center (Code 0120)
Director, Behavioral Sciences Department, Naval Medical Research
 Institute, Bethesda
Superintendent, Naval Academy
Superintendent, Naval Postgraduate School
Superintendent, United States Military Academy
Superintendent, Air Force Academy
Superintendent, Coast Guard Academy
Technical Library, Naval Postgraduate School
Human Goals Office, Naval Education and Training Center, Newport, R. I.
Army Research Institute for Behavioral and Social Sciences
U. S. Army Enlisted Evaluation Center, Fort Benjamin Harrison (2)
Human Resources Development Division, U. S. Army Personnel and
 Administration Combat Developments Activity
Chief of Research and Development, Department of the Army,
 (Attn: Army Personnel Research Office), (Attn: Human Factors and
 Operations Research Division), Washington, D. C.
Commander, Air Force Human Resources Laboratory, Brooks Air Force Base
Headquarters, Air Force (Attn: Science Division, Directorate of
 Science and Technology, DCS/Research and Development), Washington, D. C.

Assistant Director, Life Sciences, Air Force Office of Scientific Research
Personnel Research Division, Air Force Human Resources Laboratory (AFSC), Lackland Air Force Base
Occupational and Manpower Research Division, Air Force Human Resources Laboratory (AFSC), Lackland Air Force Base
Technical Library, Air Force Human Resources Laboratory (AFSC), Lackland Air Force Base
Technical Training Division, Air Force Human Resources Laboratory, Lowry Air Force Base
Defense Race Relations Institute, Patrick Air Force Base
Flying Training Division, Air Force Human Resources Laboratory, Williams Air Force Base
Advanced Systems Division, Air Force Human Resources Laboratory, Wright-Patterson Air Force Base
Headquarters, Coast Guard, (Attn: Chief, Training and Procurement Division), (Attn: Psychological Research Branch, (G-P-1/62), Washington, D. C.
U. S. Coast Guard Institute, Oklahoma City, Oklahoma
Center for Naval Analyses, Arlington, Virginia
Director, Defense Documentation Center, Alexandria (12)
Director, Utilization and Management Manpower Techniques, Washington, D. C.
National Research Council, Washington, D. C.
National Science Foundation, Washington, D. C.

NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER
SAN DIEGO, CALIFORNIA 92152

TECHNICAL LIBRARY
NAVAL POSTGRADUATE SCHOOL
MONTEREY, CA 93940

0212