NAVAL POSTGRADUATE SCHOOL
MONTEREY, CALIFORNIA

THESIS

AVIATION SELECTION TESTING: THE EFFECT OF MINIMUM SCORES ON MINORITIES

by

Brian J. Dean

March 1996

Thesis Co-Advisors: Mark J. Eitelberg
Anthony P. Ciavarelli

Approved for public release; distribution is unlimited.

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE March 1996	3. REPORT TYPE AND DATES COVERED Master's Thesis

4. TITLE AND SUBTITLE AVIATION SELECTION TESTING: THE EFFECT OF MINIMUM SCORES ON MINORITIES	5. FUNDING NUMBERS
6. AUTHOR Brian J. Dean	
7. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Postgraduate School Monterey CA 93943-5000	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

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

13. ABSTRACT

The purpose of this study is to examine the effects of Aviation Selection Test Battery (ASTB) cutoff scores on racial/ethnic minority applicants to naval aviation. The data were obtained from the Naval Aerospace and Operational Medical Institute in Pensacola, Florida. The data consist of test scores and performance measures for student pilots from 1988 through 1994, including pilots who were selected by both the 1992 ASTB and the previous version of the selection test. The study simulates the effect of a higher cutoff score on the "Old Test" portion of the data, then relates the findings to what may be occurring under present conditions. The results show that the "selected" pilots performed at a higher level, but the representation of minority groups declined markedly. The "deselected" pilots performed at a lower level and experienced higher attrition. The implication is that the relatively high cutoff score used by the Marine Corps may be improving the overall performance of selected pilots, but it may also be eliminating minority candidates at disproportionate rates. Further study of several options is recommended, including the following: additional selection procedures, intensified recruiting efforts, the use of selective waivers, and adverse impact analysis.

14. SUBJECT TERMS Aviation, Selection, Testing, Minorities	15. NUMBER OF PAGES 59
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

NSN 7540-01-280-5500 Standard Form 298 (Rev. 2-89)

Approved for public release; distribution is unlimited.

AVIATION SELECTION TESTING: THE EFFECT OF MINIMUM SCORES ON MINORITIES

Brian J. Dean
Captain, United States Marine Corps
B.S., University of Notre Dame, 1987

Submitted in partial fulfillment
of the requirements for the degree of

MASTER OF SCIENCE IN MANAGEMENT

from the

NAVAL POSTGRADUATE SCHOOL
March 1996

Author: Brian J. Dean

Approved by: Mark J. Eitelberg, Thesis Co-Advisor

Anthony P. Ciavarelli, Thesis Co-Advisor

Reuben T. Harris, Chairman
Department of Systems Management

ABSTRACT

The purpose of this study is to examine the effects of Aviation Selection Test Battery (ASTB) cutoff scores on racial/ethnic minority applicants to naval aviation. The data were obtained from the Naval Aerospace and Operational Medical Institute in Pensacola, Florida. The data consist of test scores and performance measures for student pilots from 1988 through 1994, including pilots who were selected by both the 1992 ASTB and the previous version of the selection test. The study simulates the effect of a higher cutoff score on the "Old Test" portion of the data, then relates the findings to what may be occurring under present conditions. The results show that the "selected" pilots performed at a higher level, but the representation of minority groups declined markedly. The "deselected" pilots performed at a lower level and experienced higher attrition. The implication is that the relatively high cutoff score used by the Marine Corps may be improving the overall performance of selected pilots, but it may also be eliminating minority candidates at disproportionate rates. Further study of several options is recommended, including the following: additional selection procedures, intensified recruiting efforts, the use of selective waivers, and adverse impact analysis.

TABLE OF CONTENTS

LIST OF TABLES

I. INTRODUCTION

The training of naval aviators is a serious business. Flight school is a challenging, arduous experience designed to train young officers to safely and effectively operate aircraft under the most demanding conditions. Navy and Marine Corps aircraft operate worldwide, day and night, at sea and over land, in desert heat and in arctic cold, in river basins and tens of thousands of feet in the air. The Navy and the Marine Corps place many demands on their pilots that are largely unique to the seagoing services, such as shipboard operations and aviation support of amphibious operations. The selection and training processes for naval aviators reflect those exceptional requirements. The training costs are high -- hundreds of hours in the cockpit, the simulator, and the classroom are required before a flight student is ready to wear the "wings of gold." The selection procedures must therefore be stringent, but must produce enough candidates who can succeed in training to fill fleet requirements. Many potential applicants are removed from contention by the strict physical requirements, including sensory perception, athleticism, and even anthropomorphic dimensions. Beyond these, the main selection instrument used by the Navy and Marine Corps is a selection test battery.

The current version of the test battery is the 1992 Aviation Selection Test Battery, or ASTB. Researched and developed over a number of years, the 1992 ASTB has been in use since late that year. The test generated some controversy when it was first introduced. Some recruiting commands in the Navy and Marine Corps raised concerns about possible gender and ethnic bias, specifically in the portion of the battery called the Biographical Inventory. The Naval Aerospace and Operational Medical Institute in Pensacola, Florida had developed the instrument and took the lead in demonstrating the scientific validity of the test battery. The debate in the Marine Corps centered more around the cutoff score. The Marines had elected to employ a higher cutoff score than the one used by the Navy for entry into the flight program. Suggestions for change ranged from lowering or waivering the cutoff score to eliminating portions of the ASTB altogether. The main issue was that the higher cutoff might be needlessly excluding qualified candidates. Special concerns were raised about the effect of the cutoff score on the ability to recruit racial/ethnic minority members into flight training. The Marine Corps, like the other services, has been intensifying efforts to increase racial/ethnic representation in its officer ranks, and some saw this cutoff score as an unnecessary impediment to these efforts.

Nevertheless, the Marine Corps policy stayed in place. The command decision to maintain the higher cutoff score with no allowance for waivers was based on concerns for safety, as well as keeping ASTB standards in line with overall maintenance of high performance standards throughout the Marine Corps. It was still too early to determine the specific demographic effects of the higher score, and the decision was made to resolve any difficulties with the accession of minorities by intensifying recruiting efforts. That policy remains in place at the time of this writing.

This study attempts to measure the effects of the higher Marine Corps cutoff score on minority applicants to the naval aviation. The study uses a statistical simulation on older, more extensive data and relates the findings to the current situation. Although exploratory in nature, it is hoped that this study can provide a theoretical "look ahead" for the Marine Corps to help identify potential obstacles and possible courses of action as it attempts to achieve racial/ethnic diversity and maintain high performance standards in aviation.

II. LITERATURE REVIEW

A. HISTORY

The search for variables that predict success in flight training and the measurement of those variables for the purpose of selection date back to the infancy of aviation. The Navy embarked on its first major study of aviator selection procedures during World War II, when the demand for Naval Aviators was growing rapidly. Data were collected on flight school candidates from pencil-and-paper tests, psychomotor apparatus, and interviews. This research, commonly called the "Pensacola 1,000 Aviator Study," gave the Navy its first comprehensive look at the personal attributes of successful student aviators. The results suggested that psychomotor abilities, mechanical comprehension, and general intelligence were reliable predictors of training success. Basic biographical data were derived from a questionnaire that asked about family background, personal and medical history, environmental influences, education, and vocational and aeronautical interests. These data were shown to be a weak and inconsistent predictor of success,[1] although it is important to remember that this inventory was much less sophisticated in design than the inventories used today. This research contributed to the Navy's development and implementation of the Academic Qualification Test/Flight Aptitude Rating, or AQT/FAR. The AQT portion was a general intelligence instrument designed to predict performance in the academic phase of training, often called "ground school." The FAR was a composite score based on the results of a Mechanical Comprehension Test (MCT), a Spatial Apperception Test (SAT), and a Biographical Inventory (BI). The composite FAR score was designed to predict the probability of a student's success in the flight portion of

[1] McFarland, R.A. and Franzen, R., The Pensacola Study of Naval Aviators -- Final Summary Report, Rept. 38, Division of Research, Civil Aeronautics Administration, Washington, D.C., 1944.

training. This test battery, with revisions in 1953 and 1971, was used up until 1992.

In the early 1980s, the Navy began an effort to devise a new selection test for aviation candidates. In addition to new Federal guidelines concerning selection testing, the Navy was concerned about possible compromises in the AQT/FAR over the years, and had observed a decrease in the predictive validity of the test. Additional impetus for change came from changes in the demographics of the applicant population with the onset of the All-Volunteer Force, changes in aviation training (such as increased use of simulators), and new operational aircraft.[2]

In 1992, the Navy released the Aviation Selection Test Battery (ASTB) for use in the selection of aviation candidates. The ASTB was developed using the knowledge of historically valid predictors of success as well as more modern test theory to ensure fairness and compliance with Federal guidelines. It consists of five subtests: a Math Verbal Test (MVT), a Mechanical Comprehension Test (MCT), an Aviation and Nautical Information Test (A/N), a Spatial Apperception Test (SAT), and a Biographical Inventory (BI). These raw scores are weighted and combined into three final scores: the Academic Qualification Rating (AQR), the Flight Aptitude Rating (FAR), and the Biographical Inventory (BI). These three scores are the basis for selection decisions. The test battery was developed by the Naval Aerospace and Operational Medical Institute (NAMI),with Educational Testing Services of Princeton, New Jersey providing developmental technical expertise. The AQR was designed to predict academic performance, the FAR was for flight performance, and the BI for attrition. The validation was conducted on approximately 30,000 aviation candidates who had already been selected for flight training. The cross validation correlation statistics for pilots

[2] Frank, L.H. and Baisden, A.G., The 1994 Navy and Marine Corps Aviation Selection Test Battery Development, presented at the Annual Meeting of the Military Testing Association, Williamsburg, VA.

(uncorrected for restriction of range) were 0.40 for the AQR, 0.27 for the FAR, and 0.25 for the BI.[3]

B. VALIDATION AND RESPONSE BIAS

In 1947, Donald W. Fiske reviewed some of the evidence concerning the usefulness of selection tests for Naval Aviators. He found that areas such as vocabulary, direction following, and arithmetic reasoning were useful in predicting ground school failures. Mechanical comprehension was shown to be a dependable predictor of both ground school and flight failures. The biographical inventory, which consisted of 150 items on biographical topics, habits, interests, attitudes and preferences, proved to be a relatively satisfactory predictor of flight failure. Still, Fiske expressed some concerns about the testing data. First, since the biographical inventory section is a self-reported survey, it is possibly subject to "faking," where the applicant answers questions in ways that he or she deems are more likely to gain acceptance into flight school. Second, the tests were validated on a population of flight students who *had already been accepted for flight training,* which Fiske claimed would limit the ability to assess the predictive power of the test.[4] In a review of Aviator selection, North and Griffin agreed with Fiske's first concern, noting that applicants for Naval Aviation are college graduates with above-average intelligence, and could likely be effective at guessing the "correct" responses to the Biographical Inventory.[5]

Still, these potential problems do not invalidate this sort of selection testing. Robert Thorndike referred to the "restriction of range" problem, where the validity of a selection instrument is measured on a non-random group, specifically on those

[3] Frank, L.H. and Baisden, A.G., The 1994 Navy and Marine Corps Aviation Selection Test Battery Development, presented at the Annual Meeting of the Military Testing Association, Williamsburg, VA.

[4] Fiske, D.W., "Validation of Naval Aviation Cadet Selection Tests Against Training Criteria," Journal of Applied Psychology 31, (December 1947): 601-613.

[5] North, R.A. and Griffin, G.R., "Aviator Selection 1919-1977," Naval Medical Research Laboratory, Pensacola, Florida, October 1977.

who have been selected. He reviews an unusual study where an experimental group of aviation candidates were given a selection test and then admitted to training regardless of their score. Validity statistics were compiled for the group as a whole and were compared with statistics based only on those who "passed" the test. The results showed a substantial decrease in the validity coefficients when they were calculated on the basis of the restricted group's training outcomes. The implication is that the predictive power of a test may be substantially *understated*[6] if it has been validated on a group that has already been selected (as is the case with the ASTB). Arthur Jensen supported this hypothesis, adding that the underestimation of validity becomes more severe as the selection becomes more stringent, perhaps even reducing the coefficient to zero, depending on the degree of selectivity and the strength of the correlation between test scores and the criterion.[7] (Statistical techniques are available that can correct for this restriction of range and allow estimation of the "true" validity coefficients, but for the purposes of this study it is sufficient to note that the uncorrected coefficients are likely to be biased downward, understating the predictive power of the test.)

The issue of response bias or "faking" is always a concern in a self-reported inventory such as the BI, since such inventories depend on honest answers from the test-taker. Research by Philippe Thiriart showed that people who take personality tests are more willing to accept socially desirable statements about themselves than statements that may be more scientifically accurate.[8] This finding is supported by Merydith and Wallbrown in examining systems for understanding how people may systematically distort their answers to personality inventories.[9] In

[6] Thorndike, R.L., Personnel Selection, (New York: John Wiley and Sons, 1949), 170-171.

[7] Jensen, A.R., Bias in Mental Testing, (New York: The Free Press, 1980), 311-312.

[8] Thiriart, P., "Acceptance of Personality Test Results," Skeptical Inquirer 15, (Winter 1991): 161-165.

[9] Merydith, S.P., and Wallbrown, F.H., "Reconsidering Response Sets, Test-taking Attitudes, Dissimulation, Self-deception, and Social Desirability," Psychological Reports 69, (December 1991): 891-905.

addition to these natural inclinations toward social desirability, aviation candidates have another potential motivator to bias their responses: *they are trying to get into the program.* The North and Griffin study cited above suggested that aviation candidates may be bright enough to guess the responses that lead to higher scores on the BI. An experiment by David Peltier and James Walsh asked college students to "fake" personality traits on an inventory designed specifically to prevent response bias by masking the "correct" responses. The results showed that the subjects (a population similar to aviation candidates) were able to successfully feign either the existence or non-existence of the personality traits.[10] Power and McRae reported similar results with the use of the Eysenck Personality Inventory.[11] Leary and Kowalski suggest that in order to effectively "fake" an inventory, candidates must be both able and motivated to do so[12], and aviation candidates appear to have both "motive" and "means."

The Navy conducted an experiment related to this possibility on flight students at Pensacola. Researchers gave the California Psychological Inventory (CPI) to incoming Aviation Officer Candidates (AOCs) and to flight students who had voluntarily quit the program (DORs, or Dropped On Request). Both groups were further divided and given the test under two different sets of instructions: One set of instructions asked for honest self-appraisal, and the other asked the subjects to respond "as they would like to be." Under normal instructions, incoming AOCs and DORs obtained almost identical scores. Under "ideal" instructions, however, the incoming AOCs obtained significant elevations on 11 of

[10] Peltier, B.D. and Walsh, J.A., "An Investigation of Response Bias in the Chapman Scales," Educational and Psychological Measurement 50, (Winter 1990): 803-815.

[11] Power, R. and McRae, K., "Characteristics of Items in the Eysenck Personality Inventory Which Affect Responses When Students Simulate," British Journal of Psychology 68, (1977): 491-498.

[12] Leary, M. and Kowalski, R., "Impression Management: A Literature Review and Two-component Model," Psychological Bulletin 107, (1990): 34-47.

the 18 measured scales.[13] This study was focused on analyzing the CPI for potential use in predicting DORs, which is one kind of attrition. In terms of this thesis, combining the results of the CPI study with the work of Thiriart and of Merydith and Wallbrown suggests that candidates taking the ASTB may answer questions with more of an "ideal" mindset, which could actually add to the predictive power of the test.

It is worthwhile to note, at this point, that the Biographical Inventory currently used by the Navy and Marine Corps as part of the ASTB contains some elements of personality assessment, but is <u>not</u>, strictly speaking, a personality inventory. The personality measurement focuses on attributes such as emotional stability, a historically significant predictor of aviator success that explains a unique portion of variation in student attrition.[14] The other focus of the inventory is more on life experience and past behaviors, rather than on specific traits. Still, the applicant may subject to the same response bias when taking the BI.

Consider the following example. An applicant reads the following question:

"Have you ever skied on anything other than a beginner's slope?"

a) Yes
b) No

Now suppose that this applicant lived in a southern area where snow skiing was unavailable, such as Florida. The applicant feels confident that he or she would have skied regularly and progressed to the most challenging slopes if snow skiing

[13] Bucky, S.F., "The California Psychological Inventory Given to Incoming AOC's and DOR's With Normal and 'Ideal' Instructions," 1971, Naval Aviation Medical Research Laboratory Report 1127, Pensacola, Florida.

[14] Cattell, R., Eber, H., and Tatsuoka, M., <u>Handbook for the Sixteen Personality Factor Questionnaire,</u> (Champaign, IL: Institute for Personality and Ability Testing), 1990; and Luk'yanova, N., "Personality Characteristics of Pilot-cadets With Different Marks in Flight Disciplines," (Charlottesville, VA: U.S. Army Foreign Service and Technology Center), 1977; and Fleischman, H., Ambler, R., Peterson, F., and Lane, N., "The Relationships of Five Personality Scales to Success in Naval Aviation Training," NAMI-968, (Pensacola, FL: Naval Aerospace Medical Institute), 1966.

areas had been accessible. The applicant may well answer "Yes" to the question, since he or she would have taken the more difficult slopes, had the opportunity arisen. Given the choices, the applicant may well judge that a "Yes" response, while not technically accurate, is a better reflection of his or her attitudes and interests. In this example, the candidate's judgment is correct, and the Navy actually gets more accurate data and can make a better prediction about the applicant's potential for success. Some recent research in this type of "faking" does not appear to be a major source of distortion for job applicants,[15] and does not undermine the predictive validity of the instrument.[16]

C. ETHNIC DIFFERENCES AND "FAIRNESS"

Setting aside the issue of the Biographical Inventory for the moment, the history of aviation selection testing from the Pensacola 1000 to the present has left little doubt about the usefulness of testing for mechanical comprehension, general intelligence, direction following, and reasoning skills for the screening of candidates. These skills have been shown to be sound predictors of both academic and flight performance by the Navy[17], the Army[18], and the Air Force[19]. An issue of some concern, however, relates to differences in test scores between the genders and racial/ethnic groups. As organizations attempt to diversify, concerns about fairness in selection test construction and standard-setting have

---

[15] Hough, E., Eaton, N., Dunnette, M., Kamp, J., and McCloy, R., "Criterion-related Validities of Personality Constructs and the Effect of Response Distortion on Those Validities," Journal of Applied Psychology Monograph 75, (1990): 581-595.

[16] Cunningham, M., Wong, D., and Barbee, A., "Self-presentation Dynamics on Overt Integrity Tests: Experimental Studies of the Reid Report," Journal of Applied Psychology 79, (1994): 643-658.

[17] Examiners Manual and Scoring Instructions, U. S. Navy and Marine Corps Aviation Selection Tests, NAVMED P-5098 (1971), Aerospace Operational Psychology Branch, Bureau of Medicine and Surgery, Navy Department, Washington, D. C.

[18] Kaplan, H., "Prediction of Success on Army Aviation Training," Technical Research Report 1142, U.S. Army Personnel Research Office, OCRD, 1965.

[19] Miller, R. E., "Interpretation and Utilization of Scores on the AFOQT-AFHRL-TR-69-103", Personnel Research Laboratory, Lackland AFB, Texas, 1969.

moved closer to center stage. Current guidelines from the Equal Employment Opportunity Commission (EEOC) place requirements on employers who use selection tests to ensure that they do not unfairly discriminate against demographic groups.

Here, it should be noted that there is a distinction between "discrimination" and "unfair discrimination" in the legal sense. Selection tests are specifically *designed* to discriminate, on the basis of the stated criterion (usually some measure of job performance). A test that failed to discriminate on the basis of the criterion between groups of people with a given set of attributes would be useless as a selection instrument, since it would be unable to predict job performance based on those variables (i.e., test scores). Additionally, aggregate differences in abilities and interests between certain groups are well documented[20], and are usually observed in any accurately measured variables.[21]

These aggregate differences and their impact on the fairness of hiring and promotion practices have been the subject of legal debate for many years. A landmark case in this area was heard at the United States Supreme Court in 1971. In *Griggs v. Duke Power*, the Court laid out the basic criteria for the legal claim that a particular employment practice (like a selection test), has a disparate impact on members of a protected demographic group. The Court recognized that a

[20] Anastasi, A., Differential Psychology, McMillan, New York, 1965; and Neiner, A. , "Examples of Testing Programs in the Insurance Industry," in Test Policy and the Politics of Opportunity Allocation: The Workplace and the Law, ed. Gifford, B., (Norwell: Kluwer Academic, 1989); and Dreger, R. M., "Comparative Psychological Studies of Negroes and Whites in the United States: 1959-1965," Psychological Bulletin 75, (1968): 261-269; and Wing, H. "Profiles of Cognitive Ability of Different Racial Ethnic and Sex Groups on a Multiple Abilities Test Battery," Journal of Applied Psychology 3, (1980): 289-298; and U.S. Department of Defense. Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs and Logistics). 1982. Profile of American Youth: 1980 Nationwide Administration of the Armed Services Vocational Aptitude Battery. [Washington D.C.]: U.S. Department of Defense, Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs and Logistics), 30-36.

[21] Arvey, R. D. and Faley, R. H., Fairness in Selecting Employees, 2nd edition, p. 122, Addison-Wesley, Reading, Massachusetts, 1988.

practice may not have been specifically *designed* to discriminate unfairly, but may in practice needlessly exclude certain groups of applicants.

> ...good intent or absence of discriminatory intent does not redeem employment procedures or testing mechanisms that operate as "built-in headwinds" for minority groups and are *unrelated to measuring job capability*.[22] (emphasis added)

Although the court placed the burden on the employer to demonstrate that a selection test is related to job performance, it affirmed the usefulness and equity of such a test even in the face of a disparate impact. Nevertheless, the decision provided little guidance as to *how much* of a difference in selection rates between groups is sufficient to demonstrate that disparate impact (or, as it is commonly called, "adverse impact") exists. The EEOC, the Civil Service Commission, and the Department of Labor provided some clarity in 1978 with the joint publication of the Uniform Guidelines on Employee Selection Procedures:

> A selection rate for any racial, ethnic, or sex subgroup which is less than four-fifths (4/5) (or 80 percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact...[23]

This guideline, commonly called the "four-fifths rule," is not presented as a specific requirement, but as a "rule of thumb" for the establishment of a *prima facie* case of adverse impact. The Uniform Guidelines caution against the strict adherence to this rule when sample sizes are small. Also, they recognize that some groups may, on average, not possess attributes (usually physical) that are closely related to job performance. Therefore, the existence of adverse impact under the "four-fifths" rule alone is not grounds for discontinuance of a selection test. If the test scores show significant correlation with job performance variables and no

---

[22] *Griggs v. Duke Power*, 3 FEP 175, (1971).

[23] "1978 Uniform Guidelines on Employee Selection Procedures" Section 4, pp. D.

specific discriminatory intent exists, the test is likely a valid and legally defensible instrument.

Still, the debate over what constitutes a "fair" test persists, since selection procedures can be set up so many different ways. Perhaps the most commonly used assessment of fairness comes from T. A. Cleary, who suggests that a test is fair if regression lines for population subgroups do not differ.[24] Cole suggests that separate cutoff scores should be set for minority and majority groups so that qualified members of each group have an equal likelihood of being selected.[25] A similar model suggested by Einhorn and Bass uses the concept of "equal risk," using separate regression equations and separate acceptance scores for subgroups that will equalize the probability of success on the job.[26] Darlington suggests, as one option, adding a premium (equal to one-half standard deviation) to the predicted performance of a minority group applicant, which would represent the value the organization places on the selection of minority candidates. The organization then selects candidates with the highest predicted performance.[27] Thorndike developed a combination approach, where separate regression equations are used for subgroups, and then cutoff scores are set so that the selection ratio and success ratio between majority and minority groups are equal.[28] Other than the Cleary model, these approaches to fairness all can allow the use of different cutoff scores for different groups. Selection processes like these were affected by the passage of the Civil Rights Act of 1991, which states in Section 106:

---

[24] Cleary, T.A., "Test Bias: Prediction of grades of Negro and white students in integrated colleges," Journal of Educational Measurement 5, (1968): 115-124.

[25] Cole, N.S., "Bias in Selection," ACT Research Report 51. Iowa City, Iowa: American College Testing Program, 1972.

[26] Einhorn, H.J. and Bass, A.R. "Methodological Considerations Relevant to Discrimination in Employment Testing," Psychological Bulletin 75, (1971): 261-269.

[27] Darlington, R.B., "Another Look at Culture Fairness," Journal of Educational Measurement 8, (1971): 71-82.

[28] Thorndike, R.L., "Concepts of Culture Fairness," Journal of Educational Measurement 8, (1971): 63-70.

> It shall be an unlawful employment practice for a respondent, in connection with the selection or referral of applicants or candidates for employment or promotion, to adjust the scores of, use different cutoff scores for, or otherwise alter the results of, employment related tests on the basis of race, color, religion, sex, or national origin.[29]

This prohibition of discriminatory use of test scores places the emphasis for fairness on the instrument itself, moving the Cleary model for fairness to the forefront. The American Psychological Association specifically endorses the use of the Cleary model,[30] and it was also used as a test of fairness for the ASTB. Test developers found no evidence that the ASTB had differential regression lines for population subgroups, meaning that it does not overpredict or underpredict performance of any group relative to another.

D. CUTOFF SCORES

Although the Navy and Marine Corps use the same test, they differ on the minimum required scores. The Navy established a cutoff score of 3/4/4 for the AQR, the FAR, and the BI, respectively. Some waivers for lower scores are allowable in cases of otherwise exceptionally qualified candidates. The Marine Corps uses a 4/6/4 standard, but decided not to allow waivers. The rationale for the higher score and the decision not to allow waivers was that the ASTB was designed to help minimize attrition by better predicting flight school success, and allowing waivers may negate these cost-saving benefits, as well as possibly reducing the quality of aviation students. Additionally, the 4/6/4 standard was selecting enough candidates to fill available training seats. The Uniform Guidelines state that while the use of cutoff scores (where candidates scoring below the standard have little or no chance for selection) may be appropriate, the

---

[29] Civil Rights Act of 1991, Statutes at Large, 105, sec.106, 1075 (1991).

[30] American Psychological Association, Standards for Educational and Psychological Testing, Washington, D.C. (1985): APA.

degree of adverse impact that results should be a consideration in the establishment of the minimum score.[31] Given the existence of aggregate differences between population subgroups, it becomes evident that the location of the cutoff score can have a large impact on the selection rates for the groups and the resulting degree of adverse impact. Consider the graph of the test shown in Figure 1:

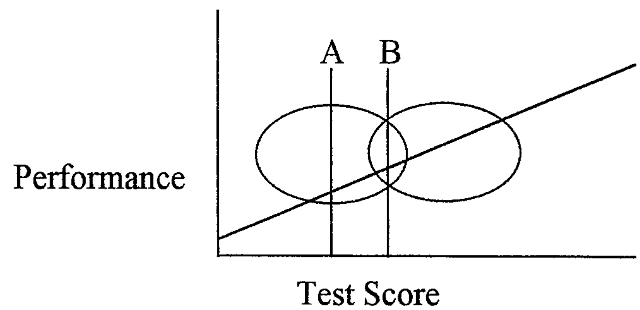

Figure 1. Cleary Model "Fair" Test With Two Cutoff Scores

Figure 1 shows a fair test under the Cleary model. Regression lines for the two groups do not differ, although the distributions of scores and performance measures are not the same. An organization considering moving the cutoff score from A to B can note two consequences. First, the aggregate performance measure for the group they select (those whose test scores are to the right of the line) will likely increase. They can expect the performance increase to have a reasonably linear relationship with the cutoff score increase. After all, the test was designed specifically to measure that relationship between test score and job performance. Secondly, the organization must note that the increase in cutoff score may have a more marked effect on the proportion of the two groups in the selected population, and this relationship my be largely non-linear. In Figure 1, the use of cutoff score A appears to achieve a population that has roughly one-third from the lower scoring group. Raising the cutoff score to B would yield a selected population

[31] "1978 Uniform Guidelines on Employee Selection Procedures," Section 5, pp. H.

with almost no representation of the lower scoring group. The organization would have to weigh the benefit of the higher performance against the potential cost of the reduced representation.

The use of cutoff scores is still the subject of much legal debate, but the key issue seems to remain focused on the validity of the instrument. The Equal Employment Advisory Committee (EEAC) has proposed that the next revision of the Uniform Guidelines include language that allows employers to set cutoff scores as high or as low as they please, so long as the test is a demonstrably valid instrument.[32]

Of course, cutoff scores are only one option in the scoring of a test battery such as the ASTB. Robert Thorndike discusses the use of multiple regression, where a single aptitude score is derived from the weighted sum of the subtests, in this case the AQR/FAR/BI scores. Thorndike suggests that the multiple regression method will yield better criterion performance (i.e., better flight students) than the use of multiple cutoff scores so long as the test scores maintain a reasonably linear relationship with the criterion.[33] To a degree, this is already done in the ASTB, since the AQR and FAR are weighted combinations of other subtests. The problem with further weighting and combining is that the AQR, FAR and BI each measure distinct abilities and attributes that relate to performance. As a result, a higher AQR score, for example, cannot "make up" for a lower BI score in terms of performance. The BI in particular, which has had extensive validation review, has been shown to account for a unique portion of variation in performance.[34] Since candidates need to have a certain measure of each attribute, the use of multiple cutoff scores here appears to be appropriate.

---

[32] "Division 14 Principles," Equal Employment Advisory Committee, 1980.

[33] Thorndike, R.L., Personnel Selection, (New York: John Wiley and Sons, 1949) 186-198.

[34] Frank, L.H., "Biographical Inventory Validation Assessment," Naval Aerospace and Operational Medical Institute, Pensacola, FL, May 1994.

Overall, there appears to be a strong historical consensus that the variables measured by the ASTB, including intelligence, mechanical comprehension, and biographical data have a strong relationship with flight school performance. The validation of the ASTB suggested that it is a useful and equitable instrument for the screening and selection of flight school candidates. The issue of cutoff scores and their relationship to the demographic distributions of applicant scores, however, is a worthwhile area of study. This is especially true as the Marine Corps continues to seek population diversity in its Officer corps while maintaining high performance standards.

Therefore, the focus of this study is to examine the population of pilot candidates that the Marine Corps is "de-selecting" and the Navy is "selecting" as a result of the different ASTB standards used by the two services. A better understanding of the effects of the different cutoff scores on the demographic mix of the selected populations can help the Marine Corps in policy decisions in the area of aviator selection.

III. METHODOLOGY

A. DATA

The data for this study was obtained from the Naval Aerospace and Operational Medical Institute in Pensacola, Florida. The data originally contained almost 6,000 observations of students who were admitted for training at Naval Aviation Schools Command in Pensacola from 1988 through 1994.

Since the focus of this study is on pilots, all applicants for the Naval Flight Officer (NFO) program were removed from the file. Although the NFO candidates take the same version of the selection test, the training they receive is too different from pilot training to allow those observations to remain. Foreign students were removed, so that the study could concentrate on United States forces. Candidates from other services, such as the United States Coast Guard, were kept in the sample, since they go through the same training and are drawn from a similar population as are Navy and Marine Corps candidates.

These restrictions left a sample of 3,800 pilots. The observations were then divided in to two data files -- "New Test" and "Old Test" -- for the purpose of this study.

1. "New Test" Data

This data file contains observations on pilots who were selected for flight training under the 1992 ASTB, and had completed primary flight training. At the time these data were obtained, 59 pilots had progressed through the primary flight stage. Although the data are the most current available, the 1992 ASTB was not released for use until late in that year. Additionally, candidates who take the test frequently do so while still in college. As a result, delays of two years or more are not unusual between the test date and the date of entry into flight school. The observed variables in this file include ASTB scores, primary flight grades, flight school academic grades, and demographic variables.

2. "Old Test" Data

This file contains observations on pilots who were selected under the pre-1992 selection test battery. There are approximately 3,700 observations in this file, including all pilots who started flight training from 1988 to 1992.[1] These pilots have all either completed training and earned their wings or have attrited from the program. Consequently, these data are much more conducive to detailed analysis than are the "New Test" data.[2] The observed variables included are much the same as those in the "New Test" data, except, of course, that the selection test scores are from the older version.

In this study, the key variables are race and flight grades, so it is important to precisely define these variables as they exist in the data and as they are used in the analysis.

3. Race

The race variable in the data takes different values for a number of different racial and ethnic groups. In this study, four groups are defined: White (Caucasian), Black (African-American), Hispanic, and Asian (including Pacific Island regions). Other groups, such as Native Americans, were identifiable in the data but were not singled out for analysis because the small numbers of observations in these categories would make any meaningful statistical analysis difficult to interpret. However, these observations are included in the analysis

---

[1] Some observations, especially older ones, had values or codes on a relevant variable that appeared unreliable (i.e. a flight grade that was outside the range of possible grades, or a race code that did not exist). These observations were excluded from the portion of the analysis pertaining to that variable, but were included in other areas where their values were reliable. As a result, the divisions of the data may not always sum to 3700.

[2] While accurately collected and coded, the "New Test" data simply have too few observations for meaningful analysis.

when minorities as a group are compared to the majority group. Table 1 shows the racial/ethnic composition of the "Old Test" data.

Table 1. Racial/Ethnic Composition of "Old Test" Data

Group	Frequency	Percent	Cumulative Percent
Asian	57	1.5	1.5
Black	119	3.3	4.8
Hispanic	78	2.2	7.0
White	3322	92.0	99
Other	31	0.8	99.8
Total[3]	3607	100.0	100.0

Source: Derived from data obtained from Naval Aerospace and Operational Institute.

4. Flight Grades

A crucial factor in an analysis of this kind involves the selection of a performance measure. Although there are several measures of flight school performance available in the data, primary flight grades were chosen for this study for the following reasons: First, and most important, primary flight grades are common to the two data files, "New Test" and "Old Test." The primary flight syllabus is the same for the two groups, and the grading criteria are also the same. The aircraft used for both groups is also the same, the T-34. Certainly the flight instructors themselves are different, as rotation schedules move personnel around the Navy and Marine Corps, but the high degree of standardization of flight procedures that drives the student learning suggests that it is simply a matter of different instructors teaching the same things. Teaching techniques certainly can differ between instructors, and these techniques may affect the grades of the students. Still, to introduce bias these differences would have to be systematic between the "New Test" and "Old Test" periods, and there is no compelling evidence to suggest that is the case. Another possibility is "grade inflation," or the

[3] Some percentage values have been rounded or truncated, and may not sum exactly to 100.

tendency over time for average grades to rise while actual student performance remains stable. This possibility is made less likely by the built-in objectivity of primary flight grades. Although the grades are assigned based on the judgment of the flight instructors, there are guidelines for instructors to follow in evaluating student performance. For example, a student who is graded as "Above Average" on a particular maneuver (a turn pattern, for example) can be assumed to have performed the maneuver within certain objective parameters (plus or minus twenty-five feet of altitude and plus or minus ten knots of airspeed, for example). This assumption of commonality of primary flight grades is essential, since this variable is the basis for comparing the two groups in this study. Many of the other performance variables (such as academic performance), while comparable in scale of measurement, differ in syllabus and are therefore not useful for this analysis. The second reason for the selection of flight grades is that primary flight performance is one of the performance measures that the selection tests (both "New "and "Old") were designed to predict. This makes the grades relevant to the discussion of the selection instrument. Third, as discussed in detail above, primary flight grades are a reasonably objective measure, reducing the likelihood of bias.

Numerically, primary flight grades can be thought of much like a Grade Point Average or GPA for primary training, with a range of one to four. On every flight, students are graded on a series of maneuvers, as well as attributes such as procedural knowledge and headwork. The four possible grades, as well as their numerical value, are listed in Table 2.

Table 2. Primary Flight Grades and Their Corresponding Numerical Values

Grade	Numerical Value
Above Average	4.0
Average	3.0
Below Average	2.0
Unsatisfactory	1.0

Source: Naval Aviation Schools Command, Pensacola, Florida.

The distribution of grades tend to be clustered close to 3.0, since most of the items on a particular flight will be graded as "Average." The students are expected to progress in skill such that a particular level of performance on a maneuver that is deemed "Above Average" on one flight might well be considered "Average" on the following flight. A student who receives three "Above Average" marks on a flight with twenty graded items would be considered very successful that day, and would likely leave the base with what some instructors call a "three-above smile."

B. PROCEDURES

This study analyzes the group of pilots who fall between the Navy and Marine Corps cutoff scores. The obvious method would be to simply look at that population as it exists today. However, the limited number of observations in the "New Test" data precludes, for the time being, any meaningful analysis of the racial/ethnic composition of that group. As an alternative, since the "Old Test" data are much more extensive, this study poses the following question: What would have happened if the higher cutoff score of today had been used in 1988? Certainly, it would have yielded a selected population with, on average, higher criterion scores. The older version of the test had seen a decrease in predictive validity, but it was still a useful instrument. Also, it is possible that a higher cutoff score might have significantly altered the racial/ethnic mix of the student population.

The next logical issue in a simulation such as this is to decide where to place the simulated cutoff score. To be useful, the cutoff score must be placed where it would have the same effect on the selected population as the higher cutoff score used by the Marine Corps today. One possibility might be to simply numerically set the simulated cutoff score in the "Old Test" data to match the difference between the two cutoff scores used today. However, this is not appropriate here, since the new and old test scores are not comparable. The procedures for weighting and combining the raw subtest scores were changed when the test was rewritten. A student who scored a "5" on a particular section of the old test could not be assumed to score the same on the new test.[4] Although raw subtest scores are available, and could possibly be recombined and scored under the new procedures, the test items themselves were changed enough that the comparability of new and old subtest scores becomes questionable.

So, although it is not possible to simulate the higher cutoff on the basis of the test scores themselves, it is possible to do so on the basis of criterion performance. Given that the previous version of the test was still valid, changing the cutoff score will change the aggregate performance of the selected population. Moreover, there must be *some* test score on the "Old Test" such that, had it been the actual cutoff, a population of pilots would have been selected with the same criterion performance as exists under the 1992 ASTB. How do we find that cutoff score? Quite simply, we do not need to. Numerically, it would have little meaning in itself. It will likely be some fraction of a score, which is not actually achievable by any individual test-taker since the scoring procedures yield only whole numbers. Again, as a numerical value, it is functionally irrelevant. All that matters is that its use would have yielded a selected population of "Old Test" pilots whose performance matches that of the "New Test" pilots. So, we just need remember that the score exists, and that it is different (likely higher) than the actual cutoff, as long as the actual "New Test" and "Old Test" criterion scores differ. Since consistent performance data are available, this methodology seeks to establish a *performance-based* simulated cutoff score since a test-based simulated cutoff score is not practicable.

---

[4] For example, the Flight Aptitude Rating (FAR) on the old test encompassed the Biographical Inventory (BI). On the 1992 ASTB, the BI score stands alone as a separate score.

The central issue, then, is the matching of the performance index, primary flight grades. The first step is to examine the grade distribution of the "New Test" and "Old Test" pilots. They are listed in Table 3.

Table 3. Primary Flight Grades: "Old Test" vs. "New Test"

Test	Mean (μ)	Standard Deviation (σ)
"Old Test"	3.055	.045
"New Test"	3.084	.037

Source: Derived from data obtained from Naval Aerospace and and Operational Medical Institute.

Since the "New Test" pilots' mean flight grades are significantly higher (t = 6.02), the simulated performance cutoff will be higher than the minimum performance achieved under the actual "Old Test" cutoff score.[5] Therefore, the simulated cutoff must "deselect" the lower portion of the performance distribution such that the "selected" group will have a performance mean that matches that of the "New Test" pilots. As it turns out, using a simulated cutoff score one standard deviation below the mean ($\mu - 1\sigma$) yields a mean performance for the selected group of 3.08, which matches the "New Test" pilots mean performance. The numeric value of ($\mu - 1\sigma$) is 3.047. What remains, then, is to examine these two groups of "selected" and "deselected" pilots to determine the effect this cutoff score would have had on the racial/ethnic mix of the student population.

[5] This is to be expected, since primary flight performance is one of the predicted criteria for both the new and old tests. The improved validity of the 1992 ASTB should lead to more effective screening and a higher criteria score for the selected population.

IV. RESULTS

A. "NEW TEST" AND "OLD TEST" PRIMARY FLIGHT GRADES

As mentioned above, the first step in this analysis was to examine the primary flight grades of the "New Test" and "Old Test" pilots.

Table 4 again presents the mean primary flight grades for both groups of pilots.

Table 4. Mean Primary Flight Grades: "Old Test" vs. "New Test"

Test	Mean (μ)	Standard Deviation (σ)
"Old Test"	3.055	.045
"New Test"	3.084	.037

Source: Derived from data obtained from Naval Aerospace and Operational Medical Institute.

Assuming that the flight grades have remained consistent over the years[1], the increase in primary grades observed under the 1992 ASTB appears to be a positive sign for the selection process. It was noted earlier, however, that a higher test standard should yield a population of students that, on average, shows higher criterion performance. The question, then, is whether the increase in primary grades is attributable to a more valid selection instrument or to a higher cutoff score on the newer test.

Is the "New Test" cutoff score higher than that of the "Old Test?" The answer is somewhat unclear. We are aware that the Marine Corps "New Test" cutoff score is higher than the Navy "New Test" cutoff score, since that is the subject of this thesis. However, Marine Corps students account for a small percentage of the persons in the aviation training pipeline, so it is unlikely that

[1] Both "New Test" and "Old Test" pilots trained in the same aircraft, and used the same primary flight syllabus. The assignment of numerical grades is based on reasonably objective performance measures that are the same for the two groups. This assumption of comparability of primary grades is discussed in detail in Chapter Three.

they are significantly raising the mean. Additionally, cutoff scores are based, in part, based on certain levels of required performance in flight school as well as on the perceived ability of the applicant population as a whole. There is no reason to assume that the required level of flight school performance is different between the "New Test" and "Old Test" pilots. Also, there is little reason to assume that there is a relevant difference between the applicant population in the 1988-1992 group and the 1992-1994 group. Even though efforts to recruit minority applicants have increased over these years, and these efforts could potentially affect the applicant pool, any differences between the "New Test" and "Old Test" groups would have to be systematic and criterion-related to bias the distribution of "New Test" flight grades. Additionally, if such a bias existed, it would likely cause the "New Test" flight grades to be *understated*, since the aggregate measures of test and performance variables on minorities tend to be lower. In any event, the "New Test" data are not extensive enough to demonstrate such bias. Overall, there does not appear to be any compelling evidence, either empirical or theoretical, to suggest that the higher primary flight grades of the "New Test" pilots compared to the "Old Test" pilots are attributable to a proportionally higher cutoff score.

We are left, then, with the increase in the validity of the selection instrument itself. As discussed earlier, the 1992 ASTB validation study revealed an increase in predictive validity over the previous version. The selection effects of increasing the validity of a selection instrument are presented in Figure 2.

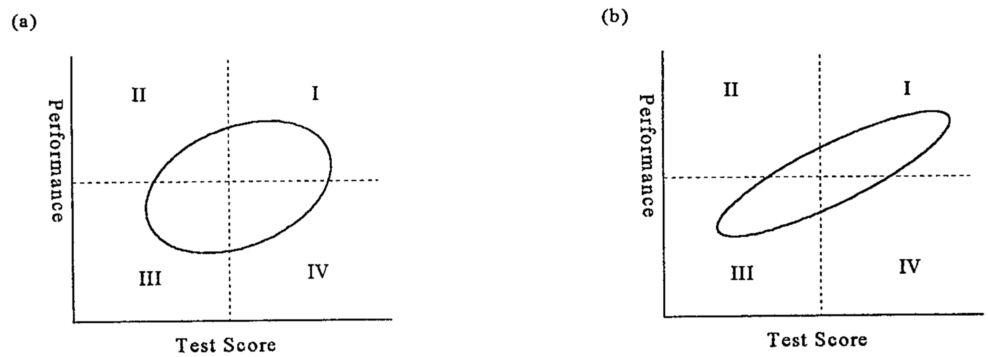

Figure 2. Two Selection Tests With Different Validities

A measure of the predictive validity of a selection test is primarily based on the degree of correlation between the test score and job performance. Figure 2 (a) and (b) are two selection tests for the same job, with test (b) having the higher validity. Note the shape of the ellipse that defines the data. The higher validity "squeezes" the distribution of test scores and performance measures in to a more linear form. The horizontal lines on each graph represent some minimum level of job performance that is deemed to be acceptable. The vertical lines represent the selection test cutoff scores. Assuming that the minimum acceptable job performance and the test cutoff score are identical for the two tests, we can see that the increased validity of test (b) will yield a higher-performing selected population. The "squeezing" of the data will reduce the number of people who pass the test but fail on the job (Quadrant IV, often called "false positives") and reduce the number of people who score below the standard on the test but would have succeeded on the job (Quadrant II, or "false negatives").[2]

Since the validation of the 1992 ASTB revealed an increase in predictive validity over the previous version, and primary flight performance is one of the criteria predicted by both the new and old test, the difference in primary flight grades between "New Test" pilots and "Old Test" pilots can reasonably be

---

[2] See Arvey, R. D. and Faley, R. H., Fairness in Selecting Employees, 2nd edition, pp. 40-43, Addison-Wesley, Reading, Massachusetts, 1988.

attributed to the increase in the effectiveness of the 1992 ASTB as a selection instrument.

B. THE EFFECT OF THE SIMULATED HIGHER CUTOFF SCORE

Once again, the central theoretical question of this study is the following: given that the 1992 ASTB data are not extensive enough for detailed analysis, what would have happened to the racial/ethnic mix of the student pilot population if the same higher cutoff score had been applied in 1988?

First, the racial/ethnic mix of the "Old Test" pilots in terms of the groups of interest, as they actually existed, are examined. The results are presented in Table 5.

Table 5. Actual Racial/Ethnic Mix of Flight Students, 1988-1992

Racial/Ethnic Group	Frequency	Percent
White	3,322	92.0
Black	119	3.3
Hispanic	78	2.2
Asian	57	1.5
All Groups	3,607	100.0

Source: Derived from data obtained from Naval Aerospace and Operational Medical Institute.

Next, a simulated, performance-based cutoff score is applied to these data and the newly "selected" and "deselected" populations are examined separately. As mentioned in Chapter Three, the use of a primary flight grade cutoff of one standard deviation below the mean (a value of 3.047) yields a "selected" population of "Old Test" pilots whose mean performance matches that of the "New Test" pilots. This "selected" population therefore consists of all pilots whose primary flight grades are greater than or equal to 3.047. The "deselected" population consists of all pilots whose primary flight grades are less than 3.047.

The two groups are examined separately and compared to the entire group of "Old Test" pilots.

C. THE "SELECTED" PILOTS

Of the original 3,607 pilots in the "Old Test" data, 2,201 were "selected" by the simulated higher cutoff score. The key question, of course, is whether the use of that score had a disproportionate effect on the minority applicants. To begin, the effect of the higher cutoff score on the percentage of minority pilots in the selected population (both real and simulated) are examined. The results are presented in Table 6.

Table 6. Percentage of Selected Minority Pilots Under Actual and Simulated Cutoff Scores in "Old Test" Data

Minority Group	Actual Cutoff	Simulated Cutoff	Percent Change
Black	3.3	1.5	-55
Hispanic	2.2	1.5	-32
Asian	1.5	1.0	-33
All Minorities	7.8	4.8	-38

Source: Derived from data obtained from Naval Aerospace and Operational Medical Institute.

These results suggest that, had the higher cutoff been used in 1988, the representation of racial/ethnic minorities among student pilots would have markedly decreased. The overall percentage of minorities (which includes the groups too small in number for separate analysis) would have decreased by 38 percent, from 7.8 to 4.8. The largest single impact is seen for Blacks, who experienced a decrease of 55 percent under the higher cutoff score. When actual and estimated numbers of pilots in these groups are examined, the effects are even more striking. The frequencies of each group are presented in Table 7.

Table 7. Numbers of Minority Pilots Selected Under Actual and Simulated Cutoff Scores

Minority Group	Actual Cutoff	Simulated Cutoff	Change	Percent Change
Black	119	32	-87	-73
Hispanic	78	34	-44	-56
Asian	57	22	-35	-61
All Minorities	285	106	-179	-63

Source: Derived from data obtained from Naval Aerospace and Operational Medical Institute.

Here, of the 285 minority pilots who were accepted for training, it is estimated that only 106 (approximately 37 percent) would have been accepted under the higher cutoff score. As seen for the percentages presented in Table 5, the largest impact is on the African-American group, where the higher cutoff score reduced the number accepted from 119 to 32. This represents almost a 75-percent decrease in the number of Black applicants accepted for training as a naval aviators.

Do these results make sense? Recall the theoretical framework presented in Chapter Two concerning the impact of raising cutoff scores. When aggregate differences in test scores and criterion measures between population subgroups exist, raising the cutoff score may have a disproportionate effect on the mix of those subgroups in the selected population. Moreover, the largest impact should be on the subgroup whose distribution is the lowest (or farthest to the left) on the regression line, since that subgroup will have the largest proportion of its distribution fall below the cutoff score. Table 8 compares the mean primary flight grades of the minority groups to the impact of the higher cutoff score on those groups.

Table 8. Minority Flight Grades vs. Impact of Higher Cutoff Score (in Ascending Order of Flight Grades)

Minority Group	Mean Flight Grade	Change in Percent Representation Under Higher Cutoff	Percent Change in Number Selected Under Higher Cutoff
Black	3.03	-55	-73
Asian	3.04	-33	-61
Hispanic	3.05	-32	-56

Source: Derived from data obtained from Naval Aerospace and Operational Medical Institute.

These results suggest, at least circumstantially, that the distributions of flight grades and the resulting impact of the higher cutoff score are behaving in accordance with the general conceptual model presented in Chapter Two. As the flight grades increase, the impact of the higher cutoff score decreases. This appears to hold true both for the number and percentage of each group in the "selected" population.

Overall, then, it appears that the simulated higher cutoff had two main effects on the "selected" population. First, the mean performance of the group increased. This, of course, was by design. Second, the representation of racial/ethnic minority groups decreased sharply both in terms of percentages and actual numbers. Also, the degree of the impact on any particular group appears to be related to the location of that group's distribution of scores, as had been suggested by the general theoretical model of a "Cleary fair" test presented in Chapter Two.

What can be said about the "deselected" group? Under normal circumstances, very little: test scores and demographics would be available, but no performance data would exist because the applicants would not have been accepted for training. Presumably, some would have been "true negatives" and some would have been "false negatives," but there is no way to tell how many or in what proportions. However, because this methodology used only a *simulated* higher cutoff score, more analysis of the "deselected" group is possible. In other words, they were "deselected" by the study but not by the Department of the Navy.

D. THE "DESELECTED" PILOTS

Of the original 3,607 pilots in the "Old Test" data, 1,406 were "deselected" by the simulated higher cutoff score.[3] The frequencies of the minority groups among "deselected" pilots are compared to the entire group of "Old Test" pilots in Table 9.

Table 9. Percentages of Minority Pilots in "Deselected" Group and Overall Group

Minority Group	Percentage	
	"Deselected"	Overall
Black	5.1	3.3
Hispanic	2.4	2.2
Asian	2.2	1.5
All Minorities	10.4	7.8

Source: Derived from data obtained from Naval Aerospace and Operational Medical Institute.

For every minority group, the percentage of persons is higher in the "deselected" sample than in the group as a whole. This is consistent with the findings in the analysis of the "selected" pilots. Since the percentage of these

[3] Of course, it is very unlikely that the Navy and Marine Corps would have accepted a shortage of this many pilots over the four-year period in question. However, the cutoff scores for the selection test are based in part on certain minimum acceptable levels of predicted criteria performance. As a result, it is unlikely that any shortage in the supply of qualified candidates would be corrected by lowering the cutoff score. Historically, the fluctuations in supply are stabilized by changing the intensity of the recruiting efforts. Moreover, if there were an actual shortage, it is likely that the recruiting efforts would be intensified across the board, rather than on certain particular racial/ethnic groups. Therefore, the percentages of each group in the applicant pool and the resulting selected population would remain unchanged.

groups in the "selected" sample decreased, they must necessarily increase in the "deselected" sample. Again, the effect is most pronounced among Black candidates, going from 3.3 percent in the overall population to 5.1 percent in the "deselected" sample.

As mentioned earlier, the use of a simulated higher cutoff score also allows analysis of the criterion performance of the "deselected" population. Table 10 presents the primary flight grades and attrition rates for the "deselected" pilots and the overall group.

Table 10. Mean Flight Grades and Attrition Rates: "Deselected" Pilots vs. Overall Group

	"Deselected" Pilots	Overall Group
Mean Flight Grade	3.019	3.055
Attrition Rate*	.30	.20

*The attrition data are for all phases of aviation training, not just primary flight. They include academic and flight failures, as well as physical disqualifications that arise after the initial screening process.

Source: Derived from data obtained from Naval Aerospace and Operational Medical Institute.

Since the pre-1992 version of the selection test was still a valid predictor, these results are not surprising. The mean primary flight grades of the "deselected" group (3.019) are approximately one standard deviation below the mean grades for the group as a whole. Since primary flight performance was one of the predicted criteria for the selection tests (both "New" and "Old") one would expect to see lower flight grades, on average, for a group with lower test scores.

As seen in Table 10, the attrition rates show a similar pattern. This "deselected" group experienced a 30 percent attrition rate, as opposed to 20 percent for the "Old Test" group as a whole. This is also expected. As with primary flight performance, both selection tests are designed to predict an applicant's likelihood of attrition. Therefore, one would expect to find that candidates with lower test scores are, on average, higher attrition risks. Attrition is expensive in any setting, but flight school attrition is of special concern because of the significant costs in training naval aviators. A difference of ten percentage points in flight school attrition can translate into significant savings. Still, there is another side to the issue that is worthy of consideration. The attrition rate of 30 percent experienced by the "deselected" pilots also means that 70 percent of them successfully completed training and earned their wings. This means that seven out of every ten pilots who were "deselected" by the study would have been "false negatives": in the simulation, they would have failed to score high enough for acceptance into training; but, in fact, they successfully completed the course. Still, this result should be interpreted with some caution, especially when relating it to what may be happening under the current test. Since the 1992 ASTB has increased validity, the numbers of "false positives" and "false negatives" are reduced. This "squeezing" of the test score/criteria data (as depicted in Figure 2) will force more of the observations into the "true positive" and "true negative" categories, therefore reducing both the number and the proportion of incorrect predictions.

V. DISCUSSION, CONCLUSIONS AND RECOMMENDATIONS

A. DISCUSSION AND CONCLUSIONS

The 1992 version of the ASTB appears to be more effective than the previous version at selecting pilot candidates, at least in terms of their performance in one of the predicted criteria, primary flight training. Primary flight grades have increased, and this increase can be reasonably attributed to the increased effectiveness of the selection test, assuming that training and grading criteria have remained constant from 1988 to 1995.

Had the Marine Corps applied the same higher cutoff score in 1988 as in 1995, the primary flight performance distribution of the selected pilots would have likely increased. However, the proportion of minority pilots in the selected population would have likely decreased markedly. The effects would have been the most dramatic among Black applicants, but the effects would also have been strong among the proportion of Asian and Hispanic candidates. The degree of the impact on a particular minority group correlates with the average performance of that group on the test: the lower the average test score, the greater the impact.

The population of candidates who were "deselected" by the simulated higher cutoff score performed at a lower level in terms of primary flight grades than did the group as a whole. Also, this "deselected" group had a higher attrition rate from flight training than did the "selected" pilots.

Since the criterion measure used to simulate the higher cutoff score on the "Old Test" data is comparable to the criterion measure in the "New Test" data, it can be inferred that similar effects may be occurring in the Marine Corps under the current cutoff score. In short, the population of pilots that the Navy is accepting for flight training, but the Marine Corps is not, may be similar to the "deselected" pilots in the study simulation.

The study has analyzed some important questions about the possible effect of the current Marine Corps ASTB cutoff score on the selection of racial/ethnic minorities into the flight program. It has also examined the relationship of that score to two aspects of flight school performance, primary flight grades and attrition. Of course, the effects are estimates, since the methodology is based on a simulation of the use of the higher cutoff score on the more extensive "Old Test" data. Still, because the simulation was based on the use of an assumed cutoff score that yielded a "selected" population of pilots whose performance mirrored that of the "New Test" pilots, it is reasonable to suggest that the simulated effects may be similar to actual effects. One point, however, has become patently clear: in any "Cleary fair" test such as the ASTB, when the distributions of test scores and criterion performance between population subgroups differ, *the location of the cutoff score can have a marked effect on the demographic mix of the selected population.* This idea was suggested by the general theoretical model and was borne out by the data.

As time passes, more and more data will become available on the "New Test" pilots. As data become available, the Marine Corps can begin to get a better feel for the specific effects of the higher cutoff score, and can make policy decisions based on the evidence. For the time being, however, the Marine Corps can certainly enhance its understanding of this issue and study other options in an attempt to take an aggressive stance in the event that the estimates of this study prove to be an accurate assessment of current trends. Some options and possible areas of further study and analysis are presented below.

B. RECOMMENDATIONS

1. Additional Selection Procedures

The 1992 ASTB is a solid, useful selection instrument. It is reliably explaining a significant portion of the observed variation in flight student academic performance, primary flight performance, and attrition. The test battery can quickly and easily be administered and graded (subject to NAMI verification) in a recruiting office or on a college campus. In short, the ASTB provides a large "bang for the buck" in comparison with other selection devices. This is true of most pencil-and-paper selection tests, which is one reason why they are so widely used. To markedly improve the selection process as a whole, however, would require more than the use of a test. The Navy and Marine Corps would need to find ways to account for the portion of the variation in performance left unexplained by the ASTB. All of the services, and even civilian airlines, have studied the use of additional selection procedures that add predictive power to the selection process above and beyond the written test. Simple psychomotor apparatus tests and computer-based risk-taking analysis are among the methods that scientists both in and out of the military are studying in an attempt to strengthen the selection process. The key for the Navy and Marine Corps is to seek out selection procedures that can explain a unique portion of performance variation and still be cost-effective. Of course, these additional kinds of selection devices can be expensive in terms of both acquisition and administration. However, if they provide additional predictive capability, they will improve the performance of our student pilots and reduce attrition, resulting in reductions in training costs that may offset the expense of the selection instrument.

2. Minority Recruiting Efforts

The central issue for the cutoff score, as stated earlier, is where that score falls along the test score/criterion performance distributions of the different racial/ethnic groups. Obviously, then, moving the cutoff score will change the mix of the selected group. However, another option is to attempt to *move the score/performance distributions themselves.* This would be a function of recruiting efforts. With a cutoff score held in place, more effective recruiting of minorities will, over time, raise the score/performance distributions of those groups, moving more and more of them to the right of the cutoff score. Although this is not likely to change the overall criterion performance of the flight students[1], it will certainly improve the percentages of racial/ethnic minority groups that are selected. Of course, it must be recognized that this can be an expensive and difficult proposition. The selection process for naval aviators is very stringent, drawing candidates from the top-performing layers of the general population. This is especially true in the Marine Corps, since the requirements to become a Marine Officer are so high by themselves. Add to the Marine Officer standards the physical, intellectual, and psychological standards for flight training and the result is an even smaller pool of qualified candidates. Now consider the nature of the labor market for racial/ethnic minorities in this group. As so many organizations attempt to expand their representation of racial/ethnic minorities, the market value of these individuals grows. They are smart, self-confident, motivated leaders who would be an asset to any organization *regardless* of their minority status. In short, naval aviation is working in a highly competitive labor market. These are not arguments against the more aggressive recruiting of minorities to increase diversity. They simply recognize that raising the overall distribution of minority groups in the applicant pool would be a significant challenge for recruiting commands, and it would likely require larger financial commitments to the recruiting process as well as the continued personal commitment of the recruiting community in the Marine Corps.

3. Selective Waivers

This may be the most controversial policy option. The idea would be to allow selected ASTB score waivers, down to the Navy standard of 3/4/4, for

[1] Criterion performance of the population selected by a "Cleary fair" test, training techniques held constant, will be a function of the location of the cutoff score alone as long as the shapes of the distributions of the different groups are not significantly different.

example. These waivers could be considered on a case-by-case basis, and would be granted for "otherwise exceptionally qualified candidates." One serious challenge, to ensure fairness in the waiver process, would be to arrive at some quantitative definition of "otherwise exceptionally qualified." Not only must these other qualifications be measurable, but they must be attributes that are not measured by the current selection process. For example, the granting of a waiver for a candidate who appears to have exceptional mechanical abilities makes little sense when mechanical comprehension (as it relates to aviation) is already measured by the ASTB. The search for criterion-related variables that exist outside of the current selection process is indeed a difficult one. After all, the whole point of the development of the selection process over the years was to define and measure those variables.

Another, and perhaps more serious problem with waivers, is one of perception. Waivers imply lowered standards. If the cutoff score is set at five, for example, then the argument is that it should apply to all candidates, not to only certain groups. This becomes even more controversial when racial/ethnic considerations become part of the decision. The use of waivers simply to access more of a desired group are legally problematic, since it implies the use of differential cutoff scores, a practice prohibited by the Civil Rights Act of 1991 (noted in Chapter Two). Using waivers targeted at "otherwise exceptionally qualified candidates," while legal, may create the same perception, especially if larger proportions of waivers are granted to minority applicants.

The problem, then, centers around other attributes that make a candidate qualified but are not part of the selection process. If waiver criteria are used that do not relate to flight school performance, then the Marine Corps would likely pay a price for the waivers. Namely, the waivered candidates would, on average, perform at a lower level and attrite at higher rates than non-waivered candidates.

This might in turn add credence to the argument that the granting of waivers is simply a lowering of entry standards to improve the demographic characteristics of the selected population at the expense of performance. Additionally, a waiver that is not based on some criterion-related measurement may be largely self-defeating, especially in the area of attrition. Although it is true that waivering may allow the selection process to capture more of the distribution of racial/ethnic minority groups, a resulting increase in attrition may negate these gains. After all, the ultimate goal is to increase the diversity of the fleet, not just flight students, and waivered candidates have a higher risk than non-waivered candidates of never reaching fleet squadrons.

Still, other options exist in the area of selective waivers that are worth studying. First, as previously mentioned, is expansion of the selection process to find and measure variables that would help explain performance variation above and beyond that of ASTB scores. This could allow waivering (or outright lowering) of minimum ASTB scores, thereby capturing more of the minority distributions without compromising performance or increasing attrition. Of course, as mentioned before, this is probably a difficult and costly proposition. Additionally, to make the waivers effective, the added selection factors would have to show less of a difference between group scores than is the case with the selection test.

Another possibility comes as a result of the unique predictive power of the 1992 ASTB. As stated above, higher attrition from flight school could be a significant cost in granting waivers. However, the prediction of attrition is concentrated in a distinct portion of the ASTB, the Biographical Inventory (BI). If the minimum BI score were held at a constant level,[2] but selective waivers were granted for the other portions of the test, the attrition costs of the waiver might be

[2] Currently, the Navy and Marine Corps use the same cutoff score for the BI.

largely controlled. Academic and flight performance, of course, would likely decline among waivered candidates. In this study, for example, average flight grades for potential waivered candidates (the "deselected" pilots) were approximately one standard deviation below the mean. The question then becomes, how much does this matter in the long run? There may be some impact on pipelines, since the selection of flight students into the different aviation communities is largely based on flight grades. But would the lower grades affect the success of these pilots in the fleet? The answer is unclear, but it is worthwhile to remember that the selection test does not claim predictive power beyond flight training, and flight grades may be similarly unreliable predictors of fleet success.

4. Adverse Impact Analysis

An analysis of adverse impact was conducted by the test developers in the original validation of the 1992 ASTB. Adverse impact is a function of the selection rates of different population subgroups. It is based on the "four-fifths rule" from the 1978 Uniform Guidelines referenced in Chapter Two. The validation showed that the selection rates for minority groups were no less than eighty percent of the selection rate for the majority group. However, this analysis was conducted based on a cutoff score of 3/4/4, not the Marine Corps cutoff of 4/6/4. One of the key issues brought out in the present study is that selection rates can vary greatly with the use of different cutoff scores. The "New Test" data for this study only included officers who had been accepted for training. Although these data were not extensive enough to allow detailed analysis of the flight grades and demographic characteristics of candidates who fell between the Navy and Marine Corps cutoff scores, the data on *all applicants* may be extensive enough for a reasonable estimate of selection rates for different racial/ethnic groups. The demonstrated validity of the 1992 ASTB will allow it to remain a legal selection instrument even if adverse impact exists; still, the nature of the current selection rates is a worthwhile area of study.

Many of the issues analyzed and discussed in this study are of ongoing concern to manpower planning organizations in the Marine Corps. Some of the topics can be controversial, and can generate a great deal of interest and scrutiny both in and out of the military. With a greater understanding of these and other personnel selection issues as they relate to minorities, the Marine Corps can continue to align policy decisions with the goals of expanding racial/ethnic representation and maintenance of the high performance standards that are the hallmark of the United States Marine.

LIST OF REFERENCES

1. Anastasi, A., Differential Psychology, McMillan, New York, 1965.

2. Arvey, R. D. and Faley, R. H., Fairness in Selecting Employees, 2nd edition, p. 122, Addison-Wesley, Reading, Massachusetts, 1988.

3. Bucky, S.F., The California Psychological Inventory Given to Incoming AOC's and DOR's With Normal and Ideal Instructions, Naval Aviation Medical Research Laboratory Report 1127, Pensacola, Florida, 1971.

4. Cattell, R., Eber, H., and Tatsuoka, M., Handbook for the Sixteen Personality Factor Questionnaire, Institute for Personality and Ability Testing, Champaign, IL, 1990.

5. Civil Rights Act of 1991, Statutes at Large, 105, sec.106, 1075, 1991.

6. Cleary, T.A., Test Bias: Prediction Of Grades Of Negro And White Students In Integrated Colleges, pp. 115-124, Journal of Educational Measurement 5, 1968.

7. Cole, N.S., Bias in Selection, ACT Research Report 51, American College Testing Program Iowa City, Iowa, 1972.

8. Cunningham, M., Wong, D., and Barbee, A., Self-presentation Dynamics on Overt Integrity Tests: Experimental Studies of the Reid Report, pp. 643-658, Journal of Applied Psychology 79, 1994.

9. Darlington, R.B., Another Look at Culture Fairness, pp. 71-82, Journal of Educational Measurement 8, 1971.

10. Division 14 Principles, Equal Employment Advisory Committee, 1980.

11. Dreger, R. M., Comparative Psychological Studies of Negroes and Whites in the United States: 1959-1965, pp. 261-269, Psychological Bulletin 75, 1968.

12. Einhorn, H.J. and Bass, A.R., Methodological Considerations Relevant to Discrimination in Employment Testing, Psychological Bulletin 75, pp. 261-269, 1971.

13. Examiners Manual and Scoring Instructions, U. S. Navy and Marine Corps Aviation Selection Tests, NAVMED P-5098, Aerospace Operational Psychology Branch, Bureau of Medicine and Surgery, Navy Department, Washington, D. C., 1971.

14. Fiske, D.W., Validation of Naval Aviation Cadet Selection Tests Against Training Criteria, pp. 601-613, Journal of Applied Psychology, 31, 1947.

15. Fleischman, H., Ambler, R., Peterson, F., and Lane, N., The Relationships of Five Personality Scales to Success in Naval Aviation Training, NAMI-968, Naval Aerospace Medical Institute, Pensacola, FL, 1966.

16. Frank, L.H. and Baisden, A.G., The 1994 Navy and Marine Corps Aviation Selection Test Battery Development, presented at the Annual Meeting of the Military Testing Association, Williamsburg, VA.

17. Frank, L.H., Biographical Inventory Validation Assessment, Naval Aerospace and Operational Medical Institute, Pensacola, FL, May 1994.

18. Hough, E., Eaton, N., Dunnette, M., Kamp, J., and McCloy, R., Criterion-related Validities of Personality Constructs and the Effect of Response Distortion on Those Validities, pp. 581-595, Journal of Applied Psychology Monograph 75, 1990.

19. Jensen, A.R., Bias in Mental Testing, pp. 311-312, The Free Press, New York, 1980.

20. Kaplan, H., Prediction of Success on Army Aviation Training, Technical Research Report 1142, U.S. Army Personnel Research Office, OCRD, 1965.

21. Leary, M. and Kowalski, R., Impression Management: A Literature Review and Two-Component Model, pp. 34-47, Psychological Bulletin, 107, 1990.

22. Luk'yanova, N., Personality Characteristics of Pilot-cadets With Different Marks in Flight Disciplines, U.S. Army Foreign Service and Technology Center, Charlottesville, VA, 1977.

23. McFarland, R.A. and Franzen, R., The Pensacola Study of Naval Aviators -- Final Summary Report, Rept. 38, Division of Research, Civil Aeronautics Administration, Washington, D.C., 1944.

24. Merydith, S.P., and Wallbrown, F.H., Reconsidering Response Sets, Test-taking Attitudes, Dissimulation, Self-deception, and Social Desirability, pp. 891-905Psychological Reports 69, 1991.

25. Miller, R. E., Interpretation and Utilization of Scores on the AFOQT-AFHRL-TR-69-103, Personnel Research Laboratory, Lackland AFB, Texas, 1969.

26. Neiner, A. , Examples of Testing Programs in the Insurance Industry, *in Test Policy and the Politics of Opportunity Allocation: The Workplace and the Law*, ed. Gifford, B., Kluwer Academic, Norwell, 1989.

27. North, R.A. and Griffin, G.R., Aviator Selection 1919-1977, Naval Medical Research Laboratory, Pensacola, Florida, 1977.

28. Peltier, B.D. and Walsh, J.A., An Investigation of Response Bias in the Chapman Scales, pp. 803-815, Educational and Psychological Measurement 50, 1990.

29. Power, R. and McRae, K., Characteristics of Items in the Eysenck Personality Inventory Which Affect Responses When Students Simulate, pp. 491-498, British Journal of Psychology 68, 1977.

30. Profile of American Youth: 1980 Nationwide Administration of the Armed Services Vocational Aptitude Battery, pp. 30-36, U.S. Department of Defense, Office of the Assistant Secretary of Defense, Manpower, Reserve Affairs and Logistics, Washington D.C., 1982.

31. Standards for Educational and Psychological Testing, American Psychological Association, Washington, D.C., 1985.

32. Thiriart, P., Acceptance of Personality Test Results, pp. 161-165, Skeptical Inquirer 15, 1991.

33. Thorndike, R.L., Concepts of Culture Fairness, pp. 63-70, Journal of Educational Measurement 8, 1971.

34. Thorndike, R.L., Personnel Selection, pp. 170-171 and pp. 186-198, John Wiley and Sons, New York, 1949.

35. Uniform Guidelines on Employee Selection Procedures, Section 5, pp. H.

36. Wing, H., Profiles of Cognitive Ability of Different Racial Ethnic and Sex Groups on a Multiple Abilities Test Battery, pp. 289-298, Journal of Applied Psychology 3, 1980.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center.. 2
   8725 John J. Kingman Road, Ste. 0944
   Ft. Belvoir, VA 22060-6218

2. Dudley Knox Library... 2
   Naval Postgraduate School
   411 Dyer Road
   Monterey, CA 93943-5101

3. Navy Manpower Analysis Center.. 2
   Code 531
   NAS Memphis
   5820 Navy Road
   Millington, TN 38054-5056

4. Director, Training and Education 1
   MCCDC, Code C46
   1019 Elliot Rd.
   Quantico, VA 22134-5027

5. Mark Eitelberg, SM/Eb.. 4
   Naval Postgraduate School
   Monterey, CA 93943-5103

6. Tony Ciavarelli.. 2
   Aviation Safety Programs
   Naval Postgraduate School
   Monterey, CA 93943-5103

7. Naval Aerospace and Operational Medical Institute.................................. 1
   Code 41
   220 Hovey Road
   Pensacola, FL 32508-1047

8. Commanding General (MROR) 1
Marine Corps Recruiting Command
Headquarters, U.S. Marine Corps
2 Navy Annex
Washington, DC 20380

9. Brian J. Dean 2
c/o John DeRosa
64 Morris Avenue
West Milford, NJ 07480