Study Report 94-02

# Examining the Self-Development Test for Race and Gender Fairness

**Jay M. Silva**
U.S. Army Research Institute

July 1994





**United States Army Research Institute for the Behavioral and Social Sciences**

# U.S. ARMY RESEARCH INSTITUTE
# FOR THE BEHAVIORAL AND SOCIAL SCIENCES

**A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel**

**EDGAR M. JOHNSON**
**Director**

---

Research accomplished under contract
for the Department of the Army

Technical review by

Peter Legree
Fred Mael

**NOTICES**

**DISTRIBUTION:** Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-POX, 5001 Eisenhower Ave., Alexandria, Virginia 22333-5600.

**FINAL DISPOSITION:** This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

**NOTE:** The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

# REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188



| 1. AGENCY USE ONLY (Leave blank) | 2. REPORT DATE | 3. REPORT TYPE AND DATES COVERED |
|---|---|---|
| | 1994, July | Final May 93 - Sep 93 |

**4. TITLE AND SUBTITLE**
Examining the Self-Development Test for Race and Gender Fairness

**5. FUNDING NUMBERS**
65803D
730
1231
H1

**6. AUTHOR(S)**
Silva, Jay M.

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**
U.S. Army Research Institute for the Behavioral and Social Sciences
ATTN: PERI-RG
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

**8. PERFORMING ORGANIZATION REPORT NUMBER**
ARI Study Report 94-02

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**
--

**11. SUPPLEMENTARY NOTES**
--

**12a. DISTRIBUTION/AVAILABILITY STATEMENT**
Approved for public release; distribution is unlimited.

**12b. DISTRIBUTION CODE**
--

**13. ABSTRACT** *(Maximum 200 words)*

The Self-Development Test (SDT) was examined for gender and race fairness. Three SDT versions with the largest male score advantage and three SDT versions with the largest White score advantage were selected for analysis. Potentially biased items were identified and analyzed. Item correlations with target construct, subject matter expert reviews, and the impact of removing all items showing large performance differences between subgroups were considered. The percentage of items showing large differences in subgroup performance ranged from 14 to 61 percent across the six SDT examined. However, few of these items showed a differential relationship with the target construct across subgroups and the subject matter experts could not identify the items that were more difficult for minority subgroups. Scoring the SDT after removing items with statistically significant differences did not generally eliminate the subgroup differences at the test score level. Although no support was found for race or gender bias in the SDT, differential assignments based on gender, along with SDT emphasis on material covered in some duty positions may give a performance advantage to males in some Military Occupational Specialties

(Continued)

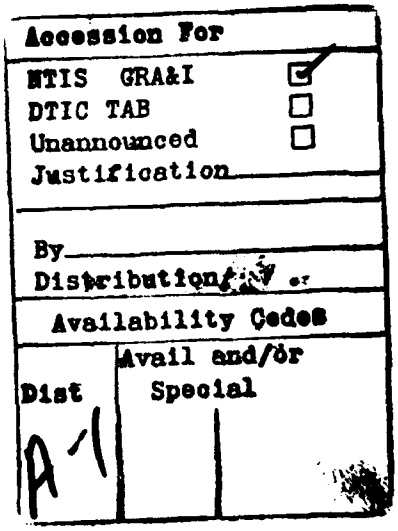
| 14. SUBJECT TERMS | | 15. NUMBER OF PAGES |
|---|---|---|
| Self-development test | Gender fairness | 50 |
| SDT | Test fairness | 16. PRICE CODE |
| Race fairness | | -- |

| 17. SECURITY CLASSIFICATION OF REPORT | 18. SECURITY CLASSIFICATION OF THIS PAGE | 19. SECURITY CLASSIFICATION OF ABSTRACT | 20. LIMITATION OF ABSTRACT |
|---|---|---|---|
| Unclassified | Unclassified | Unclassified | Unlimited |

ARI Study Report 94-02

13. ABSTRACT (Continued)

(MOS). An examination of assignment procedures for MOS showing large gender performance differences in SDT scores is recommended.



Study Report 94-02

# Examining the Self-Development Test for Race and Gender Fairness

**Jay M. Silva**
U.S. Army Research Institute

**Leadership and Organization Change Technical Unit**
**Paul A. Gade, Chief**

**Manpower and Personnel Research Division**
**Zita M. Simutis, Director**

U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

Office, Deputy Chief of Staff for Personnel
Department of the Army

**June 1994**

**Army Project Number**
**2Q162785A791**

**Manpower, Personnel and Training**

## FOREWORD

The Self-Develcpment Test was designed and implemented to help noncommissioned officers evaluate self-development progress in job-related areas, including job knowledge, leadership, and training management. In the future it may also be used as a component in systems that are used for promotion and school selection decisions. Given the potential impact of this test score on career-forming decisions, the Army wants to ensure that the test is gender and race fair.

The research reviewed in this report establishes the fairness of the Self-Development Test using analytical methodology and subject matter experts. The target minority groups were well represented within the subject matter expert groups that examined the test items for bias. The results of these extensive analyses did not show the Self-Development Test to be gender or race biased.

This clears the way for the Self-Development Test to be used in operational promotion and school selection decisions. In the future, operational data can be used to explore other issues, such as the predictive capacity of the Self-Development Test for promotion and school selection decisions.

EDGAR M. JOHNSON
Director

## ACKNOWLEDGMENTS

The assistance of the staff at the Individual Training Evaluation Directorate at the U.S. Army Training Support Center at Fort Eustis is greatly appreciated. Their help in providing Self-Development Test data and assisting in directing the analyses and in collecting subject matter expert data was critical, given the short time available to accomplish this task. I especially thank Brian Davis for his helpful insights in this process.

EXAMINING THE SELF-DEVELOPMENT TEST FOR RACE AND GENDER FAIRNESS

## EXECUTIVE SUMMARY


Requirement:

The FY 94 Self-Development Test (SDT) is nearing implementation as a component of noncommissioned officer (NCO) promotions and school selection under the Enlisted Personnel Management System (EPMS). Concerns related to the fairness of the SDT arose when preliminary analyses showed Blacks and women scored lower than Whites and men, respectively, on many FY 92 SDT. The purpose of this study was to evaluate the fairness of the SDT as it would be used in a revised EPMS.

Procedure:

Although SDT for almost 600 Military Occupational Specialties (MOS)/skill level combinations were fielded in FY 92, the limited time frame for this study allowed the author to focus on only a few of these. The three SDT with the largest stable (i.e., a sample size of at least 100 in each subgroup) score differences between Blacks and Whites were chosen. Likewise, three SDT with the largest stable score differences between males and females were chosen. Items within each of the six chosen SDT were identified as potentially biased via the differential rate of correct responses across subgroups (i.e., Blacks compared to Whites and women compared to men). These items were analyzed with respect to their relation to the construct being measured and the impact of removing these items was assessed. Finally, subject matter experts (SME) were asked to identify the potentially biased items.

Findings:

The results indicated that the worst Black-White differences were substantial. While an average of 44% of Whites scored in the upper third of the SDT score distribution, only 10% of Blacks did so on average across the three SDT with the largest Black-White differences. In standard deviation units, the mean difference between Blacks and Whites across these three SDT versions was 0.94 standard deviations in favor of Whites. The difference for women versus men was not as large. Thirty-six percent of the men scored in the upper third, while only 18% of women did so on average across the three SDT with the largest male-female differences. In standard deviation units, the mean difference between females and males across these three SDT versions was 0.39 standard deviations in favor of males. Across the six SDT the percentage of items identified as potentially biased ranged from 14 to 61 percent. These items showed low relationships with the intended construct but few showed differential relationship with the intended construct across subgroups. Those that did were examined but no basis could be ascribed for the outcome. The SMEs were not able to distinguish the potentially biased items from those items showing no subgroup performance difference on five of the six SDT. The SMEs

for SDT 63H(2) were slightly more capable of choosing the potentially biased items. Whereas SMEs were expected to correctly choose 5 of 10 items simply by chance, they were able to choose 6 of the items. Examination of the items consistently chosen by SDT 63H(2) SMEs did not identify any substantial problems. On one item, cultural differences may have given White examinees an advantage if neither group had studied the material on which the item was based. Seven of ten MOS SMEs for SDT 88M(2) and four of ten MOS SMEs for SDT 71M(2) commented that females were less likely to be assigned to duty positions that require knowledge of content covered in the SDT's MOS Knowledge section. However, these SMEs could not identify the "potentially biased" items any better than chance.

Utilization of Findings:

Item analyses did not support the position that the six SDTs examined were race or gender biased. Differential assignments based on subgroup membership, along with SDT emphasis on material covered in some duty positions may, however, give a performance advantage to males on some SDT. To determine the value of using SDT scores in personnel decisions, the predictive validity of the SDT must be established. Previous analyses with the Skill Qualification Test (SQT) revealed that it was strongly predictive of future performance (Arabian & Mason, 1986). It is, on this basis, likely that the SDT, being a similar test, would also be predictive of performance in future Army roles. It is with this in mind that the SDT is recommended for use in the EPMS on a trial basis in order to evaluate subgroup performance under operational conditions and to quantify the predictive value of the SDT. As a condition to this recommendation, MOS whose SDT show significant subgroup score differences should have their assignment procedures evaluated. Assignment procedures that disproportionately affect the distribution of subgroup members across duty positions should be changed. Concurrently, the Army should assess the impact that any gender or race differential assignment may have on other personnel decision factors such as civilian and military education, awards received, and the NCO Evaluation Report.

EXAMINING THE SELF-DEVELOPMENT TEST FOR RACE AND GENDER FAIRNESS

CONTENTS

CONTENTS (Continued)

CONTENTS (Continued)

EXAMINING THE SELF-DEVELOPMENT TEST FOR RACE AND GENDER FAIRNESS

## Introduction

The Self-Development Test (SDT) is currently being used to aid soldiers in evaluating self development progress. Performance on the test, which is divided into three sections (Leadership, Training Management, and MOS Knowledge), can be used by a soldier to focus future development and training in areas showing knowledge deficiencies. In addition to its current usage, the Army is planning to use the FY 94 SDT score as an additional factor in NCO promotion and school selection decisions.

Preliminary analysis of FY 92 SDT scores revealed that Blacks and women, on average, scored lower than men and Whites, respectively, on this instrument. If the SDT was the only factor in NCO promotion and school selection, these differences would guarantee that women and Blacks, proportionate to their NCO representation, would be less frequently promoted and selected for schooling. This highlights the need for an assessment of the SDT's fairness to Blacks and women.

The planned usage of SDT is for NCO promotion and school selection decisions (i.e., linkage to the Enlisted Personnel Management System, EPMS). Specifically, for decentralized promotions (e.g., used for promoting Sergeants to Staff Sergeants) the candidate's SDT score would be provided to the board and it would be used with other information to evaluate the candidate under the "Whole Person" concept. The board allots 200 of the 800 points that a candidate can earn. Other components that are scored separately and fall into the remaining 600 points include courses taken and awards/medals. For centralized promotions (e.g., used for promoting Staff Sergeants and Sergeants First Class) there is no formal point system. The SDT score is considered along with other information, such as the candidate's photograph, assignment history, courses taken, awards/medals, and their NCO Evaluation Report. A promotion merit score is derived from all the information for the candidate under the "Whole Person" concept.

The SDT's potential contribution to a candidate's evaluation may be large or small under such a promotion system. It depends on whether the board thinks the SDT score is a good basis for the promotion decision. If it thinks it is a good basis, it will weigh it highly, otherwise the board may give it a low weight or ignore it. If it gives the SDT a high weight, it is imperative that the SDT be fair to Blacks and women or it may serve to block the advancement of these groups. The remainder of this paper will evaluate the fairness of the SDT for Blacks and women as it would be used in promotion and school selection decisions.

## Defining Fairness

There are three interrelated types of fairness that should be addressed. The first type is content fairness. The issue with content fairness is whether members of each subgroup have equal access and experience with the material covered by the instrument. For example, are men and women given equal access to preparation materials and experience from which the instrument

content is sampled? With respect to the SDT, it is likely that there are no systematic differences in access to preparation materials, but if women and men are given different assignments within an MOS based on their gender, the instrument may not be content fair and, as a result, would unfairly disadvantage the gender whose work assignment employs knowledge and experience that is less emphasized on the instrument.

A second type of fairness is construct fairness. With construct fairness the issue is whether each test item measures the same construct in each subgroup. For example, suppose an instrument is developed to measure the construct of "Brake System Knowledge," but the items are framed at a reading level more difficult than required by their jobs. If Blacks, for example, read at a lower reading level, then even if their knowledge of the "Brake System" is equal to that of Whites, Blacks will perform lower on this instrument by virtue of the reading level of the instrument. The test will not be construct fair because the difference between White-Black performance will be the result of reading level and not "Brake System Knowledge."

A third type of fairness is predictive fairness. In general, when a score on an instrument is a component in a personnel decision such as promotion, then the score on the instrument should be related to performance on some measure of success on the promoted-to job. This is necessary to establish the validity of a selection instrument and hence its appropriateness for the personnel decision. For example, those promoted because they scored higher on the selection instrument should perform better on the new job than those not promoted. In addition, for predictive fairness to be high, the difference in subgroup performance on the selection instrument should mirror the candidate's future performance on the job to which he or she is promoted.

It is possible that the instrument is unfair in all three respects even when there is no difference in subgroup performance on the instrument being examined. For example, with respect to content fairness, women might perform the same on the instrument even when they are given differential assignments based on their gender. This could occur if women studied and practiced the material covered on the instrument on their own time. It is unfair because the women had to work harder to achieve the same score. With respect to construct fairness, Blacks may have a better knowledge of "Brake Systems" and could have demonstrated it on an instrument with an easier reading level. It is unfair because it focused on reading level rather than "Brake Systems." Finally, with respect to predictive fairness, Blacks with the same scores as Whites may perform better at the new job level. The instrument is unfair because the same score on the selection instrument does not have the same implication for performance at the new job level.

Likewise, an instrument may manifest subgroup differences and be fair. In the case of content fairness, males and females could have equal access to the same knowledge and experience and still perform differently on the same test. With respect to construct fairness, the instrument could measure the intended construct but yield subgroup differences as a result of differences in the level of the construct across subgroups. Finally, with respect to predictive fairness, subgroup differences on the selection instrument could be mirrored in job performance differences.

The issue is not whether there are subgroup performance differences on the SDT. The issue is whether the SDT is fair. Does the SDT sample accurately from the job content of men and women and Blacks and Whites? Does the SDT accurately measure the intended constructs? And, if used for predicting future performance, does the same SDT score indicate the same future performance for each subgroup? Unfortunately, the lack of resources, time, and data does not allow a comprehensive examination of SDT fairness. The present research will primarily address construct fairness, and to some extent content fairness through subject matter expert reviews of items where members of a minority subgroup are more likely to choose an incorrect response. This review is intended to identify any glaring unfairness with the SDT and is not comprehensive.

## Magnitude of SDT Score Difference by Race and Gender

A concern over fairness to subgroups arises when subgroup score differences are found on an instrument that could be used to make personnel decisions. That was the case with the SDT. The initial findings presented below were the impetus for the present research effort.

Preliminary examination of SDT score differences by race and gender were conducted on nine Military Occupational Specialties (MOS) chosen in Project A (Campbell, 1990) to be representative of Army MOS. SDT score analysis of these MOS revealed a small gender difference in favor of men (i.e., 0.12 SD units) and a larger race difference in favor of Whites (i.e., 0.45 SD units in favor of Whites compared to Blacks).

The preliminary Black-White score difference on the SDT was smaller than differences found in the literature for job knowledge tests (Campbell, Crooks, Mahoney, & Rock, 1973; Ford, Kraiger, and Schechtman, 1986; Kraiger and Ford, 1985). For example, in a meta-analysis of 16 studies using job knowledge tests Ford et al. (1986) found performance on these tests to be, on average, 0.67 standard deviation units in favor of Whites. The difference found on the nine Project A MOS was only 0.45 standard deviation units, or 0.22 standard deviation units less.

Although the magnitude of the Black-White difference is well documented, the literature is less capable of informing on why these differences exist. Most suggest that these differences are environmental. Bentz (1988), for example, proposes that improved nutrition, housing, and medical care, a nurturing social environment free of fear and violence, and enriched educational opportunities would go far in eliminating racial differences in performance. Others such as Jensen (1980) have also suggested that Black-White aptitude differences have a genetic component. This nature-nurture debate will certainly continue for as long as these subgroup differences exist.

The preliminary gender difference found was substantially smaller than the Black-White difference (i.e., nearly four times smaller). A perusal of the literature did not find any studies describing the magnitude of gender score differences on SDT-like tests (i.e., job knowledge tests). Research exists, however, on differences between men and women on verbal and mathematics performance. Hyde & Linn (1988) summarized 165 studies which compared men and women on verbal ability. The average difference between men and women averaged 0.11 of a standard deviation in favor of women. The direction of this difference held across all age groups studied,

which ranged from pre-kindergarten to post-college age. The range of the difference across the age groups was from 0.06 (for 6-10 year olds) to 0.20 (for those 26 and older).

Hyde, Fennema, and Lamon (1990) studied gender differences on mathematics ability tests. They summarized 100 studies examining this issue, and reported an average performance difference of 0.05 of a standard deviation in favor of men across all age groups. However, this average does not capture the essence of the true difference. Their results indicate that women actually perform better than men until age 15. Starting at age 15 men perform better than women. Men between the ages of 15 and 18 perform 0.29 of a standard deviation better than women of the same age. For those aged between 19 and 25 this difference increases to 0.41 of a standard deviation advantage for men. And for those older than 25 the male advantage increases to 0.59 of a standard deviation.

Unfortunately, because the data is cross-sectional and not longitudinal, it is not clear whether the increasing male-female difference as a function of age is a result of age (i.e., women lose mathematical capabilities faster than men) or the result of the educational system in place at the time when older women were schooled (i.e., older women were less likely to be encouraged to enroll in mathematics courses). Some recent data, however, suggests that the latter explanation is more plausible. Kimball (1989) reports that starting in high school women are less likely to enroll in mathematics courses. This effect, magnified for older individuals would, at least in part, account for the lower performance by females in the older age groups.

# Method

## Data Source

The SDT data examined were from the 1992 data collection. This was the first year that the SDT was administered. In one respect these data may not reflect performance on the SDT in subsequent years because at that time the SDT was not being used to make personnel decisions. Since their scores would have no impact, SDT examinees were not motivated to prepare for this examination. This view is supported by survey data collected at testing time. Data on over 80,000 SDT examinees indicated that over three-fourths of the examinees studied less than 10 hours. Of these, nearly one third did not study at all.

In addition, some of the materials from which test items were drawn were relatively new to the non-commissioned officers (NCO). For example, the materials to be used to prepare for the "Training Management" portion of the test had only recently been made available. This is reflected in the mean "Training Management" score: The average NCO answered just over half of these items correctly. Scores on this and other sections of the test should improve as NCOs obtain greater familiarity with the preparation materials and as their motivation to perform well on the SDT improves after the SDT becomes an integral part of personnel decision systems.

The lack of preparation prior to testing on the SDT will differentially affect average subgroup scores to the extent that different subgroups may have different average aptitude or be

predisposed to the correct answer due to their previous experience, cultural differences in interests, values, and beliefs, or differential duty assignments within the MOS. With increased preparation before SDT testing, group differences in mean SDT scores should be reduced provided that access to preparation materials is not related to subgroup membership.

## Identifying SDT Versions to Study

The SDT is a paper-and-pencil job knowledge test covering three content areas: Leadership (20 items), Training Management (20 items), and MOS Knowledge (50 or more items). However, it is not a single test given to all MOS at all ranks. Rather it is a multitude of tests. Each combination of MOS and rank tested on the SDT uses its own version of the instrument to address differences in the MOS Knowledge requirement. The portions of the SDT covering Leadership and Training Management are shared across MOS (i.e., cover the same content) within a rank and differ across some SDT versions only for maintaining test security.

The SDT content originates from materials that NCOs are expected to study when preparing for their duty position. The SDT is currently based on three leadership manuals, one training management manual, and the Soldier's Manual. SDT items originate from these three sources for the Leadership, Training Management, and MOS Knowledge sections, respectively. The SDT was designed to match the emphasis given to the various topics in these manuals. However, if a duty position within an MOS is not well represented in the Soldier's Manual, then the SDT may also not accurately represent that duty position.

An SDT version for a specific rank is labeled the SDT's skill level. Even within an MOS for a specific skill level (i.e., rank) there may be various versions of the SDT if different equipment is used or different tasks are performed within an MOS at that skill level. This is labeled the SDT's track. Only three or four MOS have different SDT tracks. In FY 92 there were 597 versions of the SDT. This number will grow to about 650 in future years. Given this large number of SDT versions, it was impossible, in the time frame available, to examine all versions of the SDT. Instead, it was decided that the three SDT versions with the largest stable (i.e., a sample size of at least 100 in each subgroup) score differences between Blacks and Whites would be chosen. Likewise, three SDT versions with the largest stable score differences between males and females would be chosen.

Because so many versions of the SDT existed it was decided to initially use only the total test score rather than the three section scores available. This strategy was unlikely to have had much impact on the choice of SDT versions to study since two of three sections targeted identical content areas across SDT versions within a skill level. The next issue was what SDT score statistic would be used. The mean score for each subgroup could be compared across subgroups. A second possibility was to compare the proportion of each subgroup performing in the upper end of the SDT score distribution. The latter option is more reasonable given that the proposed usage is for personnel decisions such as promotion. For example, if a third of those who took the SDT were promoted, the SDT was fair, and women and Blacks performed equally on the constructs the SDT purported to measure, then it would be fair that of those promoted, Blacks and women would be represented to the same extent as their representation in the Army. It is not practically

important how the other two thirds of the SDT examinees performed. Comparing means, however, would include the other two thirds of the SDT examinees. A focus on the upper end of the SDT score distribution was thus chosen.

The final issue was where to set the SDT score cutoff. The cutoff should be set at the same point as the rate involved in the personnel decision. For example, if a third of the SDT examinees will be promoted, then the SDT cutoff should be set to include the highest scoring third of the examinees. A problem exists, in that, the rate involved in the personnel decision will vary by the exact nature of the decision (e.g., promotion or school selection), by MOS needs, and by rank of the promotion (i.e., higher rank promotions may show lower promotion rates). Robinson & Prevette (1992) reported that in 1990 19% of Staff Sergeant candidates were promoted while only 14% of Sergeant First Class candidates were promoted. The cutoff was set at the upper third because it was expected that the promotion rate would be higher for candidates at lower ranks.

## Quantifying Magnitude of SDT Score Difference by Race and Gender in SDT Versions with Largest Differences

The percentage of each subgroup that scored in the upper third and upper half of the SDT score distribution in the six SDT versions identified as having the largest differences was computed. In addition, the nature of the subgroup distribution differences were examined with respect to whether the differences were due solely to negative shifts in the minority subgroup score distribution or also a result of negative skew within the minority subgroup distribution. Finally, means for each subgroup for the six SDT versions and subgroup differences in standard deviation units were computed to allow future comparison to the literature which uses this metric.

## Using Subgroup Performance on Items to Identify Potentially Biased Items

The probability of correctly answering each SDT item in each SDT version for each subgroup was computed. Those items having a statistically significant lower probability of being answered correctly by a minority subgroup were identified. A statistical significance level of 0.01 was chosen because of the large number of comparisons required. Using this statistical significance level reduced the likelihood of identifying items that showed a difference across subgroups solely as a function of sampling fluctuation.

Adjusting SDT Scores to Not Include Potentially Biased Items and Comparing Correlations of Potentially Biased Items with "Bias Adjusted Scores" Across Subgroups. If the items identified were actually biased, then by computing the total score from the remaining items it was possible to construct a score which presumably was construct related and not biased. Correlating this "bias adjusted construct" score with each of the potentially biased items provided insight into whether the identified items may be measuring something other than the intended construct.

Examining the Impact of Removing All Potentially Biased Items From the SDT. The items identified as potentially biased, by definition, showed the greatest differential between subgroups in their probability of being answered correctly (i.e., Blacks and women were more likely to answer these items incorrectly). Other items also demonstrated smaller subgroup differences in their probability of being answered correctly. One important issue is whether these smaller subgroup differences will cumulate to preserve a large portion of the SDT score differential between the subgroups.

## Using Subject-Matter-Experts to Identify Potentially Biased Items

Approach. Identified the 10 items exhibiting the largest subgroup differences in probability of being answered correctly (i.e., less likely to be answered correctly by Blacks or women). In addition, identified 10 items showing nearly no subgroup differences in probability of being answered correctly (i.e., no more than .03 difference in subgroup probabilities of answering an item correctly). In order to sample items from each SDT section the following procedure was used. For each group of 10 items within each SDT version studied, to the extent possible, three items were chosen from the Leadership portion of the test, three items from the Training Management section, and four items were chosen from the MOS Knowledge section. Sometimes the lack of potentially biased items within the Leadership and Training Management sections forced the inclusion of more than four MOS Knowledge items. In addition, within this constraint the items showing the largest subgroup differences were chosen first in a top-down fashion. Finally, items across the two sets of items were also matched for average difficulty (i.e., computed as the proportion of those who answered the item correctly out of all those who responded to the item).

A test form composed of the 20 items (i.e., 10 potentially biased and 10 showing no difference in subgroup probability of answering the item correctly) was constructed for each test version. These test versions were distributed to Subject Matter Experts (SME). SMEs were either enlisted MOS incumbents, usually at the Sergeant rank or higher, or testing experts. The MOS-specific SMEs were tasked only with test versions in their own MOS, while testing expert SMEs were tasked with three or six test versions. A goal was set to obtain 10 MOS-specific SMEs distributed in equal number across subgroups for which that SDT version was identified as being potentially biased. In all cases this goal was accomplished although failure to perform the task as requested reduced usable responses to less than 10 SMEs for some MOS. Five testing expert SMEs were found who agreed to perform the task and returned usable responses.

The task for the SMEs was to identify 10 items that they believed would be biased for one subgroup. They were to identify the item and the subgroup they believed would perform worse as a result of the bias. The nature of potential biases was not specified to the SMEs in order not to lead them. Expecting that this would be a difficult task, they were told to guess if they, after careful consideration, had no basis for making their choices. In addition, they were asked to make notes next to each question they selected as to why they believed the item was biased or unfair to a particular subgroup. At the end of this task they were asked to answer a few questions and make four ratings as follows: 1) the difficulty of the task, 2) their confidence in their choices, 3)

the extent that item wording portrayed the minority subgroup worse, 4) the amount of balance in positive roles given to minority and majority subgroups. In addition, they were asked to identify that were offensive to members of the minority subgroup.

Examining Ability to Identify Potentially Biased Items. It was expected that if SMEs were not able to key in on relevant dimensions, if such dimensions existed, then the number of potentially biased items they would identify would be at a chance level. In this case the number of items one would expect SMEs to identify correctly purely through guessing or by chance was 5. In addition to comparing the mean number of items correctly identified across all SMEs within SDT version, comparisons were also made within each gender and race within SDT version.

Examining Consistency of Items Chosen. Even in cases where no overall ability exists to identify the majority of the potentially biased items, it is possible that individual items may be consistently identified across SMEs . In such cases it is important to examine the rationale given by the SMEs in choosing these items. Similar sound rationales would target items which merit closer scrutiny.

Examining Rationales for Choosing Items. Comments were examined for general trends in rationales for choosing items as being potentially biased. General comments were also examined and valuable information was culled.

Examining Responses to Other Questions. The confidence of the SMEs in their choices, their ratings of how the minority subgroup was portrayed relative to the majority subgroup, and their ratings of the balance in positive roles for the minority subgroup versus the majority subgroup were examined.

## Results

### Magnitude of SDT Score Difference by Race and Gender in SDT Versions with Largest Differences

Table 1 shows the SDT versions identified as having the largest male-female and Black-White score differences when the cutoff was set at either the upper third or upper half. The Black-White difference ranged from 31-39% when the cutoff used was the upper third and from 37-44% when the cutoff used was the upper half. What this means is that when only those who scored in either the upper third or upper half were selected, Blacks were represented anywhere from 31-44 percentage points less. For example, a percentage difference of 31 points in the SDT for 12C(2) indicates that 11% of Blacks scored in the upper third while 42% of Whites scored in the upper third.

When the upper third cutoff was used to examine race score differences, the SDT versions with the largest differences were 63H(2) (i.e., SDT for MOS 63H Skill Level 2), 63B(2), and 12C(2). When the cutoff was moved to the upper half, the SDT versions with the largest

**Table 1**

**Percent of Subgroup Members Scoring in the Upper Third and Upper Half of the SDT Score Distribution in the Three SDT Versions Exhibiting the Largest Subgroup Differences**

| | SDT Cutoff | | | | | |
|---|---|---|---|---|---|---|
| | Upper Third | | | Upper Half | | |
| | SDT Version | | | SDT Version | | |
| Subgroup | 63H(2) | 63B(2) | 12C(2) | 63H(2) | 63B(2) | 96B(3) |
| Black | 6.08 | 12.67 | 10.77 | 20.55 | 20.24 | 22.38 |
| White | 45.21 | 44.32 | 41.76 | 64.84 | 59.45 | 59.17 |
| Difference | | | | | | |
| | SDT Version | | | SDT Version | | |
| | 71M(2) | 91C(4) | 88M(2) | 71M(2) | 88M(2) | 91C(4) |
| Male | 40.57 | 33.97 | 32.42 | 50.29 | 48.27 | 50.00 |
| Female | 17.65 | 18.42 | 17.14 | 28.43 | 30.11 | 33.55 |
| Difference | | | | | | |

Notes. A cell entry in a subgroup row represents the percent of the subgroup scoring above the cutoff on the SDT. A cell entry in a "difference" row indicates the difference in subgroup percentages scoring above the cutoff. A positive "difference" indicates that the majority group (i.e., males or Whites) was, proportionate to subgroup size, more likely to score above the cutoff. See Table 2 for a description of the MOS associated with each SDT version.

differences were the same except that version 12C(2) had been replaced by version 96B(3). Table 2 identifies the MOS associated with each SDT version.

Figures 1, 2, and 3 illustrate the differences in score distributions by race for SDT versions 12C(2), 63B(2), and 63H(2). The Black score distributions for these SDT versions, compared to those for Whites, were clearly shifted to the lower scores. This shift clearly accounts for the lower proportion of Blacks, as compared to Whites, scoring in the top third of the SDT score distributions.

The largest male-female differences were substantially less compared to the Black-White differences. They ranged from 15-23% when the cutoff was set at the upper third and from 16-22% when the cutoff was set at the upper half. The same SDT versions were identified as having the largest gender differences at both the upper third and upper half cutoffs. The SDT versions with the largest gender differences were 71M(2), 91C(4), and 88M(2).

Figures 4, 5, and 6 illustrate the differences in score distributions by gender for these three SDT versions. The female score distributions for these SDT versions, compared those for males, were somewhat shifted to the lower scores. This shift clearly accounts for the lower proportion of females, as compared to males, scoring in the top third of the SDT score distributions.

Table 3 shows the sample sizes, means, and standard deviations for each of the six SDT versions with the largest Black-White or male-female differences in the upper third of the SDT score distribution. The mean SDT score difference in the SDT versions with the largest Black-White differences ranged from 0.76 to 1.25 standard deviation units in favor of Whites. The male-female difference in the respective SDT versions was substantially less, ranging from 0.36 to 0.40 of a standard deviation unit in favor of males.

AFQT Adjustments. Examination of Armed Forces Qualification Test (AFQT) scores for SDT examinees late in the analysis process enabled some revealing analyses. The AFQT is a general trainability measure that, via its emphasis on math and verbal ability, predicts the learning capability of incoming recruits. There were substantial average AFQT differences between Blacks and Whites, and males and females in this study. Adjustment of SDT scores using the AFQT essentially asks the question "how do two individuals who scored the same on the AFQT comparatively score on the SDT?" Analysis revealed that for those who were tested on these six SDT versions, the average entry AFQT score of Blacks was 0.93 of a standard deviation lower than Whites, and the average score of females was 0.48 of a standard deviation higher than males. This suggested that perhaps Black-White differences on the SDT could be explained by original entry performance differences on the AFQT. AFQT adjustments for SDT male-female differences would only increase the average male-female SDT score difference since females, on average, scored higher on the AFQT. This latter analysis was needless and was not conducted.

Analyses which adjusted each examinee's score based on their AFQT score revealed that SDT race differences could not be entirely explained by AFQT differences. The difference in Black-White representation in the top-third of the SDT score distribution was only reduced after adjusting for AFQT score. The reduction ranged from 4 to 9 percentage points across the three

**Table 2**

**MOS Associated with Each SDT Version Referenced**

| SDT Version | Associated MOS |
|---|---|
| 12C(2) | Bridge Crewmember |
| 63B(2) | Light Wheel Vehicle Mechanic |
| 63H(2) | Track Vehicle Repairer |
| 71M(2) | Chaplain Assistant |
| 88M(2) | Motor Transport Operator |
| 91C(4) | Practical Nurse |
| 96B(3) | Intelligence Analyst |

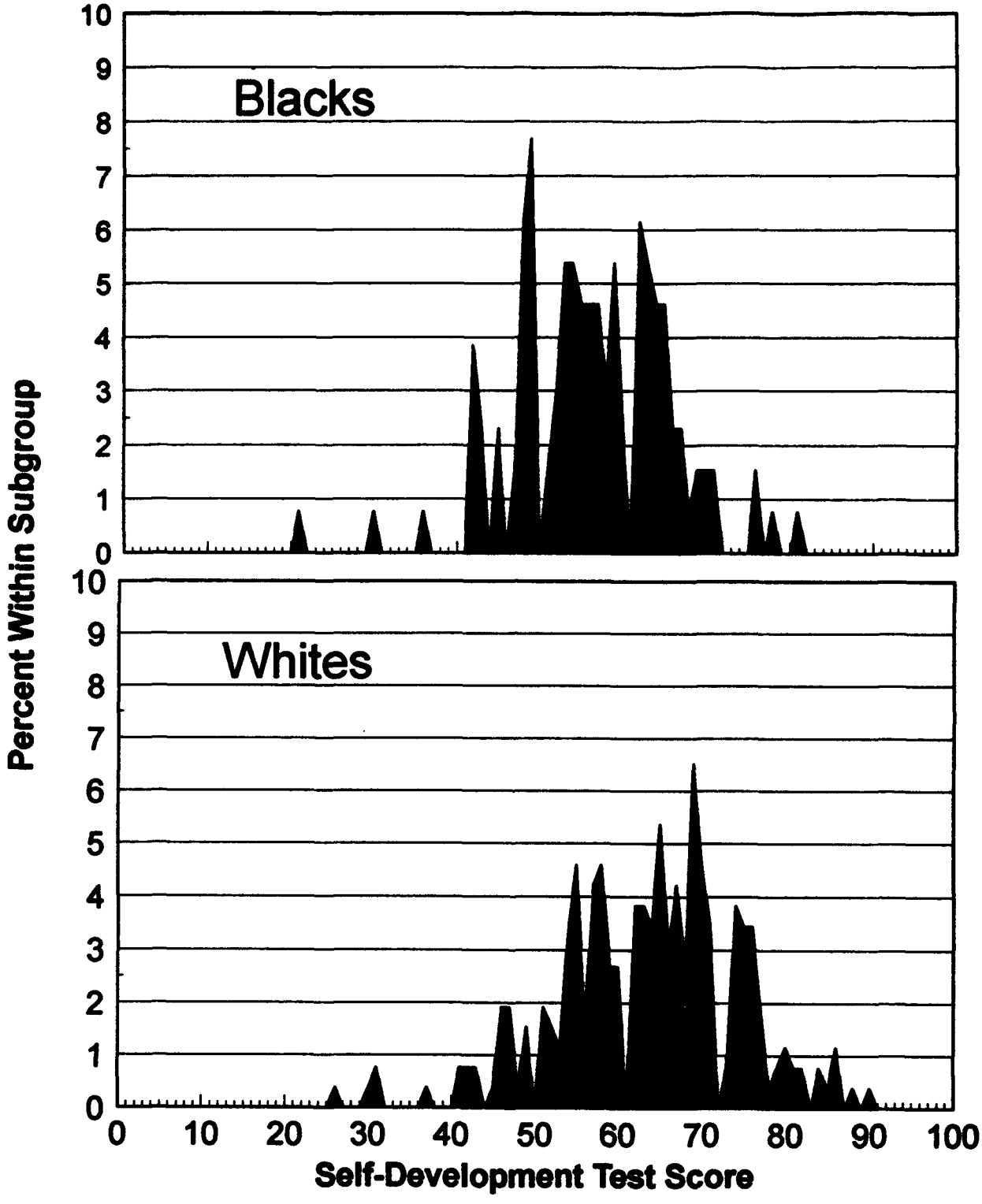



**Figure 1. SDT Score Distribution for Blacks and Whites for SDT Version 12C(2).**

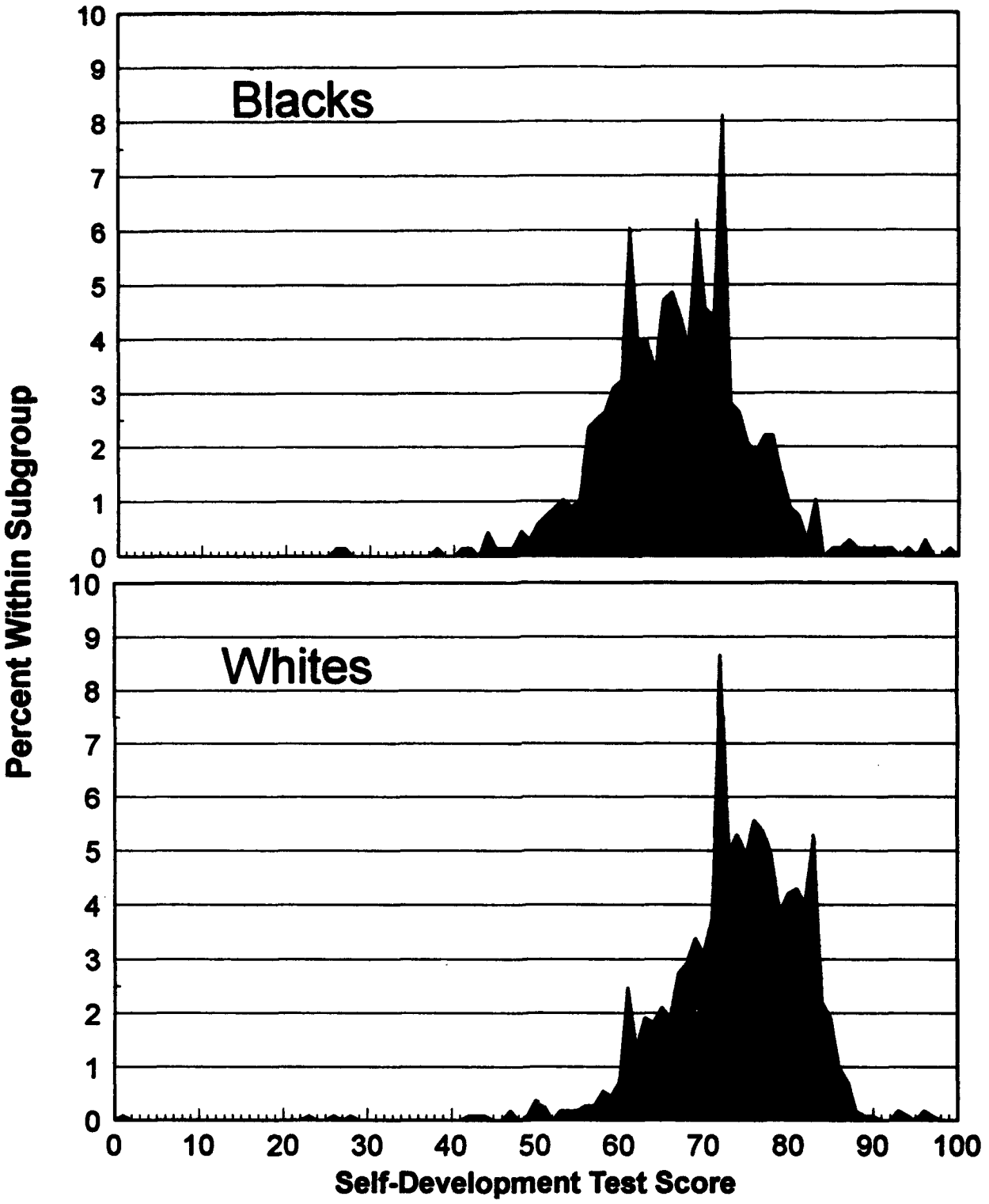



**Figure 2. SDT Score Distribution for Blacks and Whites for SDT Version 63B(2).**

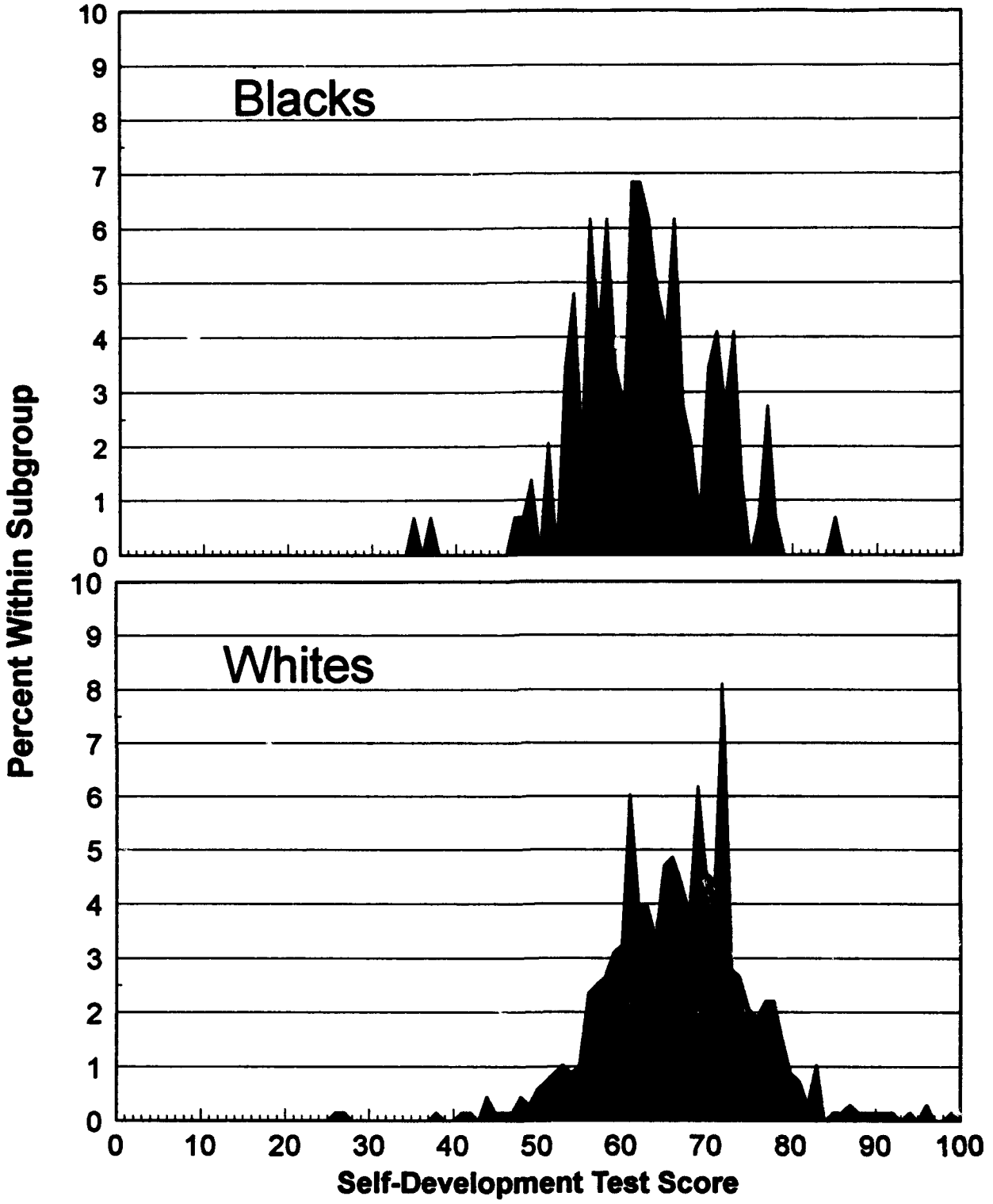



**Figure 3. SDT Score Distribution for Blacks and Whites for SDT Version 63H(2).**

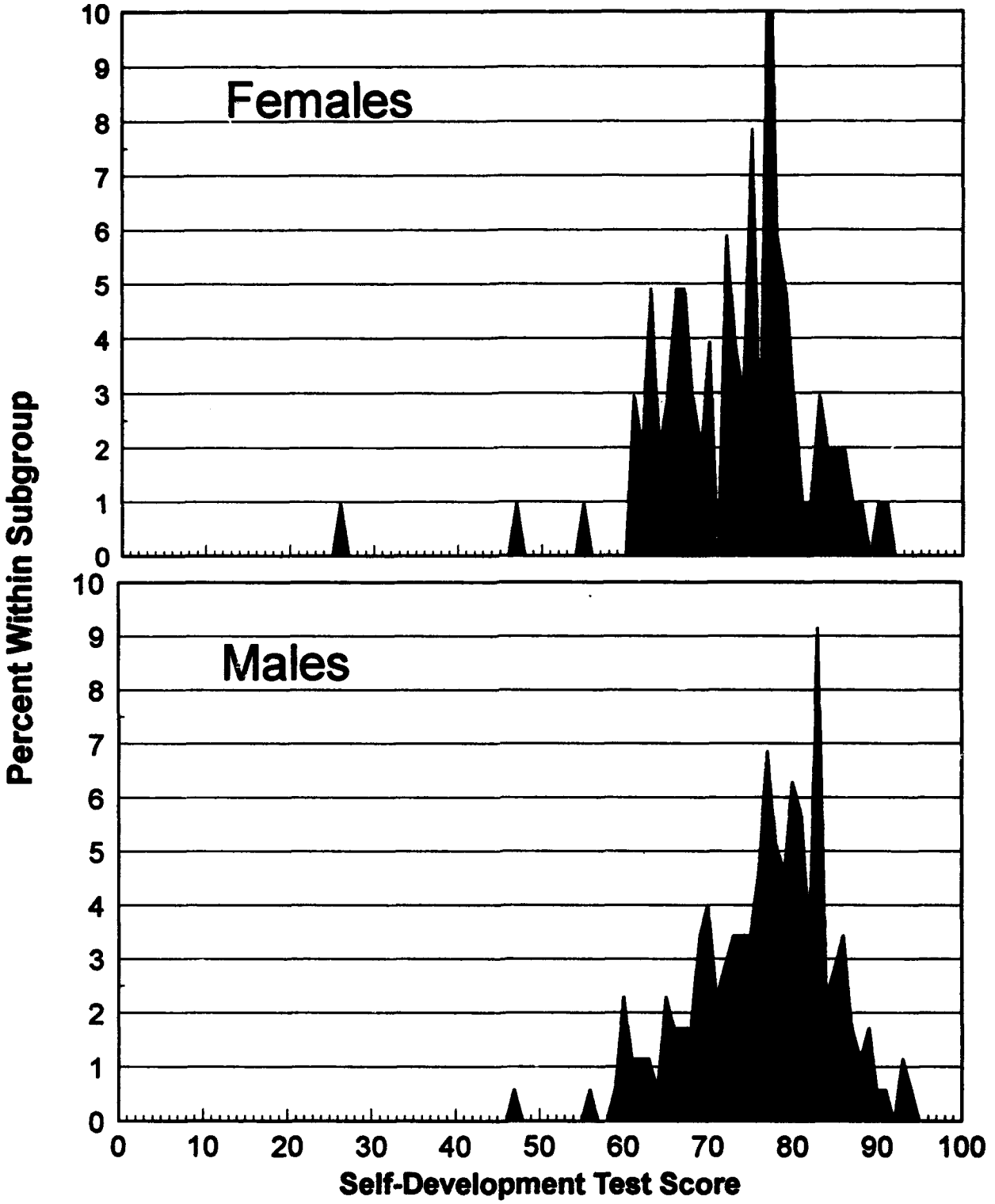



**Figure 4. SDT Score Distribution for Females and Males for SDT Version 71M(2).**

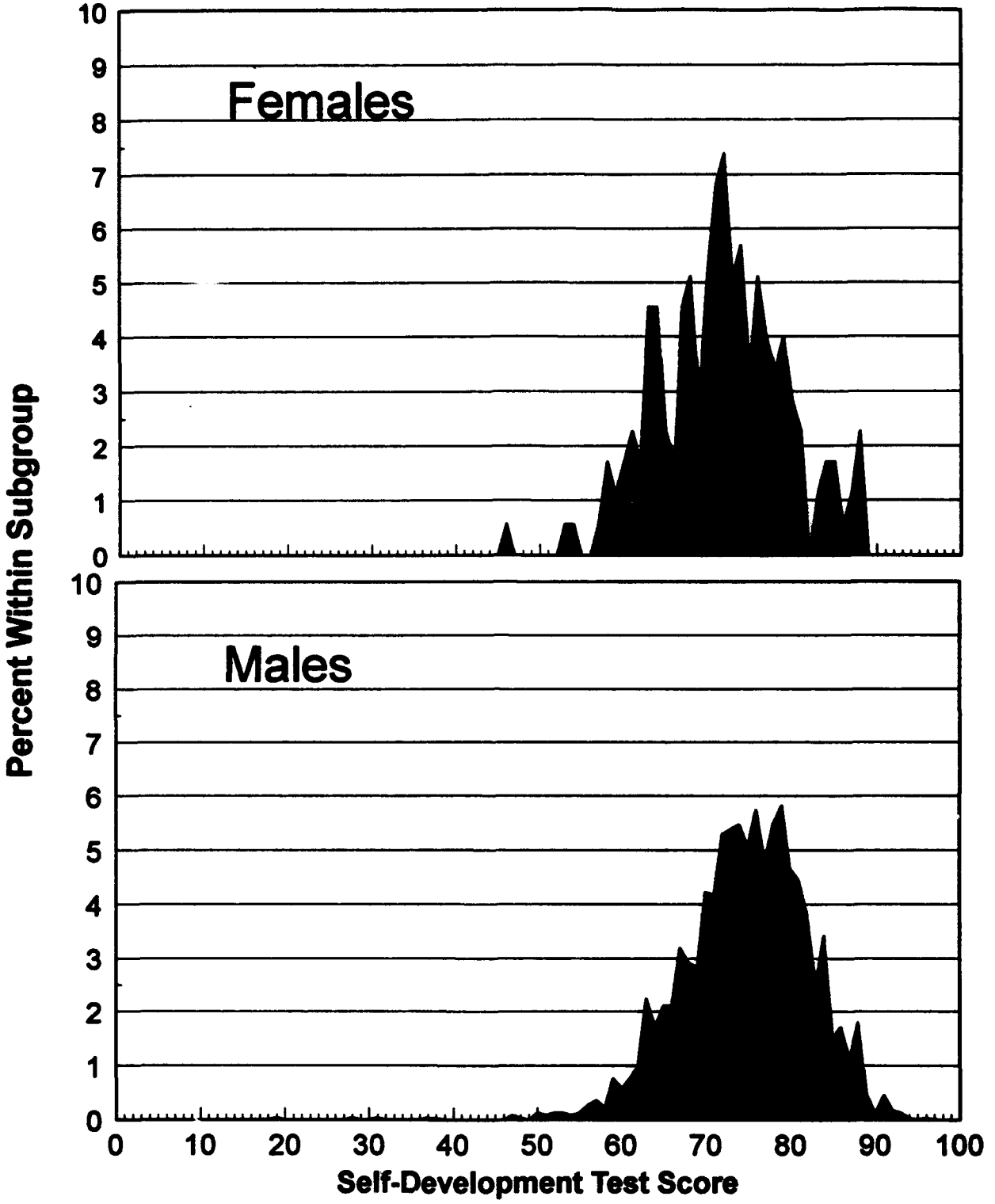



**Figure 5. SDT Score Distribution for Females and Males for SDT Version 88M(2).**

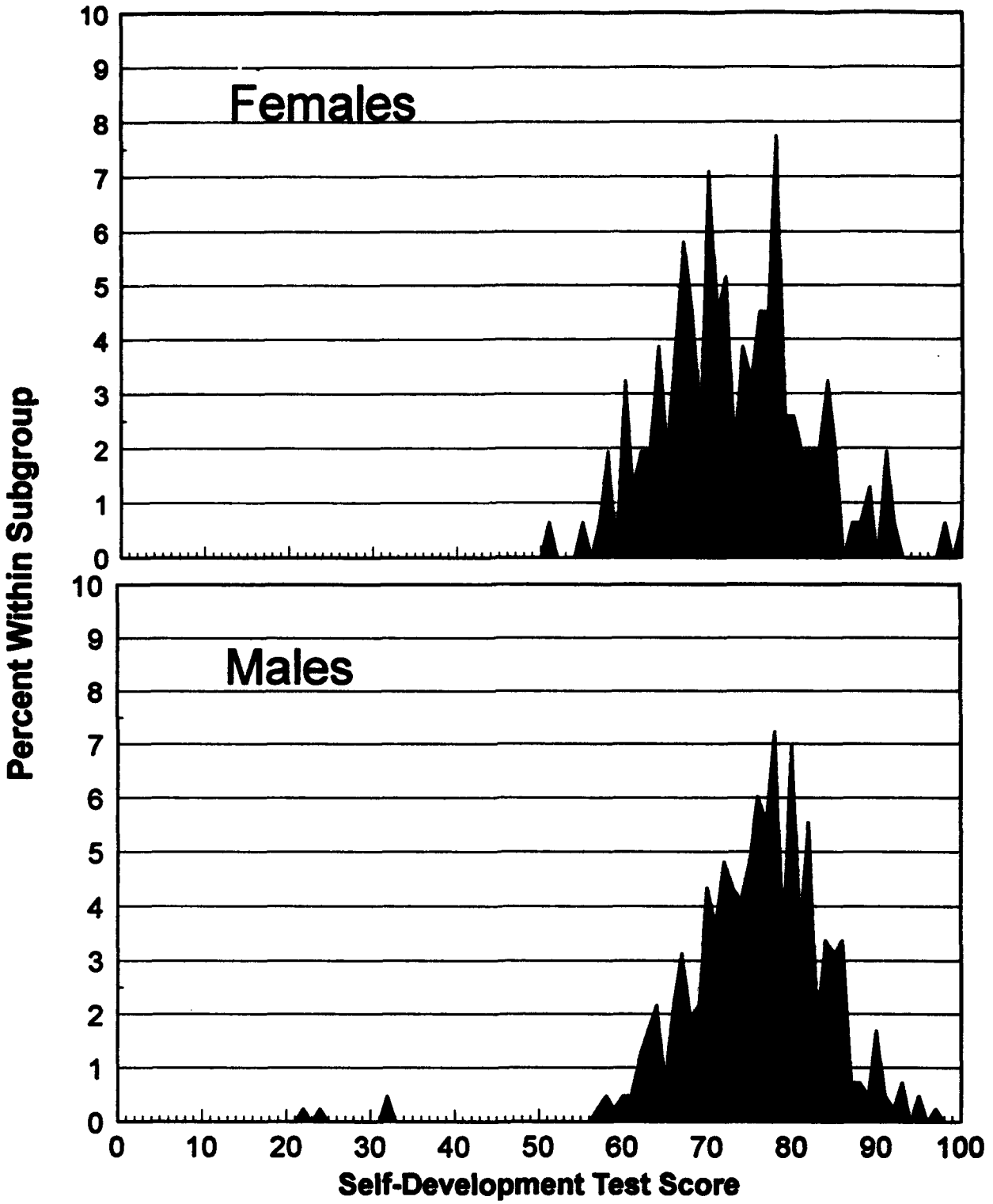



**Figure 6. SDT Score Distribution for Females and Males for SDT Version 91C(4).**

**Table 3**

**Means, Standard Deviations, and Sample Sizes for SDT Total Score by Subgroup and SDT Version**

| | Subgroup | | | |
|---|---|---|---|---|
| SDT Version | Minority | Majority | Δ | Overall |
| 12C(2) | 56.38<br>[9.23]<br>(130) | 63.43<br>[10.85]<br>(261) | -0.76 | 61.09<br>[10.85]<br>(391) |
| 63B(2) | 66.48<br>[8.59]<br>(677) | 73.52<br>[8.54]<br>(1096) | -0.82 | 70.83<br>[9.22]<br>(1773) |
| 63H(2) | 62.50<br>[7.83]<br>(146) | 72.25<br>[7.62]<br>(219) | -1.25 | 68.35<br>[9.06]<br>(365) |
| 71M(2) | 72.94<br>[9.11]<br>(102) | 76.59<br>[8.02]<br>(175) | -0.40 | 75.24<br>[8.60]<br>(277) |
| 88M(2) | 71.58<br>[7.52]<br>(176) | 74.62<br>[7.61]<br>(2228) | -0.40 | 74.39<br>[7.64]<br>(2404) |
| 91C(4) | 72.74<br>[8.44]<br>(152) | 75.77<br>[8.64]<br>(414) | -0.36 | 74.96<br>[8.68]<br>(566) |

Notes. In minority and majority columns unsurrounded entries are means, entries surrounded by [ ] are standard deviations, and entries surrounded by ( ) are sample sizes. Entries in Δ column are standardized differences between the minority and majority subgroups where a negative entry indicates lower mean scores for the minority subgroup. Δ was computed as follows: (minority mean - majority mean) / minority standard deviation. Overall column is based on examinees irrespective of subgroup membership. The minority subgroup for SDT versions 12C(2), 63B(2), and 63H(2) was Blacks and the minority subgroup for 71M(2), 88M(2), and 91C(4) was females.

SDT versions with the largest Black-White differences. For example, the largest reduction occurred in version 12C(2) where the Black-White difference was reduced from 31 to 22 percentage points after SDT score adjustment based on the AFQT. Most of the minority-majority subgroup differences still persisted.

Race and Gender Adjustments. The sex and race of an examinee are not mutually exclusive characteristics. That is, males and females can be Black or White and vice versa. It is possible that what appears to be a gender difference may actually be mainly a race difference. Consider the following extreme example where a gender effect was found. If all the males were White and all the females were Black then what appears to be a gender effect may actually be a race effect. To examine the degree to which the three largest gender differences were a function of gender rathger than race, the distribution of Blacks and Whites within each gender was examined.

The relative distribution of Blacks and Whites was similar within each gender for two of three SDT versions. The third, 71M(2), however, showed a substantially higher proportion of Black females compared to Black males (i.e., 72% of the females were Black while only 31 percent of the males were Black). Adjusting SDT scores for race eliminated the observed gender difference in 71M(2), but had no substantial impact in versions 88M(2) and 91C(4).

The converse was not true for the three SDT versions with largest race effects. Adjusting Black-White differences using examinee gender did not eliminate or even substantially reduce the Black-White differences. The differences in their relatively unchanged state persisted.

The results below report analyses for 71M(2) based on gender. However, given the distribution of Blacks within each gender and the elimination of a gender difference after adjusting for race makes it clear that the observed difference across genders was actually a function of race for SDT version 71M(2).

## Identification of Potentially Biased Items

Items and item counts within SDT section are presented in Table 4. The number of potentially biased items within an SDT version ranged from 14 to 67, or 14% to 61%, respectively. Within the Leadership section 0 to 10 items were identified, representing 0% to 50% of these items, respectively. Within the Training Management section 3 to 9 items were identified, representing 15% to 45% of these items, respectively. Finally, for the MOS Knowledge section 10 to 47 items were identified, representing 13% to 68% of these items, respectively.

Although SDT versions 63B(2) and 63H(2) share the same Leadership and Training Management items and were both examined with respect to Black-White differences, the same items were not consistently flagged as potentially biased across the two SDT versions. For the Leadership section, 5 of the 6 items flagged in 63H(2) matched 5 of the 10 items flagged in

**Table 4**

**Item Numbers and Counts of Items Identified as Potentially Biased for Each SDT Section by SDT Version**

| SDT Version | Leadership | Training Management | MOS Knowledge | Total Item Count |
|---|---|---|---|---|
| 12C(2) | [5/25%] | [3/15%] | [12/24%] | [20/22%] |
| | 4 7 11 14 16 | 27 31 33 | 42 44 49 59 60 63 64 65 66 67 69 84 | |
| 63B(2) | [11/55%] | [9/45%] | [47/68%] | [67/61%] |
| | 1 4 6 7 11 12 14 16 17 18 20 | 21 23 25 27 31 33 37 38 39 | 41-46 49-56 58-65 67-71 73-75 77-79 82 85-87 89-91 93-94 99 104-106 108 | |
| 63H(2) | [6/30%] | [7/35%] | [37/64%] | [50/51%] |
| | 1 6 11 16 17 18 | 21 24 25 27 37 38 39 | 41 43 44 46 47 51 53-55 61-64 67 68 70 73 74 76 77 80-83 87-90 92-98 | |
| 71M(2) | [1/5%] | [5/25%] | [10/13%] | [16/14%] |
| | 7 | 24 25 27 38 39 | 50 54 72 75 85 110 111 112 114 115 | |
| 88M(2) | [0/0%] | [3/15%] | [21/33%] | [24/23%] |
| | | 24 25 37 | 51 54 57 60 62 64 66 71 75 83 84 87 89-92 95 96 99 100 104 | |
| 91C(4) | [1/5%] | [3/15%] | [10/17%] | [14/14%] |
| | 9 | 21 22 38 | 49 54 55 67 76 84 91 94 99 100 | |

Notes. The bracketed numbers inside the table can be interpreted as follows: [number of items identified as potentially biased/percent of items in the section identified as potentially biased]. Unbracketed numbers represent actual item numbers in SDT version. SDT versions enclosed in box share the same Leadership and Training Management items.

63B(2). For the Training Management section, 6 of the 7 items flagged in 63H(2) matched 6 of the 9 items flagged in 63B(2).

Likewise, although SDT versions 71M(2) and 88M(2) share the same Leadership and Training Management items and were both examined with respect to male-female differences, the same items were not consistently flagged as potentially biased across the two SDT versions. For the Leadership section, no items were flagged in 88M(2) while one item was flagged in 71M(2). For the Training Management section, 2 of the 3 items flagged in 88M(2) matched 2 of the 5 items flagged in 71M(2).

## Correlations of Potentially Biased Items with "Bias Adjusted Scores" Across Subgroups

For each subgroup, correlations between potentially biased items and SDT section and total scores (i.e., after removing all potentially biased items) were computed. Correlations of the potentially biased items with their respective section score revealed correlations in the .05 to .30 range with the bulk near .20. Prior to removing the potentially biased items these correlations were closer to .30. This reduction in the item-section point-biserial correlations may indicate that the removal of these items adversely affected the internal consistency of the SDT section. The lower point-biserial correlations observed after removal of the potentially biased items may indicate that the removed items may have served to more fully and reliably define the intended construct.

A few of the potentially biased items correlated differently with the construct across subgroups. These are examined next. Items identified as showing differential correlations across subgroups were categorized in terms of direct knowledge items and application of knowledge items. A direct knowledge item requires only that the examinee know the information. For example, "What do you use to check tire pressure?" Application of knowledge items require knowledge of certain information, as well as, the choice of a course of action or to otherwise apply that knowledge in a specific context. For example, "What should you do if you find the battery discharged?"

In SDT version 12C(2) only 2 of the 20 items showed subgroup-different correlations with the construct. Both of these items were in the MOS Knowledge section and involved application of knowledge in the form of mathematical calculations. SDT version 63B(2) contained the largest number of items with construct correlations which differed between Blacks and Whites. Two items were identified in the Leadership section and eight in the MOS Knowledge section. Both items in the Leadership section required application of knowledge. Of the eight items in the MOS Knowledge section, 5 items involved direct knowledge and 3 required the application of knowledge in a specific situation.

SDT version 63H(2) contained three such items all in the MOS Knowledge section. All three involved direct knowledge. SDT version 71M(2) contained one such item in the Leadership

section and one in the MOS Knowledge section. The Leadership item required application of knowledge and the MOS Knowledge item involved direct knowledge.

SDT version 88M(2) contained three such MOS Knowledge items. Two required application of knowledge and one was a direct knowledge item. Finally, SDT version 91C(4) contained only one such item in the MOS Knowledge section and it involved direct knowledge.

Overall consideration of the items identified as correlating differently with the intended construct across subgroups did not reveal any discernable basis for the effect. The items were clearly written using simple vocabulary and grammatical structure. They spanned direct knowledge items and more complex application of knowledge items. One possible explanation is that members of the subgroups have different exposure to the material covered by these items.

## Impact of Removing All Potentially Biased Items from the SDT

The purpose of this analysis is not to suggest that all potentially biased items be removed from the SDT. Rather it is meant to examine the impact of a very extreme action one could consider. If large differences still remain, then perhaps the answer lies not in these specific items but rather in some aspect of the context. For example, is one subgroup differentially assigned to duty positions that provide less exposure to the test content? Does one subgroup prepare less for the test? Or is one subgroup less capable of acquiring the knowledge being tested?

Figures 7 and 8 show the impact of removing those items identified as potentially biased. Some SDT versions showed a great reduction and even complete elimination of subgroup differences after all potentially biased items were removed. Others, however, indicated the impact to be minimal. Figure 7 shows differences in the percentage scoring in the upper third across subgroups before and after removing all potentially biased items. Based on the Total score, versions 63B(2) and 63H(2) showed the greatest gains for the minority subgroup. However, even after removing these items, version 63H(2) still retained a 12 percentage point differential. In other test versions, 12C(2) and 71M(2) retained most of the original subgroup difference. Versions 88M(2) and 91C(4) reduced their subgroup difference by nearly two-thirds, dropping the percentage difference from about 15% to 5%. A similar but not identical pattern can be observed in the mean score difference across subgroups presented in Figure 8.

Examining the impact on the three individual SDT sections revealed that the greater the subgroup difference within each section and across the three sections, the more the subgroup difference reduction in the total score. For example, 63B(2) in addition to showing the largest subgroup difference reduction on the Total score also showed large reductions in all three SDT sections.

Large subgroup differences still remained for most SDT versions. On the Total score version 12C(2) retained a 20 percentage point difference, version 63H(2) retained a 12 percentage point difference, version 71M(2) retained an 18 percentage point difference, and versions 88M(2) and 91C(4) retained 6 percentage point differences. Although the removed items accounted for

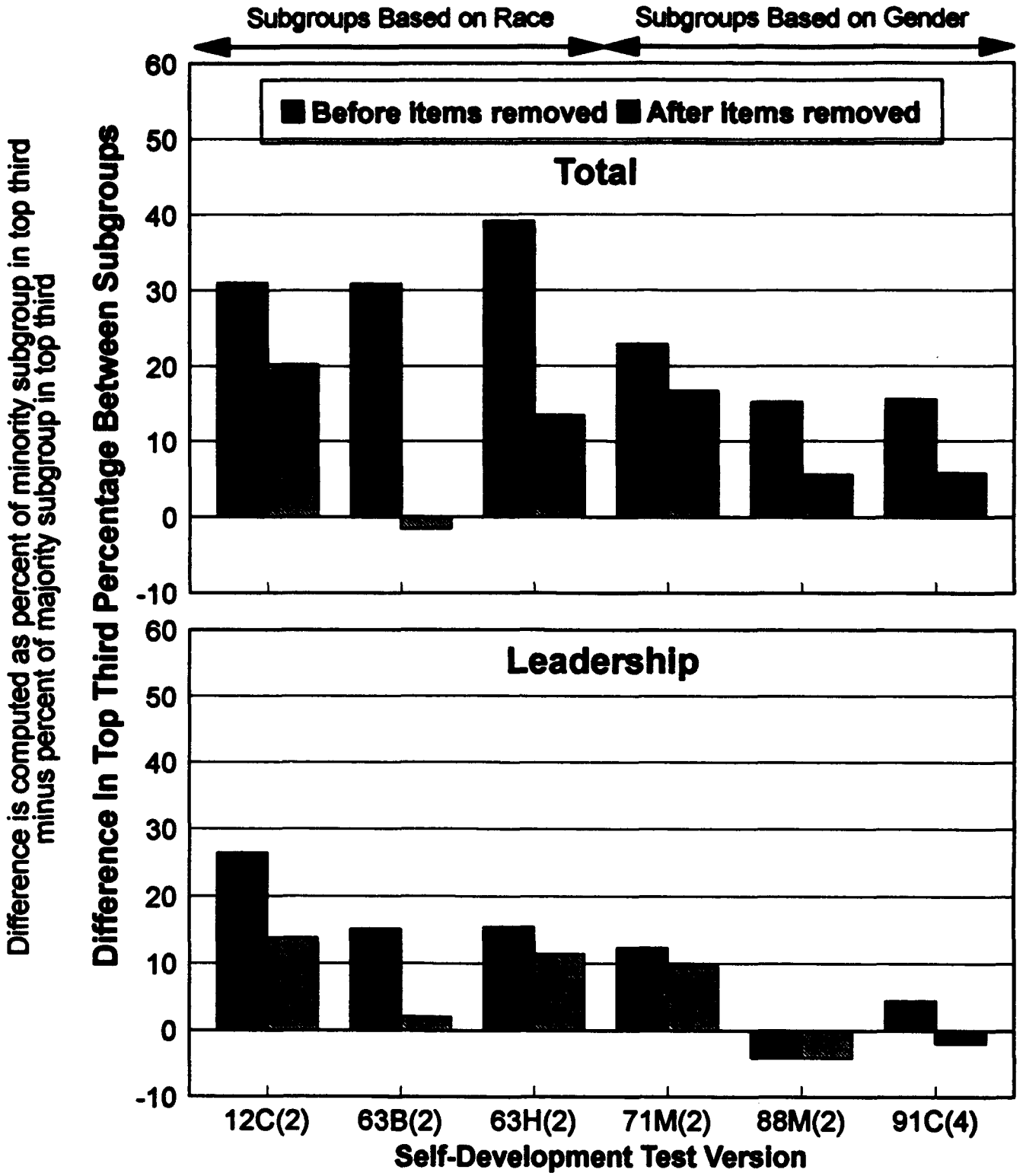



**Figure 7. Difference in Subgroup Representation in Upper Third of Score Distribution Before and After Removing Potentially Biased Items for Each SDT Version.**

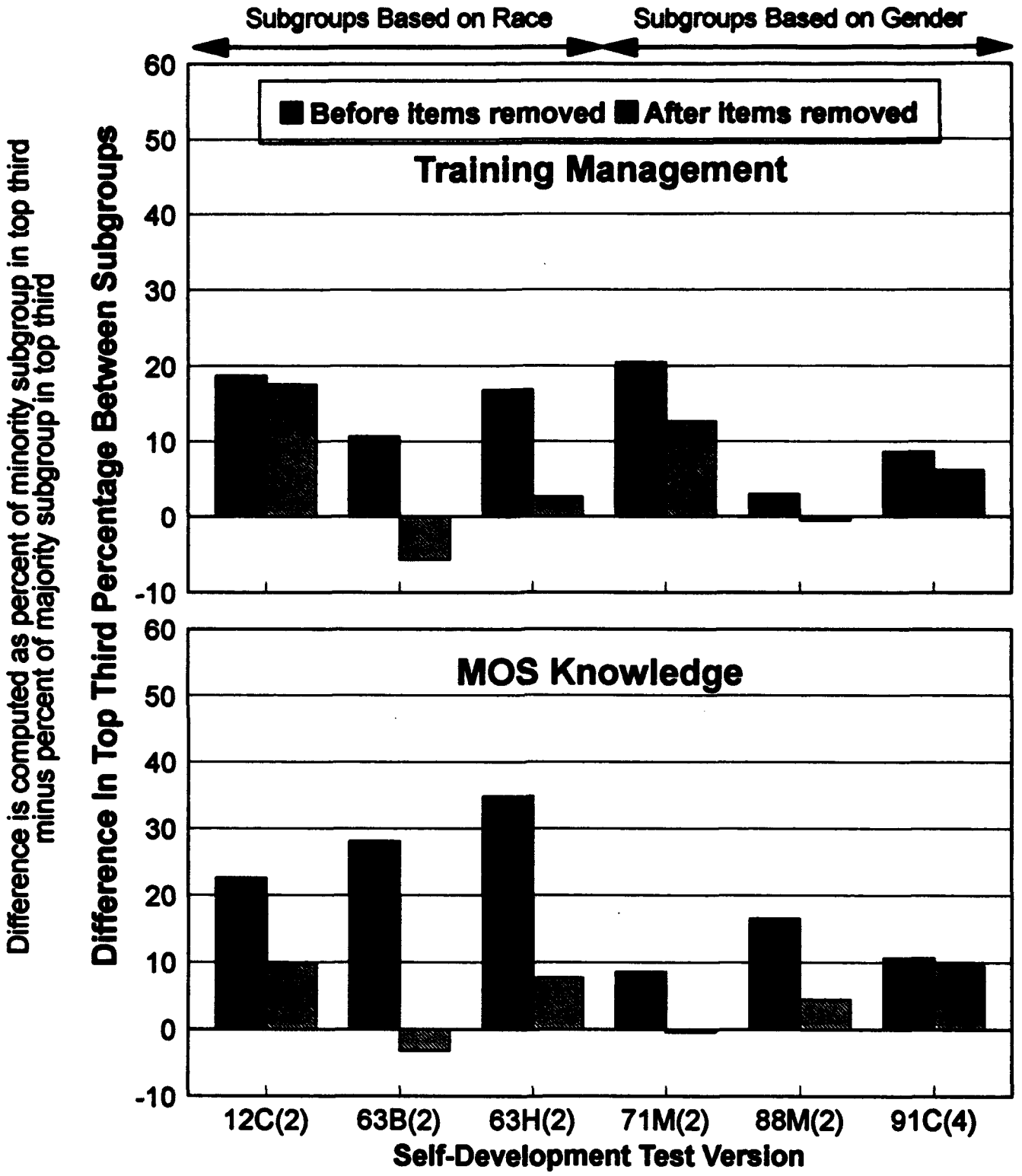



**Figure 7 (Continued). Difference in Subgroup Representation in Upper Third of Score Distribution Before and After Removing Potentially Biased Items for Each SDT Version.**

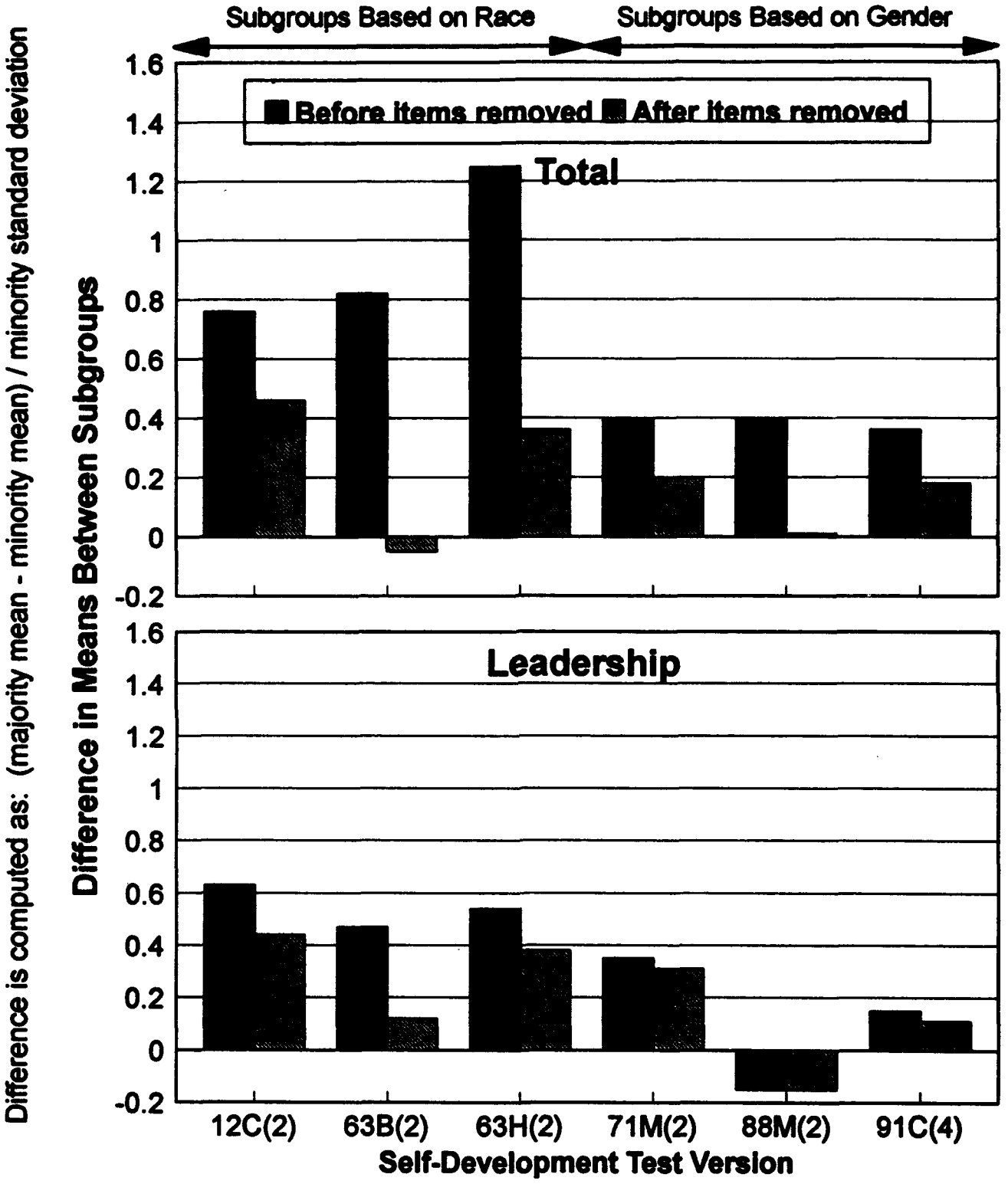



**Figure 8. Difference in Subgroup Means Before and After Removing Potentially Biased Items for Each SDT Version.**

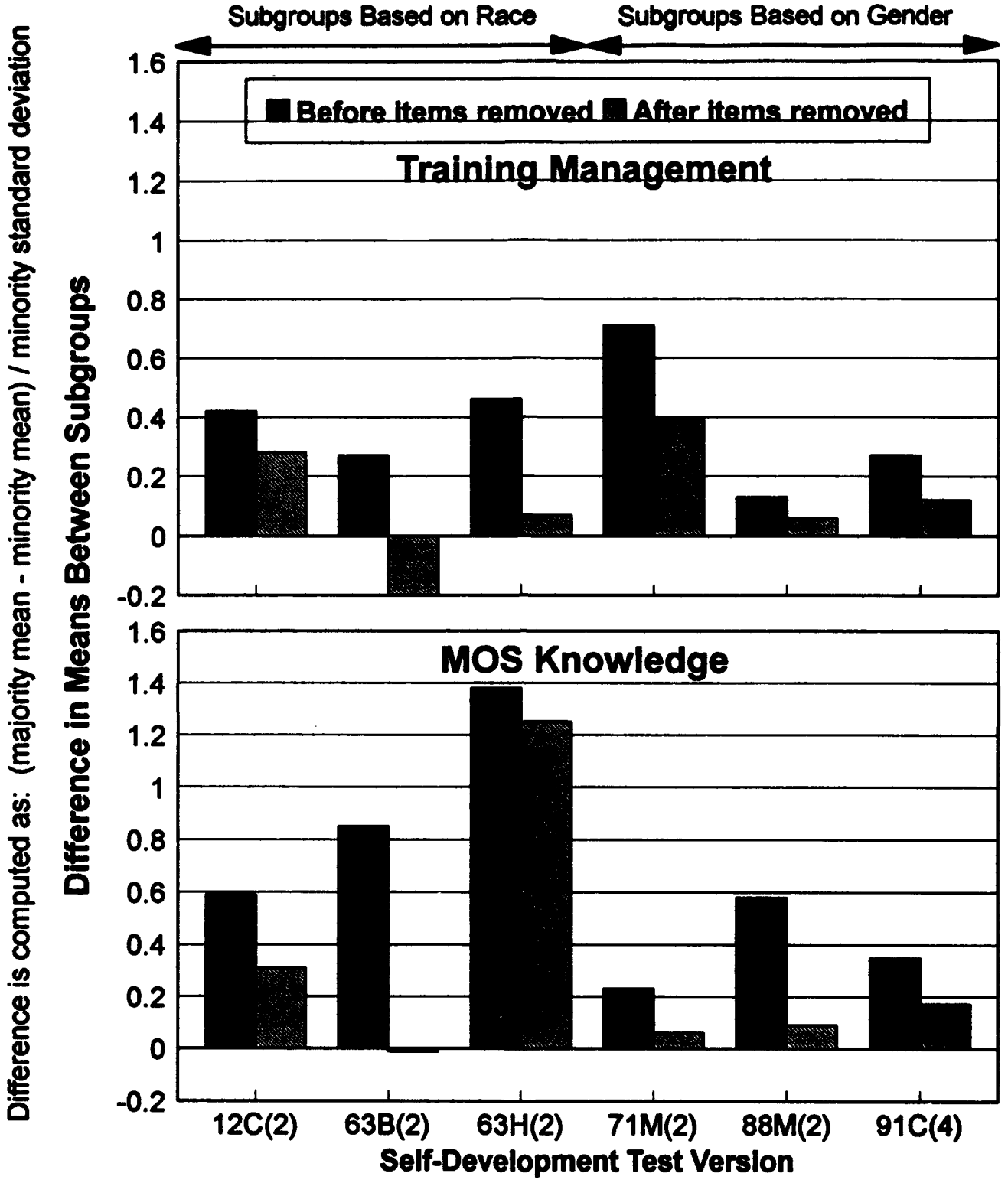



**Figure 8 (Continued). Difference in Subgroup Representation Means Before and After Removing Potentially Biased Items for Each SDT Version.**

some of the original subgroup difference, the remaining difference indicates that they may have only been indicators of a more general underlying problem.

## Subject Matter Expert Responses

Table 5 shows the distribution of SMEs by source, race, and sex. The only SDT versions not examined by more than 10 SMEs were 12C(2) and 71M(2). All 10 12C(2) MOS SMEs refused to do the task as requested. They wrote that they thought the items presented were fair to both groups and refused to identify any specific items. Some comments also made it clear that they thought the task was offensive in that it suggested that some of the items were biased. The reason that all 10 responded the same way to the task was a direct result of one of the SMEs standing up and refusing to do the task for the aforementioned reasons.

Black and White SMEs were nearly equally represented in all SDT versions. For SDT versions with the largest male-female differences, the number of female SMEs was slightly greater than male SMEs. However, in the SDT versions with the largest Black-White differences, the number of female SMEs was substantially less than male SMEs except in SDT version 12C(2) where all the SMEs (i.e., all males) refused to perform the task. All female SMEs for these SDT versions were Testing Expert SMEs; none were MOS SMEs.

Ability to Identify Potentially Biased Items. Table 6 presents the mean number of potentially biased items correctly identified by the SMEs. These means are shown overall, as well as by race and gender. Across all SMEs, only for SDT version 63H(2) did the SMEs identify potentially biased items at a rate which exceeded the rate of chance guessing and was statistically significant. The mean number of potentially biased items identified for all other SDT versions hovered around the chance level.

For 63H(2) both Blacks and Whites identified a statistically significant number of potentially biased items in excess of the chance level. The same holds true for males. The mean for females, however, was not in excess of the chance level (i.e., not statistically significant). This was because only two female SMEs participated in the 63H(2) item evaluation.

All other means which were different from chance level and were statistically significant were below chance level. Male SMEs in 63B(2), for example, identified less items than one would expect if they had just guessed;. similarly, female SMEs who evaluated 71M(2) and 91C(4) SDT versions identified less items than one would expect if they had just guessed. This suggests that SMEs may have focused on item dimensions which were unrelated to subgroup performance differences. For example, many females who evaluated version 71M(2) selected item 4 as being biased in favor of males because it referred to the subject as a male. Performance on this item, however, showed no performance difference between males and females.

The mean number of items identified by SMEs who evaluated version 63H(2) is of concern. In the analyses reported below this will be explored further.

**Table 5**

**SME Counts by Characteristic for Each SDT Version**

| | Source | | Race | | | Gender | |
|---|---|---|---|---|---|---|---|
| SDT Version | MOS | Testing Expert | Black | White | Other / Unknown | Male | Female |
| 12C(2) | 0 | 4 | 2 | 2 | 0 | 2 | 2 |
| 63B(2) | 10 | 4 | 7 | 6 | 1 | 12 | 2 |
| 63H(2) | 9 | 4 | 7 | 6 | 0 | 11 | 2 |
| 71M(2) | 5 | 4 | 4 | 4 | 1 | 4 | 5 |
| 88M(2) | 9 | 4 | 6 | 6 | 1 | 6 | 7 |
| 91C(4) | 9 | 4 | 6 | 6 | 1 | 6 | 7 |

**Table 6**

**Mean Number of Potentially Biased Items Identified by SMEs**

| | | Race | | | Gender | |
|---|---|---|---|---|---|---|
| SDT Version | Overall | Black | White | Other / Unknown | Male | Female |
| 12C(2) | 5.75 | 5.00 | 6.50 | ---- | 5.00 | 6.50 |
| 63B(2) | 4.43 | 4.14 | 4.67 | 5.00 | 4.25* | 5.50 |
| 63H(2) | 6.15** | 6.43* | 5.83* | ---- | 6.09* | 6.50 |
| 71M(2) | 4.56 | 5.00 | 3.75 | 6.00 | 4.75 | 4.40* |
| 88M(2) | 5.31 | 5.33 | 5.17 | 6.00 | 5.67 | 5.00 |
| 91C(4) | 4.62 | 4.50 | 4.50 | 6.00 | 4.83 | 4.43* |

Notes. A mean not significantly different from 5 was expected when SMEs were guessing. A mean below 5 may indicate that SMEs focused on item dimension(s) that were unrelated to the potential bias in the items (i.e., items with the characteristic the SMEs focused on were not the potentially biased items). A mean significantly above 5 may indicate that SMEs focused on item dimension(s) which were related to the potential bias in the items (i.e., items with the characteristic the SMEs focused on were the potentially biased items). Significance test examines difference of the mean from 5. * $p < .05$. ** $p < .01$.

Consistency of Items Chosen. Table 7 presents, for the three SDT versions with the largest gender differences, the percentage of SMEs who selected each of the 10 potentially biased items included in the SME instrument. In addition, it shows the percentage of the SMEs who correctly identified the subgroup which had more difficulty with the item. Table 8 presents the same information for the three SDT versions with the largest Black-White differences.

The four testing expert SMEs who evaluated SDT version 12C(2) flagged potentially biased items with a consistency ranging from 25% to 100%, and an average of 58% consistency across all 10 potentially biased items. One would expect the consistency to be 50% if identification of items was at chance level. The observed mean of 58% is well within the limits of chance guessing because of the low number of SMEs who provided usable responses for this SDT version. When they correctly identified a potentially biased item they were, on average, likely to correctly identify the disadvantaged subgroup 95% of the time. This is expected because the SMEs understood that these items were being studied because of lower minority subgroup performance.

For SDT version 63B(2), the consistency with which SMEs flagged potentially biased items ranged from 14% to 57%, and averaged 44% (i.e., less but not statistically different than one is expected to identify by chance guessing). When SMEs correctly identified a potentially biased item, they were able to correctly identify the lower performing group 68% of the time (i.e., better than chance). Even the MOS SMEs probably understood that these items were being studied because of lower minority subgroup performance. Overall, no items stood out as being a potential problem.

SMEs who evaluated version 63H(2) showed higher average consistency as comparted to previous results. One item, item 18, was identified at a rate in excess of chance guessing. SMEs identified nothing specific about the items that would lead Blacks to perform worse. Over half of the SMEs who chose this item indicated they were guessing. Only one SME wrote that the wording might be a problem for Blacks.

Two other items in version 63H(2) were chosen consistently but not above the statistical chance level. Since 63H(2) showed higher than average SME consistency in identifying the potentially biased items, these two items were also examined. Items 11 and 44 both were identified by 77% of the SMEs as potentially biased items. Item 44 is a seemingly simple knowledge question. Item 11 asks about the importance of values and beliefs. This is perhaps more affected by the cultural background of the examinees than the other items. However, the reading material with which examinees are provided clearly identifies the correct answer. Cultural background may have an impact when individuals have not studied. Perhaps Whites who have studied to the same extent as Blacks are more predisposed to the correct answer because of some as of yet unidentified aspect of their cultural background. However, no specific aspect of the cultural background of Blacks was offered by either the White or Black SMEs to explain the lower performance of Blacks on this item. Overall there is nothing specific to indicate that any of the items are inherently biased and no specific cultural bases for test performance differences were identified by the SMEs for these or any other items.

Table 7

**Percent of Potentially Biased Items Correctly Selected and Percent Where Worst Performing Subgroup was Correctly Identified by SMEs for the Three SDT Versions with the Largest Gender Subgroup Differences**

| SDT Version | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 12C(2) | | | 63B(2) | | | 63H(2) | | |
| Item | % Chose Item | % Direction Correct | Item | % Chose Item | % Direction Correct | Item | % Chose Item | % Direction Correct |
| 7 | 25 | 100 | 7 | 57 | 75 | 11 | 77 | 100** |
| 11 | 75 | 100 | 14 | 50 | 71 | 17 | 62 | 63 |
| 14 | 50 | 100 | 18 | 57 | 88* | 18 | 85* | 73 |
| 27 | 50 | 100 | 27 | 57 | 50 | 24 | 31 | 100* |
| 31 | 25 | 100 | 31 | 50 | 57 | 25 | 62 | 88* |
| 33 | 25 | 100 | 37 | 57 | 50 | 37 | 46 | 83 |
| 42 | 50 | 50 | 51 | 14 | 50 | 44 | 77 | 70 |
| 59 | 75 | 100 | 53 | 29 | 75 | 76 | 46 | 67 |
| 67 | 100 | 100 | 87 | 50 | 43 | 95 | 69 | 67 |
| 84 | 100 | 100 | 89 | 21 | 67 | 96 | 62 | 75 |
| Mean | 58 | 95** | | 44 | 68** | | 62* | 79** |

Notes. Tested the hypothesis that the percentage was based on random guessing (i.e., 50%).
* $p < .05$. ** $p < .01$.

Table 8

Percent of Potentially Biased Items Correctly Selected and Percent Where Worst Performing Subgroup was Correctly Identified by SMEs for the Three SDT Versions with the Largest Race Subgroup Differences

| SDT Version | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 71M(2) | | | 88M(2) | | | 91C(4) | | |
| Item | % Chose Item | % Direction Correct | Item | % Chose Item | % Direction Correct | Item | % Chose Item | % Direction Correct |
| 7 | 44 | 50 | 24 | 38 | 20 | 9 | 77 | 40 |
| 24 | 33 | 33 | 25 | 38 | 20 | 21 | 69 | 89* |
| 25 | 44 | 75 | 37 | 54 | 14 | 22 | 23 | 67 |
| 27 | 33 | 67 | 60 | 62 | 75 | 38 | 54 | 57 |
| 38 | 0* | ----- | 62 | 54 | 86** | 55 | 23 | 67 |
| 50 | 78* | 86 | 64 | 54 | 86** | 84 | 85* | 0 |
| 75 | 22 | 50 | 83 | 23 | 100** | 91 | 31 | 100* |
| 85 | 56 | 80 | 84 | 54 | 43 | 94 | 15* | 50 |
| 111 | 67 | 100* | 99 | 62 | 88** | 99 | 38 | 40 |
| 115 | 78 | 100* | 100 | 69 | 89** | 100 | 46 | 33 |
| Mean | 45 | 71** | | 51 | 54 | | 46 | 54 |

Notes. Tested the hypothesis that the percentage was based on random guessing (i.e., 50%).
* $p < .05$. ** $p < .01$.

For 71M(2) item 50 was identified by 78% of the SMEs as potentially biased. It is a straightforward knowledge item about field operations which appears inherently unbiased. One explanation for the identification of this item would be if females are less likely to be assigned to duty positions where they would acquire this knowledge.

SMEs who evaluated 88M(2) did not identify any one item very consistently. For 91C(4) item 84 was consistently identified as being more difficult for males because it dealt with female anatomy. It actually was more difficult for females.

Rationale. Choosing Items. Few SMEs wrote many specific comments. The comments that were made were typically general such as "wording." These comments were just as likely to be made for items which showed subgroup differences as for items which did not. However, some comments were noteworthy. For example, seven of the ten MOS SMEs stated that females would do worse in MOS Knowledge items because they often "did not work in their MOS." Rather females were more likely to be assigned to administrative tasks. Four of the ten MOS SMEs for 71M(2) made similar comments. For example, one stated "the test is not gender bias[ed] but is bias[ed] [against] those 71Ms serving in garrison units only; if a 71M has not been in TOE units, he or she will not pass this test." Other 71M(2) SMEs stated that females do not have as much opportunity to serve in combat related units and this adversely affects their knowledge of the material tested. However, the SMEs who made these comments were not able to identify the "potentially biased" items.

These comments suggest that at least in these two MOS, 71M and 88M, the differences may be a result of different opportunities for males and females. However, no such differences in duty opportunities were mentioned by SMEs for Blacks and Whites.

Responses To Other Questions. The responses to questions at the conclusion of the primary task (i.e., "identify items on which subgroups might perform differently" task) indicate that SMEs were not confident that their choices were correct. The average rating was similar across SME groups (i.e., based on SDT version they evaluated) and on a scale of 1 to 10 averaged 4.51. On the questions which asked about their perceptions of how the minority subgroup was portrayed and balance in positive roles for the minority subgroup, the average rating indicated that SMEs thought Blacks and women were portrayed somewhat worse and given less positive roles in the items they evaluated. These differences are statistically significant (i.e., from a null hypothesis of no difference). Finally, SMEs did not find the items offensive other than the references to "he" or "him." The offensive nature of these sex-typed references suggests that they should be eliminated from future versions of the test even though they are not perceived as affecting the performance of females on those items. It is therefore recommended that references to "he" or "him" should either be eliminated or balanced with an equally frequent "she" or "her."

# Discussion

## Evaluation of Results

After an evaluation of the six SDT versions showing the largest race and sex effects, it is possible to conclude that there is no support for a general bias in the items in the instruments examined. In addition, since these are the versions showing the largest subgroup differences, it also possible to cautiously suggest that if any item bias exists in the remaining SDT versions, it would be very minimal at most.

In the introduction to the issue of test fairness it was stated that at least three issues related to fairness could be identified. They were labeled content, construct, and predictive fairness. The SMEs primarily addressed issues related to both content fairness (i.e., access to materials on which test versions are based and the job-relatedness of the test items) and construct fairness (i.e., is reading level appropriate? does the item tap the same construct across subgroups?). The comparison of the relationship of each test item to its respective section across subgroups also addressed issues of construct fairness (i.e., does the item tap the same construct across subgroups). The examination of impact of removing those items on which minority subgroups perform lower is also indirectly related to construct fairness. The fact that performance on the construct remained lower for minority subgroups after removing the potentially biased items suggested that the removed items measured components also tapped by the remaining items.

Nevertheless, some questions still remain unanswered. First, would subgroup-specific SMEs identify the same items as defining the constructs tapped by the SDT? If they would not, then perhaps differential experiences in the MOS or cultural differences in assimilating MOS experiences cause the differential definition of constructs. Differential definition of constructs may indicate that different subgroups place different importance on various knowledge used in performing the job, which would in part explain differential performance across subgroups.

Second, do minority individuals with the same job knowledge as majority subgroup members, as measured from observed behavior and supervisor estimates, perform differently on the SDT? If so, this suggests that the measurement medium (i.e., paper-and-pencil tests) differentially impacts minority subgroups.

Third, it was not possible to explore predictive fairness with the current data. Would members of two different subgroups who scored the same on the SDT, on average, perform the same on a job to which they were both promoted? If not, then the same SDT score, for predictive purposes, has different implications across subgroups. If, on average, a minority individual with an equally high SDT score performed better on the job to which he or she was promoted to, then perhaps a minority subgroup individual scoring lower on the SDT would perform equal to a majority subgroup individual scoring higher on the SDT. In that case, predictive fairness would require that separate predictive equations be used for the two subgroups.

Finally, there were some problems that were highlighted by this research, that although unrelated to the SDT in a perfect world, must be considered in the context in which it will operate. SME comments on SDT versions 71M(2) and 88M(2) suggest that females may, in relative frequency, be assigned to different duty positions than males. If the test content focuses on job content that is primarily contained in male occupied duty positions, then this may be a problem when the SDT is used to make promotion decisions.

Generally it is desirable for any component in a promotion point system to predict performance on the job at the next rank. For the SDT, it is not the knowledge itself that may be predictive, but rather what the knowledge indicates. For example, a soldier who is knowledgeable in the tasks in their duty position may indicate strong motivation and ability to learn the required material. If the occupants of a specific duty position are not given the opportunity to demonstrate their level of job knowledge applicable to their duty position, then relative to occupants of duty positions which are given this opportunity on a job knowledge test, they will appear unmotivated and unable to learn. When the occupants of these different duty positions differ on the basis of gender or race, then the use of this job knowledge instrument in the promotion process may disadvantage race and gender subgroups. It may not be gender or race biased if the specific knowledge content, and not its purpose as an indicator of motivation and ability to learn, is what gives it its power to predict performance at the next rank. However, in this case race or gender differential assignment would lower the rate of promotion for females and Blacks becuase they would be more often assigned to duty positions that have not prepared them for future promotional opportunities.

## Needed Research

Analyses should be conducted to explore the remaining research issues identified above. Namely, subgroup differences in defining job knowledge constructs, impact of test anxiety and other contextual factors on minority subgroup performance, and the potential of subgroup specific predictive validity should be studied.

Second, the differential assignment (i.e., within an MOS) of women and Blacks to duty positions is speculative, especially for Blacks. What we do know is that eleven of twenty SMEs in two MOS suggested that females are less likely to be assigned to duty positions that would allow them to acquire the knowledge and experience to perform equal to their male counterparts on the SDT. This should be explored further for MOS which show differential performance on the SDT, especially the MOS Knowledge section. Assignment procedures should be examined for these MOS and changed if there is no legitimate basis for the differential assignment.

Finally, it is recommended that future research examine the current six, as well as additional SDT versions for potentially biased items if and when operational data becomes available. It would be useful to compare subgroup differences on the six studied SDT versions under non-operational and operational conditions. A subset of the analyses conducted in the present research, as well as other analyses to be suggested, could be applied.

The most enlightening analyses in the present research were the SME reviews. This requires that potentially biased items be first identified. The procedure used in this research could be used. Altertnatively, a more sophisticated procedure which focuses on the performance on each item as a function of performance on an "unbiased" construct score could be used. One procedure of this type which is highly recommended because of its low sample size requirement and low computational cost is the Mantel-Haenszel procedure (Hills, 1989). In this procedure it would also be possible to easily incorporate Black-White AFQT differences in the process of identifying potentially biased items. This would be highly recommended. This would have been the procedure of choice in the present research if resources had permitted it.

The SME questionnaire should be modified in at least three ways. First, SMEs should be provided with a more extensive background into why they are being asked to do this task. This may reduce their resistance to perform the task as requested. Second, the confidence scale should be reworded in terms of how many items they believe they have identified correctly. Finally space should be added next to where the item numbers are entered for SMEs to write a short reason for their choice. They will be more likely to enter something if the space is specifically there for that purpose.

## Conclusions

No evidence was found for race or gender item bias on the SDT versions studied. However, more versions should be studied with the recommended modified procedure. While the SDT items may not be biased, the possibility that males and females, and perhaps Blacks and Whites, are assigned to duty positions based in part on their gender or race, suggests that the SDT, if included in promotion and school selection decisions (i.e., EPMS), may lower the promotion rates of Blacks and females via its content sampling of male and White dominated duty positions within an MOS. However, it is worthwhile to note that the MOS SMEs who suggested the existence of gender based assignment in MOS 71M and 88M also could not identify the "potentially biased" items. Further analysis of assignment procedures and subgroup distribution across duty positions is necessary to eliminate this potential form of adverse impact which may be impacting SDT scores, as well as other components of the existing promotion system.

Under EPMS, the SDT may be given a large weight or a small weight under either the centralized or decentralized promotion systems. For example, if the boards believe that the SDT is a good differentiator of key leadership aptitude, they may give it a large weight, otherwise it may receive a low weight or be ignored. It is not possible to know at this time which direction the boards will take. Other components of the existing EPMS also have potential for bias. For example, the attractiveness of the candidate as conveyed by the submitted picture may be considered. Additionally, other components of EPMS such as assignment history, courses taken, and NCO Efficiency Rating may also show subgroup differences as large as those observed on the SDT. If that is the case, then the proportion of women and Blacks promoted would not be subgroup proportional (i.e., women and Blacks would be promoted less often) regardless of whether the SDT is included. Iit is possible that the inclusion of the SDT under such a scenario could improve the promotion rates of Blacks and women. This would occur if the SDT

differentiated less between Blacks and Whites or women and men than the other components currently in use.

Analysis of Army data by Robinson & Prevette (1992) indicates that a disparity in promotion rates exists for Blacks and Whites. Robinson & Prevette found, for example, that the promotion rate for Black males was lower than for other groups in 14 of 15 Army enlisted promotion boards from 1987 to 1991. Robinson & Prevette also reported that enlisted women in the Army were promoted at rates higher than those for other groups from 1987 to 1991. These data indicate that the currently configured enlisted promotion system contains substantially disproportionate promotion rates for Black males and women. A focus on the rating factors presently included in the EPMS and the procedures board members employ in integrating this information would be useful for the purpose of evaluating the fairness of the EPMS for all subgroups.

Even though thorough analysis of six SDT versions does not indicate that it contains biased item content (e.g., the content is appropriate and items are clearly and simply worded), the merit of including the SDT within EPMS hinges on two factors. First, is differential assignment of subgroup members to various MOS assignments occurring? Eleven of twenty SMEs for 88M(2) and 71M(2) suggested it is. If it is occurring, is it based on the requirements of the assignment (e.g., physical strength, combat involvement)? If there is no performance or statutory basis for differential assignments then the use of the SDT may unfairly hinder the promotion opportunities of minority subgroups. However, if minority subgroups are not assigned to duty positions which teach the content tested on the SDT, then the assignment procedures are not fair.

Second, is SDT performance related to future role performance? If it is predictive then it should be used. If it is not predictive, then it should not be used regardless of its fairness. Related to the predictive power of the SDT is the mechanism by which it predicts. If it predicts because the SDT score reflects a motivated individual capable of learning the information associated with their duty position, then an individual whose duty position is not well reflected in the SDT content will be at an unfair disadvantage. If promoted, these individuals would perform well in their future roles. However, if the predictive power of the SDT arises out of specific knowledge targeted on the SDT, then the SDT is fair. What is unfair is that certain subgroups are assigned to duty positions which may not prepare them for advanced Army roles.

Since predictive studies of the SDT have not been conducted it is not possible to know whether it is predictive or the mechanism by which it is predictive. This applies to other components of the EPMS as well. Previous analyses with the Skill Qualification Test (SQT) revealed that it was strongly predictive of future performance (Arabian & Mason, 1986). It is, on this basis, likely that the SDT, being a similar test, would also be predictive of performance in future Army roles. It is with this in mind that the SDT is recommended for use in the EPMS on a trial basis in order to evaluate subgroup performance under operational conditions and to quantify the predictive value of the SDT. As a condition to this recommendation, MOS whose SDT versions show significant subgroup score differences should have their assignment procedures evaluated. Assignment procedures which disproportionately affect the distribution of subgroup members across duty positions should be changed.

# References


Arabian, J. M., & Mason, J. K. (1986, November). Relationship of SQT scores to Project A measures. Paper presented at the meeting of the Military Testing Association, New London, CT.

Bentz, V. J. (1988). Comments on papers concerning fairness in employment testing. Journal of Vocational Behavior, 33, 388-397.

Campbell, J. P. (1990). An overview of the Army selection and classification project (Project A). Personnel Psychology, 43, 231-239.

Campbell, J. T., Crooks, L. A., Mahoney, M. H., & Rock, D. A. (1973). An investigation of sources of bias in the prediction of job performance: A six year study. (Final Project Rep. No. PR-73-37). Princeton, NJ: Educational Testing Service.

Ford, J. K., Kraiger, K., & Schechtman, S. L. (1986). Study of race effects in objective indices and subjective evaluations of performance: A meta-analysis of performance criteria. Psychological Bulletin, 99, 330-337.

Hills, J. R. (1989). Screening for potentially biased items in testing programs. Educational Measurement: Issues and Practice, 5-11.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. Psychological Bulletin, 107, 139-155.

Hyde, J. S., & Linn, M. C. (188). Gender differences in verbal ability. Psychological Bulletin, 104, 53-69.

Jensen, A. R. (1980). Bias in mental testing. New York: Free Press.

Kimball, M. M. (1989). A new perspective on women's math achievement. Psychological Bulletin, 105, 198-214.

Kraiger, K., & Ford, J. K. (1985). A meta-analysis of ratee race effects in performance ratings. Journal of Applied Psychology, 70, 56-65.

Robinson, C. A., & Prevette, S. S. (1992). Disparities in Minority Promotion Rates: A Total Quality Approach (Fiscal Years 1987-1991). Defense Equal Opportunity Management Institute (Directorate of Research) Research Series Pamphlet 92-5.