Clearance number: HSC96-062

Group Differences on US Air Force Pilot Selection Tests[1,2]

Thomas R. Carretta

Sex and ethnic group differences were examined on the operational composites and tests used to select applicants for U. S. Air Force officer commissioning programs and for pilot training. Results showed that large mean score differences in applicant samples were substantially reduced among the pilot trainees. Despite differences in test performance, there was no evidence of differential validity for groups. When group differences in predicted pilot training completion rate were observed, performance was overestimated for the minority group relative to the majority group. When regression equations were adjusted for unreliability of the predictors, the observed differences in intercepts were reduced or eliminated. No prediction bias was observed against the minority groups.

The performance of sex and ethnic groups on ability tests has come under increasing scrutiny, especially when the tests are used for personnel selection in educational and occupational settings (Burke, 1995; Hartigan & Wigdor, 1989; Jaeger, 1976; Jensen, 1980; Linn, 1982; Wing, 1980). Differences in cognitive test performance are documented for sex (Burke, 1995; Carretta, 1990; Hyde, 1981; Jensen, 1980; Maccoby & Jacklin, 1974; Siem & Sawin, 1990) and racial/ethnic groups (Brody, 1992; Coleman et al., 1966: Jensen, 1973, 1980; Loehlin, Lindzey, & Spuhler, 1975; Mathews, 1977). Despite

---

[1] Previously published as Carretta, T. R. (1997). Group differences on US Air Force pilot selection tests. *International Journal of Selection and Assessment, 5,* 115-127.

[2] The views expressed are those of the author and not necessarily those of the United States Air Force, Department of Defense, or United States Government. The author thanks Dr. Malcolm James Ree, Dr. Patrick C. Kyllonen, Dr. William E. Alley, and Dr. Rod McCloy for their many helpful comments. Send correspondence and requests for reprints to AFRL/HECI, 2210 8th Street, Area B, Bldg. 146, Room 122, Wright-Patterson AFB, OH 45433-7511. Address e-mail to thomas.carretta@wpafb.af.mil.

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
1997	**N/A**	**-**

4. TITLE AND SUBTITLE **Group Differences on US Air Force Pilot Selection Tests**	5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Air Force Research Laboratory Wright-Patterson AFB, OH 45433-7022**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S) 11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**23**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

group differences in test performance, there is little convincing evidence that well constructed tests are more valid predictors of academic, training, or occupational criteria for the majority (usually males or Whites) than for minority groups (usually females or racial/ethnic minorities).

Even though the U. S. Air Force (USAF) has shown a long-term interest in ensuring equal opportunity and fair employment practices, few studies have been conducted to examine group differences in performance on officer and aircrew selection tests (Carretta, 1990; Mathews, 1977; Roberts & Skinner, 1996; Siem & Sawin, 1990). The present study investigated the performance of applicants on the Air Force Officer Qualifying Test (AFOQT) and examined its validity for the prediction of pilot training performance for males and females and for Whites, Blacks, and Hispanics.

Air Force pilot training applicants must first qualify for an officer commissioning program through the Air Force Academy (AFA), Officer Training School (OTS), or Reserve Officer Training Corps (ROTC). Pilot candidate selection factors include medical and physical fitness, academic performance, aptitude test scores, commander's recommendations, and previous flying experience. The AFOQT (Arth, Steuck, Sorrentino, & Burke, 1990; Carretta & Ree, 1995; Skinner & Ree, 1987) has been used by OTS and ROTC boards for commissioning and pilot selection since 1957. During that period, the AFOQT has been revised including changes in content, with new forms implemented about every 7 years.

The AFOQT Pilot composite contributes to a pilot candidate selection composite known as the Pilot Candidate Selection Method or PCSM (Carretta, 1992). To date, few studies have examined the validity of AFOQT scores against pilot training performance for females (Carretta, 1990; Siem & Sawin, 1990) and ethnic minorities (Whites vs. Blacks only; Mathews, 1977). The Carretta and the Siem and Sawin studies reported similar results. Male pilot trainees had higher mean scores than female pilot trainees on selection factors (i.e., AFOQT Pilot and Navigator-Technical composites) and were more likely to complete pilot training. However, when males and females were matched on test scores, in most instances, they performed equally well in pilot training. When sex differences occurred, training performance was overpredicted for females.

Mathews (1977) observed that White pilot trainees had higher AFOQT scores than Black pilot trainees and were more likely to complete training successfully. Further, pilot training performance was overpredicted for Blacks when compared to Whites with similar AFOQT scores. Mathews (1977) concluded that the AFOQT was not biased against Blacks.

The present study extended previous research by investigating mean score differences for sex and three ethnic groups for USAF officer applicants and pilot trainees for the AFOQT composites and tests. Further, the validity of both the composites and tests was examined in this study. Previous studies (Carretta, 1990; Mathews, 1977; Siem & Sawin, 1990) focused on pilot trainees and the AFOQT composites. Examination of test-level (in addition to composite-level) differences on the AFOQT was done because previous research has shown group differences to be

dependent on kind of ability measured (Burke, 1995; Brody, 1992; Coleman et al., 1966; Hyde, 1981; Jensen, 1973, 1980; Loehlin et al., 1975; Maccoby & Jacklin, 1974).

It should be noted that none of the previous studies corrected their results for range restriction or other statistical artifacts that may have affected the observed correlations and therefore, interpretation of the results. The present study examined both observed correlations and correlations corrected for range restriction.

Method

Participants

The participants were 269,968 Air Force applicants who were tested on equivalent forms of the AFOQT (Form O or P) between 1981 and 1992. They were mostly male (81.4%) and White (78.7%).

The pilot training validation sample consisted of 9,476 Air Force officers attending Undergraduate Pilot Training (UPT). All participants were chosen for pilot training in the same manner. A board of senior officers reviewed and ranked applicants on the basis of their AFOQT scores, educational achievement, and other signs of accomplishment. All participants had completed at least a baccalaureate degree before entering pilot training. As with the Air Force applicants, the pilot trainees were mostly male (97.5%) and White (94.6%).

Measures

The AFOQT (Skinner & Ree, 1987) is a paper-and-pencil multiple aptitude battery used to select civilian or prior service military applicants for officer precommissioning training programs (i.e., OTS, ROTC) and classify commissioned officers into aircrew training programs (i.e., pilot or navigator). It has 16 tests that measure general cognitive ability and the 5 lower-order factors of verbal, quantitative, spatial, aircrew interest/aptitude, and perceptual speed (Carretta & Ree, 1996). As shown in Table 1, the 16 tests are combined into 5 composites: Verbal (V), Quantitative (Q), Academic Aptitude (AA), Pilot (P), and Navigator-Technical (N-T). AFOQT composite and test reliabilities are given by Arth (1986) and Skinner and Ree (1987).

Verbal tests. Verbal Analogies (VA) requires the ability to recognize the relationships between words and provides a measure of reasoning ability. Reading Comprehension (RC) measures the ability to read and understand paragraphs. Word Knowledge (WK) assesses the ability to understand words through the use of synonyms.

Quantitative tests. Arithmetic Reasoning (AR) measures the ability to solve arithmetic word problems. Data Interpretation (DI) assesses the ability to interpret data presented in tables and charts. Math Knowledge (MK) requires the ability to use mathematical terms, formulas, and relationships to solve problems.

Spatial tests. Mechanical Comprehension (MC) assesses mechanical knowledge and reasoning. Electrical Maze (EM) measures spatial ability based on choice of a path through a maze. Block Counting (BC) assesses spatial ability through interpretation of a three-dimensional representation of a set of blocks. Rotated Blocks (RB) measures the ability to visualize and manipulate objects in space. Hidden Figures (HF) requires the detection

of simple figures embedded in complex drawings.

Aircrew interest/aptitude tests. These are the only AFOQT tests that measure specific knowledge (Dye, Reck, & McDaniel, 1993; Olea & Ree, 1994). Instrument Comprehension (IC) measures the ability to determine the attitude of an aircraft in flight from illustrations of instruments. Aviation Information (AI) assesses familiarity with aviation concepts and terminology. General Science (GS) measures knowledge and understanding of scientific terms, concepts, principles, and instruments.

Table 1. Composition of AFOQT Aptitude Composites

		Composite				
Test	Reliability[a]	Pilot (P)	Navigator-Technical (N-T)	Verbal (V)	Quantitative (Q)	Academic Aptitude (AA)
VA	.80	X		X		X
AR	.81		X		X	X
RC	.88			X		X
DI	.71		X		X	X
WK	.88			X		X
MK	.88		X		X	X
MC	.71	X	X			
EM	.81	X	X			
SR	.84	X	X			
IC	.84	X				
BC	.83	X	X			
TR	.92	X	X			
AI	.77	X				
RB	.77		X			
GS	.70		X			
HF	.69		X			

Notes. [a]Test reliabilities were based on coefficient alpha from a normative sample of 3,000 USAF officer applicants (Skinner & Ree, 1987).

Perceptual speed tests. Scale Reading (SR) measures the ability to read and interpret scales, dials, and meters. Table Reading (TR) assesses the ability to read tables quickly and accurately and to extract information from them.

Criterion

USAF pilot trainees complete a 53 week program including instruction in job knowledge and job skills. Undergraduate Pilot Training (UPT) consists of a ground school phase, an initial jet phase (T-37 subsonic aircraft), and an advanced jet phase (T-38 transonic aircraft). Training includes about 190 hours of flying. After graduation from training, pilots are assigned to fly aircraft ranging from multi-engine transports to high performance jet fighters. Final training outcome (passing/failing; UPT P/F) is determined by academic and flying performance. UPT P/F was coded as 0 for eliminees and 1 for graduates. Over the last several years, UPT attrition rates have been steady at about 22%. Most attrition occurs for flying training deficiencies (i.e., inability to fly the aircraft) and, to a lesser extent, self-initiated withdrawal from training and airsickness. Elimination for other reasons such as academic, disciplinary, or medical, are less common. Trainees who failed for reasons other than flying training deficiencies were not included in the study.

Procedures

Test scores and flying performance criteria were collected from official records. Test scores were collected prior to entrance into commissioning programs and flying training. Most participants entered UPT within 2 years after AFOQT testing, but for some, the interval was as much as 5 years. UPT final outcome (passing/failing) was awarded at the end of the 53 week training program.

Analyses

All t-tests were one-tailed (i.e., majority group - minority group). An overall Type I error rate of .05 was used for each group of related tests (e.g., differences in means for male vs. female pilots). A Bonferroni approach (Miller, 1981) was used to control for the experiment-wise error rate within each group of related tests (males vs. females, Whites vs. Blacks, Whites vs. Hispanics). As a result, each t-test of mean differences used a .0024 Type I error rate (.05/21 tests) for the comparisons of the AFOQT scores. Mean differences in UPT pass/fail rate were tested using a .05 Type I error rate. Correlation of the AFOQT composites and tests with the UPT P/F criterion and group differences in correlations with the criterion (males - females, Whites - Blacks, Whites - Hispanics) used a .0024 Type I error rate (.05/21 tests). The Type I error rate was set to .01(.05/5 tests) for the regression models because only the 5 AFOQT composites were examined.

Results

Means

AFOQT composite and test means were examined within each sex and ethnic group for applicants and pilot trainees. The magnitude of the differences between means (i.e., effect size) was expressed in standard deviation units or d (Cohen, 1988). The standard deviation for *d* was defined as the within-group standard deviation (SD = $(Sp^2/n_1 + Sp^2/n_2)^{1/2}$), where S_p^2 is the pooled variance calculated from the weighted average of the variances for the two groups

being compared (i.e., males vs. females, Whites vs. Blacks, Whites vs. Hispanics). S_p^2 is defined as $(SS_1 + SS_2)/(n_1 + n_2 - 2)$. Subscripts 1 and 2 indicate two independent groups being compared. See for example, McNemar (1969, p. 115) Thus, $d = (Mean_1 - Mean_2) / SD$. Cohen (1988) characterizes a d of .20 as small, .50 as medium, and .80 as large. It should be noted, however, that even "small" *d* values can have a large impact on the proportion of applicants in the lower mean group that would meet or exceed some minimum cut score for selection. Group mean differences were tested using one-tailed t-tests (majority group - minority group).

Means and standard deviations for AFOQT scores are shown for officer applicants (Table 2) and pilot trainees (Table 3) by sex and ethnicity. The proportion passing UPT ranged from .65 to .80 with

males and Whites being most likely to complete training.

Sex differences. Among the applicants, males significantly outperformed females on all composites and 15 of 16 tests. The exception was an effect size of .02 on VA. The d values for the composites ranged from .69 (Pilot) to .08 (Verbal) with a mean of .422 and a median of .440. The d values for the 16 tests ranged from .02 (VA) to .95 (MC) with mean and median values of .435.

Comparisons of the composite and test means for the pilot trainees revealed some interesting findings. The *d* values for the 5 composites ranged from -.48 to .20, with a mean of -.096 and a median of -.050. The mean Navigator-Technical composite score for male pilot trainees was greater than that for females (*d* = .20), but the difference was not nearly as large as in the applicant groups. Male means did not exceed female means on the other 4 composites (Pilot, .17; Verbal, -.48; Quantitative, -.05; Academic Aptitude, -.32).

The *d* values for the 16 tests for pilot trainees ranged from -.63 (VA) to .84 (MC), with a mean of .078 and a median of .020. Male means exceeded female means on 6 tests (MC, EM, IC, AI, RB, and GS). In general, the means for males exceeded those for females only on the aircrew interest/aptitude tests and some spatial tests: IC (.45), AI (.22), GS (.47), MC (.84), EM (.41), and RB (.56).

Ethnic differences. In the applicant sample, the White means exceeded the Black and Hispanic means on all composites and tests. All differences were statistically significant. On the composites, White means were greater than Black means by about 1.35 SD for Pilot (1.40) and Navigator-Technical (1.33) and about 1 SD for Verbal (1.07), Quantitative (1.10), and Academic Aptitude (1.18). White means exceeded the Hispanic means by about .80 SD on all composites (Pilot, .80; Navigator-Technical, .77; Verbal, .80; Quantitative, .72; Academic Aptitude, .84). For the 16 tests, White-Black mean differences ranged from .68 (EM) to 1.20 (SR), with a mean d value of .961 and a median value of .955. White-Hispanic mean differences ranged from .30 (EM) to .86 (VA)with a mean of .551 and a median of .510.

White pilot trainee means exceeded those for Blacks and Hispanics on all composites and most tests, but by a much smaller amount than in applicant samples. The mean and median *d* values for the White-Black composite comparisons were .522 and .460 with d values of: Pilot (.67), Navigator-Technical (.67), Verbal (.28), Quantitative (.53), and Academic Aptitude (.46). The mean and median d values for the White-Hispanic composite comparisons both were .400 and the *d* values were: Pilot (.46), Navigator-Technical (.45), Verbal (.36), Quantitative (.33), and Academic Aptitude (.40).

For pilot trainees, the mean White-Black difference on the 16 tests ranged from .11 (WK) to .59 (SR) with a mean of .337 and a median of .340. The White-Hispanic differences were somewhat smaller than those for Whites and Blacks. They ranged from .07 (AI) to .42 (SR) with a mean of .246 and a median of .250.

Correlations

Observed correlations between AFOQT scores and UPT P/F were tested as well as differences in observed correlations for pairs of groups (i.e., males - females, Whites - Blacks, Whites - Hispanics; majority - minority > 0; see Glass & Stanley, 1970).

Significance tests were not done for correlations corrected for range restriction.

ble 2. Means and Standard Deviations for AFOQT Scores by Sex and Ethnicity (US Air Force Applicants)

	Sex					Ethnicity							
	Male		Female		*d*	White		Black		*d*	Hispanic		*d*
ore	Mean	SD	Mean	SD	M-F	Mean	SD	Mean	SD	W-B	Mean	SD	W-H
	49.42	25.53	32.15	21.37	0.69*	51.64	23.95	19.13	17.26	1.40*	32.46	22.17	0.80*
Γ	49.17	26.65	32.59	23.49	0.63*	51.43	25.28	18.83	18.25	1.33*	32.06	23.18	0.77*
	48.57	27.30	46.26	28.47	0.08*	53.20	26.15	25.71	22.31	1.07*	32.29	24.77	0.80*
	46.40	27.09	34.55	25.19	0.44*	48.83	26.11	20.74	20.02	1.10*	30.03	23.59	0.72*
ι	46.64	27.19	39.07	26.58	0.27*	50.50	25.86	20.63	19.93	1.18*	28.73	23.39	0.84*
ι	14.25	4.34	14.12	4.72	0.02	15.07	3.97	10.43	4.41	1.15*	11.61	4.44	0.86*
ι	12.51	5.00	10.28	4.80	0.44*	12.91	4.81	7..91	4.16	1.05*	9.78	4.62	0.65*
ι	15.79	5.62	15.24	5.86	0.09*	16.77	5.19	10..91	5.36	1.12*	12.43	5.69	0.83*
	12.88	4.65	11.48	4.51	0.30*	13..35	4.48	9.10	3..91	0.96*	10.44	4.38	0.65*
K	13.86	5.73	13.35	5.94	0.08*	14.63	5.54	9.85	5.21	0.87*	11.26	5.41	0.60*
K	15.09	5.91	13.34	5.95	0.30*	15.53	5.74	10.47	5.27	0.89*	12.57	5.71	0.51*
C	10.22	3.67	6.83	2.86	0.95*	10.25	3.62	6.41	2.99	1.08*	7.94	3.34	0.64*
I	8.08	4.05	5.87	3.06	0.56*	8.08	4.00	5.42	3.06	0.68*	6.87	3.64	0.30*
.	21.54	6.60	18.27	6.64	0.49*	22.14	6.25	14.67	5.91	1.20*	17.93	6.33	0.67*
	10.47	5.03	6.68	3.85	0.78*	10.51	4.96	5.91	3.73	0.95*	8.35	4.71	0.43*
ι	11.40	4.28	9.53	4.46	0.43*	11.77	4.10	7.21	3.92	1.11*	9.71	4.27	0.50*
ι	26.89	7.22	26.38	7.46	0.07*	27.77	6.83	21.47	7.39	0.91*	24.58	7.29	0.46*
	9.11	4.13	6.04	2.78	0.78*	9.13	4.06	5.74	3.02	0.86*	7.11	3.81	0.49*
ι	8.14	3.26	5.74	3.01	0.74*	8.18	3.19	5.11	3.02	0.96*	6.90	3.24	0.40*
	9.03	3.62	6.59	2.87	0.69*	9.09	3.56	6.08	2.84	0.86*	7.27	3.30	0.51*
	9.70	2.91	9.00	2.90	0.24*	9.89	2.81	7.81	2.88	0.73*	8.98	2.89	0.32*

tes: There were 219,887 male and 50,081 female, and 212,238 White, 32,798 Black, and 12,647 Hispanic in USAF plicants. *d* was used to express group mean differences in standard deviation units. *$p \leq .05$: One-tailed t-tests were used test the difference between pairs of means. A Bonferroni approach was used to control for the experiment-wise error rate hin each group (males vs. females, Whites vs. Blacks, Whites vs. Hispanics). As a result, each t-test used a .0024 (.05/ 21 ts) Type I error rate.

ble 3. Means and Standard Deviations for AFOQT Scores and UPT Final Outcome by Sex and Ethnicity 'S Air Force Pilot Trainees)

	Sex					Ethnicity							
	Male		Female		*d*	White		Black		*d*	Hispanic		*d*
ore	Mean	SD	Mean	SD	M-F	Mean	SD	Mean	SD	W-B	Mean	SD	W-H
	75.02	16.71	72.03	15.83	0.17	75.33	16.49	64.12	19.71	0.67*	67.72	17.21	0.46*
Γ	70.15	18.86	66.38	18.79	0.20*	70.38	18.79	57.66	20.78	0.67*	61.86	19.70	0.45*
	60.48	22.92	71.67	21.92	-0.48	61.35	22.83	54.82	25.03	0.28*	53.13	23.98	0.36*
	62.63	21.78	63.79	22.07	-0.05	63.02	21.72	51.44	22.57	0.53*	55.72	22.02	0.33*
	62.27	22.03	69.42	21.43	-0.32	62.88	21.90	52.73	23.94	0.46*	54.09	22.37	0.40*
	15.59	3.01	17.31	2.79	-0.63	15.70	3.00	14.62	3.51	0.35*	14.54	3.20	0.38*
	13.53	3.51	13.41	3.69	0.03	13.58	3.50	11.87	3.46	0.48*	12.65	3.47	0.26*
	18.60	4.39	20.54	3.85	-0.42	18.75	4.33	17.07	5.00	0.38*	17.10	4.97	0.38*
	13.84	3.27	13.79	3.20	0.15	13.89	3.26	12.45	3.40	0.44*	12.74	3.19	0.35*
K	14.88	4.89	16.45	4.91	-0.32	14.99	4.88	14.43	5.30	0.11	13.92	4.95	0.21*
K	17.71	4.65	18.42	4.58	-0.15	17.78	4.66	15.98	4.56	0.38*	16.65	4.57	0.24*
	12.28	2.98	9.77	2.62	0.84*	12.27	2.98	10.82	3.21	0.48*	11.33	3.13	0.31*
I	10.26	4.14	8.54	3.60	0.41*	10.25	4.14	9.10	3.87	0.27*	9.69	3.84	0.13
	25.08	5.07	25.00	5.15	0.01	25.17	5.05	22.16	4.91	0.59*	23.00	5.19	0.42*
	14.95	3.68	13.29	3.72	0.45*	14.93	3.68	13.93	4.01	0.27*	14.54	3.77	0.10
	13.66	3.43	13.69	3.31	-0.01	13.70	3.41	12.15	3.76	0.45*	12.70	3.54	0.29*
	31.14	5.64	32.89	5.44	-0.30	31.24	5.63	29.84	6.25	0.24*	29.72	5.66	0.27*
	13.71	3.95	12.82	4.10	0.22*	13.72	3.94	12.98	4.23	0.18	13.42	4.27	0.07
	9.86	2.68	8.35	2.73	0.56*	9.85	2.68	8.96	2.91	0.33*	9.34	2.88	0.19
	10.41	3.29	8.85	2.77	0.47*	10.41	3.28	9.33	3.17	0.32*	9.97	3.56	0.13
	11.09	2.48	11.24	2.51	-0.06	11.10	2.48	10.76	2.64	0.13	10.56	2.64	0.21
T P/F	0.79	0.40	0.72	0.44	0.17*	0.80	0.39	0.72	0.45	0.20*	0.65	0.47	0.38*

tes: There were 9,239 males, 237 females, 8,955 Whites, 186 Blacks, and 172 Hispanics in the validation samples.
ference between means (*d*) were expressed in standard deviation units. *$p \leq .05$: One-tailed t-tests were used to test the
erence between means. A Bonferroni approach was used to control for the experiment-wise error rate within each group.
ch t-test used a .0024 Type I error rate. The t-test for the UPT P/F criterion also used a .05 Type I error rate.

The pilot trainees represent a range restricted sample as they were screened on several factors (i.e., physical and medical fitness, academic performance, aptitude test scores, previous flying experience). Range restriction causes observed correlations to be biased estimators. The Lawley (1943) correction for range restriction was applied within each sex and ethnic group to correct the mean, variance, and correlation estimates of the tests back to that particular group's applicant pool. The unrestricted estimates of the AFOQT means, variances, and correlations came from officer commissioning applicant records. The Lawley correction procedure estimates the means, variances, and correlations as they would be observed in the unrestricted population. No assumptions beyond linearity of regression form and homoscedasticity are required for the Lawley procedure. Our experience has shown (Olea & Ree, 1994) that violation of the homoscedasticity assumption such as found in dichotomous criteria, has a generally benign effect. The Thorndike case 2 correction (Thorndike, 1949) was applied to the AFOQT composites within each group. It was inappropriate to use the Lawley procedure on the set of composites due to linear dependence among the variables. Linear dependence prohibits matrix inversion and computation of the Lawley correction.

The degree of range restriction in the pilot training validation sample is exemplified by the effect of selection on the Pilot composite. The restriction in range for the Pilot composite was such that the variances in the sex and ethnic groups averaged only about 58 percent of the USAF applicants' variances. However, the degree of range restriction varied substantially across sex (males, 43% and females, 55%) and ethnic groups (Whites, 47%; Blacks, 130%; Hispanics, 60%). Note that there was an increase in variance for Black participants. While selection usually decreases variance, an increase in variance sometimes occurs (Levin, 1972).

The correlations of the composites with the pass-fail criterion varied by sample. In the observed data (Table 4), the Pilot, Navigator-Technical, Quantitative, and Academic Aptitude composites were significantly correlated with the criterion for males and Whites and the Pilot composite was significant for Hispanics. It is important to note that no group differences (male - female, Whites - Blacks, Whites - Hispanics) in correlations were observed between the AFOQT scores and UPT P/F. After correction for range restriction (Table 5), the correlations for the Pilot composite increased .054 on average across all groups and the negative correlation for the Verbal composite, though almost zero, remained negative.

On the test level, 12 of 16 observed correlations were significant for males and Whites with only VA, RC, WK (from the Verbal composite), and GS (also highly verbal in content) not showing significance. None of the 16 tests showed significant correlations for females, Blacks, or Hispanics. Tests of differences in correlations between the AFOQT scores and UPT P/F showed no differences in correlations for males and females, Whites and Blacks, and Whites and Hispanics.

Most of the correlations increased in value when they were corrected for range restriction. For males, the strongest test

predictor in both the uncorrected and corrected for range restriction form was the test of specialized flying knowledge, IC. For females, MK was most predictive for both the uncorrected and corrected for range

Table 4. Observed Correlations Between AFOQT Scores and UPT Outcome by Sex and Ethnic Group

Score	Sex			Ethnicity				
	Male	Female	M-F	White	Black	W-B	Hispanic	W-H
P	.155*	.143	.012	.147*	.082	.065	.224*	-.077
N-T	.129*	.166	-.037	.126*	.111	.015	.182	-.056
V	-.010	-.021	.011	-.020	.070	-.090	.050	-.070
Q	.095*	.149	-.054	.092*	.178	-.086	.108	-.016
AA	.044*	.072	-.028	.037*	.149	-.112	.084	-.047
VA	.010	-.084	.094	-.003	.039	-.042	.107	-.110
AR	.083*	.155	-.072	.081*	.168	-.087	.102	-.021
RC	-.006	.053	-.059	-.014	.101	-.115	.030	-.044
DI	.085*	.047	.038	.077*	.178	-.101	.165	-.088
WK	-.022	-.030	.008	-.029	.033	-.062	.036	-.065
MK	.067*	.156	-.089	.068*	.104	-.036	.032	.036
MC	.080*	.019	.061	.079*	.023	.056	.082	-.003
EM	.062*	.062	.000	.061*	.017	.044	.013	.048
SR	.109*	.133	-.024	.105*	.131	-.026	.105	.000
IC	.144*	.118	.026	.151*	.005	.146	.089	.062
BC	.074*	.089	-.015	.070*	.077	-.007	.159	-.089
TR	.091*	.038	.053	.085*	-.006	.091	.171	-.086
AI	.072*	.140	-.068	.073*	.064	.009	.164	-.091
RB	.071*	.066	.005	.073*	-.028	.101	.096	-.023
GS	.028	.118	-.090	.028	.089	-.061	.091	-.063
HF	.050*	.095	-.045	.055*	-.077	.132	.066	-.011
Mean Pilot 8	.080	.064	.016	.078	.044	.034	.094	-.016
Mean AFOQT 16	.062	.073	-.009	.060	.057	.003	.111	-.051

Notes. "Mean Pilot 8" is the mean correlation of the 8 tests in the Pilot composite with UPT P/F. "Mean AFOQT 16" is the mean correlation of all 16 tests with UPT P/F. Neither Mean Pilot 8 nor Mean AFOQT 16 were tested for statistical significance. There were 9,239 males, 237 females, 8,955 Whites, 186 Blacks, and 172 Hispanics. *$p \leq .05$: One-tailed t-tests were used to test the correlations within a group ($r > 0$) and the difference between pairs of correlations ($r_1 - r_2 > 0$). A Bonferroni approach was used to control the experiment-wise error rate within each group of tests. Each t-test used a .0024 (.05/ 21 tests) Type I error rate.

Table 5. Corrected Correlations Between AFOQT Scores and UPT Final Outcome by Sex and Ethnic Group

	Sex			Ethnicity				
Score	Male	Female	M-F	White	Black	W-B	Hispanic	W-H
P	.227	.192	.035	.211	.072	.139	.283	-.072
N-T	.180	.206	-.026	.168	.098	.070	.212	-.044
V	-.012	-.027	.015	-.023	.063	-.086	.052	-.075
Q	.118	.169	-.051	.110	.158	-.048	.116	-.006
AA	.055	.085	-.030	.043	.125	-.082	.088	-.045
VA	.106	-.020	.126	.059	.105	-.046	.286	-.227
AR	.172	.165	.007	.157	.234	-.077	.256	-.099
RC	.067	.065	.002	.022	.127	-.105	.182	-.160
DI	.169	.076	.093	.144	.226	-.082	.318	-.174
WK	.037	-.034	.061	-.004	.093	-.097	.172	-.176
MK	.161	.225	-.064	.148	.195	-.047	.172	-.024
MC	.156	.050	.106	.149	.036	.113	.174	-.025
EM	.130	.096	.034	.125	.081	.044	.110	.015
SR	.200	.171	.029	.185	.201	-.016	.243	-.058
IC	.230	.130	.100	.232	.111	.121	.217	.015
BC	.171	.164	.007	.152	.152	.000	.266	-.114
TR	.169	.089	.080	.141	.054	.087	.263	-.122
AI	.132	.136	-.004	.128	.097	.031	.232	-.104
RB	.155	.137	.018	.152	.045	.107	.200	-.048
GS	.107	.124	-.017	.098	.157	-.059	.218	-.120
HF	.124	.155	-.031	.115	-.058	.173	.164	-.049
Mean Pilot 8	.162	.102	.060	.146	.105	.041	.224	-.078
Mean AFOQT 16	.143	.108	.035	.125	.116	.009	.217	-.092

Note. "Mean Pilot 8" is the mean correlation of the 8 AFOQT tests in the Pilot composite with UPT P/F. "Mean AFOQT 16" is the mean correlation of all 16 AFOQT tests with UPT P/F. Observed correlations (see Table 4) were corrected for range restriction within each sex and ethnic group. The sample sizes for the reference groups were 219,887 males, 50,081 females, 212,238 Whites, 32,798 Blacks, and 12,647 Hispanics. Corrected correlations were not tested for significance.

restriction correlations. The average uncorrected and corrected for range restriction correlations of the 8 tests in the Pilot composite was .080 and .162 for males and .064 and .102 for females. The average uncorrected and corrected for range restriction correlations of all 16 tests was .062 and .143 for males and .073 and .108 for females.

It should be noted that the correlations did not vary very much across the ethnic groups. The ranges of uncorrected and corrected for range restriction average correlations for the 8 tests in the Pilot composite were respectively, Whites (.078 and .146), Blacks (.044 and .105), and Hispanics (.094 and .224). The same ranges for all 16 tests uncorrected and corrected for range restriction were: Whites (.060 and .125), Blacks (.057 and .116), and Hispanics (.111 and .217). The correlation matrices of the 16 tests and UPT P/F for each sex and ethnic group are available by request from the author.

Regression Analyses

All regressions used observed correlations. Each of the 5 composites was tested to investigate bias for sex and ethnic groups in the prediction of UPT P/F (Cleary, 1968; Jensen, 1980). Stauffer and Ree (1996) have explained the use of linear versus logistic regression (LOGR) with dichotomous criteria, noting that there are instances when the linear probability model (LPM) is preferable. In general, parameters in the LOGR model are more efficient than those in the LPM. However, Tatsuoka (1988, p. 228) notes that when multivariate normality holds, LPM is more efficient than LOGR. Additionally, the coefficients of LPM are easily interpretable. For theoretical and practical reasons, LPM was used.

Regression models were estimated separately for each sex and ethnic group. Estimates of regression intercept, slope, and standard error of estimate were obtained for each group. A test of the equality of the variance error of estimate (SE^2_{est}) for the majority and minority groups (e.g., males and females) was done to determine whether the groups being compared had equal SE^2_{est}s. The test is the ratio of the larger SE^2_{est} divided by the smaller SE^2_{est} and is distributed as F (Jensen, 1980; Reynolds, 1982). If the SE^2_{est}s for the two groups are equal, linear models may be used to test the equality of the regression slopes and intercepts (Cleary, 1968; Jensen, 1980). If the SE^2_{est}s of the regression lines are not equal, some argue (Linn, 1973) that testing linear models is inappropriate. However, others (Hunter & Schmidt, 1976) do not consider the testing of variance errors of estimate and conclude "that any purely statistical approach to the problem of test bias is doomed to rather immediate failure. ...Furthermore, even among those who agree on values there will be disagreements about the validity of certain relevant scientific theories..." (p. 1069). Given this professional disagreement, linear models were tested and presented.

The testing of linear models involved comparing a "full model" to a "restricted model" that contained a subset of the variables from the full model. An F statistic was used to evaluate the change in predictive efficiency between the full and restricted models using the hierarchical step-down method of Lautenschlager and Mendoza (1986). The starting (full) model (Model 1) for each analysis contained separate

estimates for the slopes and intercepts for the two groups (males vs. females, Whites vs. Blacks, Whites vs. Hispanics). The first restricted model (Model 2) removed the separate slope estimates, and the second restricted model (Model 3) removed the separate intercept estimates. First, each AFOQT composite was tested for slope bias. If evidence of slope bias was found, the analysis sequence was terminated. If no slope bias was found, the composite was tested for difference in intercepts.

Regressions of UPT P/F on each of the composites were done within each sex and ethnic group to obtain the SE^2_{est} values. Comparisons of the SE^2_{est} values of the regression equations for males versus females, Whites versus Blacks, and Whites versus Hispanics showed that the SE^2_{est} were not equal. In all instances the SE^2_{est} values were larger for the minority group. The difference in the SE^2_{est}s are well characterized as .04 to .06 (i.e., .20 vs. .16 for females and males; .20 vs. .16 for Blacks and Whites; .22 vs. .16 for Hispanics and Whites). Humphreys (1986) has suggested that this may be a consequence of heteroscedastic error.

The combined-group linear models are presented in Table 6. Examination of the models indicated that the Pilot, Navigator-Technical, Quantitative, and Academic Aptitude composites were predictive of UPT P/F, but the Verbal composite was not. As a result, no tests of differential slopes or intercepts were done for Verbal. Comparisons of Model 1 versus. Model 2 for the other AFOQT composites indicated there were no group differences in slopes. Comparisons of Model 2 and Model 3 showed some significant intercept differences. In all instances where intercept differences occurred, performance was overpredicted for the minority group. For the sex group comparisons, UPT P/F was overpredicted for females for the Quantitative and Academic Aptitude composites. For the ethnic group comparisons, no intercept differences were found for Whites versus Blacks. However, UPT P/F was overpredicted for Hispanics relative to Whites for the Pilot, Navigator-Technical, Quantitative, and Academic Aptitude composites.

Discussion

The results showed that the composites and tests were not biased against females or ethnic minorities. Group mean differences in test scores for officer applicants and pilot trainees may be due to differential attraction of the military and the job of pilot for members of the groups. The effects of differential attraction should be considered when drawing conclusions about group differences.

Pilot samples resulting from differential attraction led to group differences in range restriction and statistical power. The female, Black, and Hispanic samples were most severely affected. Statistical power is a joint function of sample size, effect size, directionality of the hypothesis (one- vs. two-tail), and Type I error rate. The lack of statistical power hampers studies. Small sample sizes played a role in the inability to detect significant correlations for the female and ethnic minority group samples. Using the Bonferroni inequality to control the experiment-wise error rate also reduced statistical power.

Dichotomization of the criterion (Cohen, 1983) and range restriction (Hunter &

Schmidt, 1990) also reduced statistical power. Although other flying criteria were available (e.g., academic grades, flying training work samples), a large portion of the participants lacked these data, and as a result would have been removed from the study. The loss of minority participants (had

Table 6. Correlations, Squared Multiple Correlations, and F Tests for the Models

Model/Groups/Score	R	R^2	df_1	df_2	F	ΔR^2	$F_{\Delta R^2}$
Males vs. Females							
Pilot							
M1	.15281	.02335	3	9454	75.34*		
M2	.15279	.02334	2	9455	113.00*	.00001	0.05
M3	.15095	.02278	1	9456	220.48*	.00056	5.34
Navigator-Technical							
M1	.13307	.01771	3	9454	56.80*		
M2	.13277	.01763	2	9455	84.83*	.00008	0.76
M3	.13061	.01706	1	9456	164.11*	.00160	5.42
Verbal							
M1	.02991	.00089	3	9454	2.82		
M2	.02984	.00089	2	9455	4.21	.00000	not tested
M3	.01258	.00016	1	9456	1.49	.00073	not tested
Quantitative							
M1	.10107	.01021	3	9454	32.52*		
M2	.10049	.01010	2	9455	48.22*	.00011	1.11
M3	.09628	.00927	1	9456	88.47*	.00083	7.87*
Academic Aptitude							
M1	.05338	.00285	3	9454	9.00*		
M2	.05307	.00282	2	9455	13.35*	.00003	0.31
M3	.04367	.00191	1	9456	18.07*	.00091	8.61*
Whites vs. Blacks							
Pilot							
M1	.14849	.02205	3	9137	68.67*		
M2	.14804	.02191	2	9138	102.37*	.00014	1.23
M3	.14733	.02171	1	9139	202.76*	.00020	1.93
Navigator-Technical							
M1	.12868	.01656	3	9137	51.28*		
M2	.12867	.01656	2	9138	76.91*	.00000	0.02
M3	.12762	.01629	1	9139	151.30*	.00027	2.47
Verbal							
M1	.03629	.00132	3	9137	4.01*		
M2	.03344	.00112	2	9138	5.11*	.00020	not tested
M3	.01641	.00027	1	9139	2.46	.00085	not tested

Quantitative							
M1	.09881	.00976	3	9137	30.02*		
M2	.09770	.00954	2	9138	44.03*	.00022	2.00
M3	.09534	.00909	1	9139	83.83*	.00045	4.18
Academic Aptitude							
M1	.05175	.00268	3	9137	8.17*		
M2	.04853	.00236	2	9138	10.78*	.00032	2.95
M3	.04110	.00169	1	9139	15.46*	.00067	6.09
Whites vs. Hispanics							
Pilot							
M1	.15724	.02473	3	9123	77.09*		
M2	.15650	.02449	2	9124	114.53*	.00024	2.14
M3	.15145	.02294	1	9125	214.21*	.00155	14.36*
Navigator-Technical							
M1	.13659	.01866	3	9123	57.81*		
M2	.13611	.01853	2	9124	86.10*	.00013	1.20
M3	.12980	.01685	1	9125	156.38*	.00168	15.45*
Verbal							
M1	.05323	.00283	3	9123	8.64*		
M2	.05212	.00272	2	9124	12.42*	.00011	not tested
M3	.01576	.00025	1	9125	2.26	.00247	not tested
Quantitative							
M1	.10402	.01082	3	9123	33.26*		
M2	.10390	.01080	2	9124	49.78*	.00012	0.22
M3	.09382	.00880	1	9125	81.03*	.00200	18.28*
Academic Aptitude							
M1	.06222	.00387	3	9123	11.82*		
M2	.06164	.00380	2	9124	17.39*	.00007	0.65
M3	.04021	.00162	1	9125	14.77*	.00218	19.95*

Note. ΔR^2 is the change in the squared multiple correlation from the previous model to the current model (i.e., Model 1 vs. Model 2, Model 2 vs. Model 3).

these other flying performance criteria been used) would have led to even lower statistical power.

Mean Differences

Officer and pilot selection procedures acted to reduce, but not eliminate, group differences in mean scores. Officer commissioning and pilot selection regulations set minimum scores for the AFOQT and the selection boards use top-down selection. Mean score differences were smaller among pilot trainees than among officer applicants.

Male-female mean differences were consistent with earlier studies (Burke, 1995; Hyde, 1981; Jensen, 1980). In a meta-analysis of sex differences on pilot aptitude tests, Burke (1995) noted relatively small mean differences on verbal tests (-.1 *d* favoring women), with larger differences on quantitative (.5 *d* favoring men) and spatial tests (.5 *d* favoring men). Further, Burke noted that the magnitude of the male-female differences within these broad ability categories (verbal, quantitative, and spatial) varied by specific test content.

The means of the composites and tests for the pilot trainees revealed some interesting findings. All mean scores in the applicant sample favored males, but in several instances female pilot trainees had higher means than male trainees. On the Verbal, Quantitative, and Academic Aptitude composites, female means exceeded those for males (d = -.49, -.06, and -.32 respectively). The biggest mean difference on the composite level was found for Verbal, but the most surprising mean difference favoring females was on Quantitative. This appeared to be the result of prior selection. In most studies, female means on quantitative measures are less than male means (Burke, 1995; Jensen, 1980).

The ethnic group comparisons were consistent with previous studies (Brody, 1992; Coleman, et al., 1966; Jensen, 1973; Loehlin et al., 1975; Mathews, 1977). For instance, Jensen (1973) and Loehlin et al. (1975) reported White-Black differences in intelligence and aptitude test scores of approximately one standard deviation. These values are similar to those in the USAF applicant samples. As with the sex groups, ethnic group mean differences were reduced in the selection process.

Female and ethnic minority applicants were less likely to reach or exceed minimum scores on the AFOQT. To the extent that group differences on mean test scores occur (see also Burke, 1995 and Hyde, 1981), the potential for adverse impact exists. It is possible that well qualified females and ethnic minorities are less inclined to view the Air Force as an attractive career choice. Another possibility is that females and ethnic minorities are less likely to take courses or pursue leisure interests that might increase their performance on the AFOQT.

One method to reduce mean differences might be to make information regarding test content readily available. Examples of AFOQT test content are provided in free information pamphlets. Those interested in applying for pilot training can easily determine test content and adopt an appropriate preparation strategy. However, some of the tests in the Pilot composite rely on special flying job knowledge (e.g., Instrument Comprehension, Aviation Information) that is not readily available or may require a substantial financial investment on the part of the applicant (e.g., enrolling in an aircraft training course). This

may not be feasible for applicants that are economically disadvantaged.

Another method might be to replace tests with large mean differences with others less susceptible to group mean differences (e.g., chronometric measures, cognitive components with novel content). The validity of tests based on cognitive components for predicting pilot training (Carretta, 1996) and operational pilot criteria (Carretta, Perry, & Ree, 1996) has been shown. The effects of replacement remain speculative, pending group mean differences studies.

Correlations

Although the observed correlations of the composites and tests with the criterion appeared to differ between groups, statistical testing showed them not to differ. When the correlations were corrected for range restriction, most increased in value. The strongest test predictor for males, both uncorrected and corrected for range restriction, was IC, the instrument comprehension test of specialized knowledge. For females, MK was most predictive among both the uncorrected and corrected for range restriction correlations. It is worth noting that the most predictive tests for females were based on information taught in the usual educational curriculum and were mostly measures of general cognitive ability. In contrast, the content in the most predictive test for males is not likely to be learned in the usual curriculum. IC content may be learned from aviation books or magazines, or in specific aviation courses. The gathering of this specialized knowledge while certainly a function of ability, may also be a function of interest, opportunity, and motivation. Comparison of mean test scores for USAF applicants suggests that women are much less likely to acquire the specialized aviation knowledge found in IC and AI. Further, this situation appears to be analogous to the observation of male-female mean differences in the technical knowledge tests in the Armed Services Vocational Aptitude Battery as noted by Ree and Carretta (1995) who observed that those differences parallel differences in technical course enrollment for the sexes.

The observed correlations between the AFOQT scores and the criterion were lower than might be expected. There exists several potential reasons for this. Among these reasons are sampling variability, range restriction, unreliability of the predictors, unreliability of the criterion, and dichotomization of the criterion. On average, correcting for range restriction increased correlations with the criterion by about .05. The other study artifacts (unreliability of the predictors, unreliability of the criterion, and dichotomization of the criterion) were not corrected for. This is left to future meta-analysts.

Regression Analyses

The regression analyses were consistent with previous studies of sex (Carretta, 1990; Siem & Sawin, 1990) and White-Black (Mathews, 1977) differences in USAF pilot selection factors. When group differences in expected pilot training completion rate were observed, performance was overestimated for the minority group. Therefore, no prediction bias was observed against the minority groups.

Interpretation of regression intercepts is hazardous when the predictor is not perfectly reliable. Jensen (1980) has

presented the clearest explanation. See also Aiken and West (1991) and Fuller, (1987) for a general description of regression with variables containing measurement error. Jensen demonstrated where there are two groups and a single regression line, there must be two intercepts found solely on the basis of the unreliability of the predictors. He provided the following equation for the expected difference in intercepts as a function of group means, regression coefficient, and predictor reliability:

$\Delta(k_A - k_B) = b_{yx}(1 - r_{xx})(\text{Mean}_A - \text{Mean}_B)$

where k_A and k_B are the intercepts for groups A and B and Mean_A and Mean_B are the means. Further, b_{yx} is the raw score regression coefficient for the regression of y on x and r_{xx} is the reliability of predictor x.

For example, if the regression coefficient were 1 and the Mean_A and Mean_B were 10 and 5 for a test (x) with reliability of .8, the expected difference in intercepts would be 1.0. If the reliability were increased to .9, the expected difference in intercepts would decrease to .5. Conversely, if the reliability decreases to .5, the expected intercept difference increases to 2.5. Contrasting this to the circumstance where reliability is perfect and a zero difference in intercepts is found, the nature and magnitude of the artifact is made clear. The uncritical interpretation of different intercepts as bias is unwarranted.

When group differences in prediction were examined for Models 2 and 3, in some instances there was overprediction for females (.07) and Hispanics (.12). After correction for unreliability of the predictors (Jensen, 1980, p. 384), all differences were reduced to a trivial .0004 or less.

Conclusion

Although group mean differences were found in composite and test scores, no differences were found in validity. The AFOQT composites and tests were not biased against females and ethnic minorities when used for pilot training selection. This study should be repeated when larger samples of females and minority group members have been accumulated. The accumulation of these samples promises to be a long term effort, given the rate of female and minority participation in pilot training. Meta-analyses (Hunter & Schmidt, 1990) or other aggregating approaches should be considered.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Arth, T. O. (1986). *Air Force Officer Qualifying Test (AFOQT) retesting effects* (AFHRL-TP-86-8). Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Arth, T. O., Steuck, K. W., Sorrentino, C. T., & Burke, E. F. (1990). *Air Force Officer Qualifying Test (AFOQT): Predictors of undergraduate pilot training and undergraduate navigator training success* (AFHRL-TP-89-52). Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Brody, N. (1992). *Intelligence: Nature, determination, and consequence.* San Diego: Academic Press.

Burke, E. F. (1995). Male-female differences on aviation selection tests: Their implications for research and practice. In

N. Johnston, R. Fuller, & N. McDonald (Eds.), *Aviation Psychology: Training and Selection* (pp. 188-193). Aldershot, England: Avebury Aviation.

Carretta, T. R. (1990, April). *Gender differences in USAF pilot training performance*. Paper presented at the 12th Symposium on Psychology in the Department of Defense, Colorado Springs, CO.

Carretta, T. R. (1992). Recent developments in U. S. Air Force pilot candidate selection and classification. *Aviation Space and Environmental Medicine, 63,* 1112-1114.

Carretta, T. R. (1996). *Preliminary validation of several US Air Force computer-basedcognitive pilot selection tests* (AL/HR-TP-1996-0008). Brooks AFB, TX: Armstrong Laboratory Human Resources Directorate, Manpower and Personnel Research Division.

Carretta, T. R., Perry, D. C., Jr., & Ree, M. J. (1996). Prediction of situational awareness in F-15 pilots. *The International Journal of Aviation Psychology, 6,* 21-41.

Carretta, T. R., & Ree, M. J. (1995). Air Force Officer Qualifying Test validity for predicting pilot training performance. *Journal of Business and Psychology, 9,* 379-388.

Carretta, T. R., & Ree, M. J. (1996). Factor structure of the Air Force Officer Qualifying Test: Analysis and comparison. *Military Psychology, 8,* 29-42.

Cleary, T. A. (1968). Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement, 5,* 114-124.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7,* 249-253.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Coleman, J. S., et al. (1966). *Equality of educational opportunity survey.* Washington DC: National Center for Educational Statistics.

Dye, D. A., Reck, M., & McDaniel, M. (1993). The validity of job knowledge measures. *International Journal of Selection and Assessment, 1,* 153-162.

Fuller, W. A. (1987). *Measurement error models*. New York: Wiley.

Glass, G. V., & Stanley, J. C. (1970). *Statistical methods in education and psychology.* Englewood Cliffs, NJ: Prentice-Hall.

Hartigan, J. A., & Wigdor, A. K. (Eds.) (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery*. Washington, DC: National Academy Press.

Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. *Journal of Applied Psychology, 71,* 327-333.

Hunter, J. E., & Schmidt, F. S. (1976). Critical analysis of the statistical and ethical implications of various definitions of test bias. *Psychological Bulletin, 83,* 1053-1071.

Hunter, J. E., & Schmidt, F. S. (1990). *Methods of meta-analysis.* Newbury Park, CA: Sage.

Hyde, J. S. (1981). How large are cognitive gender differences? A meta-analysis using ω^2 and *d*. *American Psychologist, 36*, 892-901.

Jaeger, R. M. (Ed.) (1976). On bias in

selection. *Journal of Educational Measurement, 13*, 1-99.

Jensen, A. R. (1973). *Educability and group differences*. London: Metheun.

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

Lautenschlager, G. J., & Mendoza, J. (1986). A step-down hierarchical multiple regression analysis for estimating hypotheses about test bias in prediction. *Applied Psychological Measurement, 10,* 133-159.

Lawley, D. N. (1943). A note on Karl Pearson's selection formulae. *Proceedings of the Royal Society of Edinburgh, Section A, 62, Part 1,* 28-30.

Levin, J. (1972). The occurrence of an increase in correlation by range restriction. *Psychometrika, 37,* 93-97.

Linn, R. (1973). Fair test use in selection. *Review of Educational Research, 43,* 139-161.

Linn, R. (1982). Ability testing: Individual differences, prediction, and differential prediction. In A. K. Wigdor & W. R. Garner (Eds.), *Ability testing: Uses, consequences, and controversies: Part II* (pp. 335-388). Washington, DC: National Academy Press.

Loehlin, J. C., Lindzey, J. N., & Spuhler, N. (1975). *Race differences in intelligence*. San Francisco: Freeman.

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences*. Stanford, CA: Stanford University Press.

Mathews, J. J. (1977). *Racial equity in selection in Air Force Officer Training School and undergraduate flying training* (AFHRL-TR-77-22). Lackland AFB, TX: Air Force Human Resources Laboratory, Personnel Research Division.

McNemar, Q. (1969). *Psychological statistics* (4th ed.). New York: Wiley.

Miller, R. G., Jr. (1981). *Simultaneous statistical inference*. New York: Springer-Verlag.

Olea, M., & Ree, M. J. (1994). Predicting pilot and navigator criteria: Not much more than *g*. *Journal of Applied Psychology, 79,* 845-851.

Ree, M. J., & Carretta, T. R. (1995). Group differences in aptitude factor structure on the ASVAB. *Educational and Psychological Measurement, 55,* 268-277.

Reynolds, C. E. (1982). Methods for detecting construct and predictive bias. In R. A. Berk (Ed.) *Handbook of methods for detecting test bias,* Johns Hopkins University Press, Baltimore, MD,199-227.

Roberts, H. E., & Skinner, J. (1996). Gender and racial equity of the Air Force Office Qualifying Test in officer training school selection decisions. *Military Psychology,* 8, 95-113.

Siem, F. M., & Sawin, L. L. (1990, April). *Comparison of male and female USAF pilot candidates.* Paper presented at the AGARD Symposium on Recruitment, Selection, Training, and Military Operations of Female Aircrew, Tours, France.

Skinner, J., & Ree, M. J. (1987). *Air Force Officer Qualifying Test (AFOQT): Item andfactor analysis of form O* (AFHRL-TR-86-68). Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Stauffer, J. M., & Ree, M. J. (1996). Predicting with logistic or linear regression: Will it make a difference in who is selected for pilot training? *The*

International Journal of Aviation Psychology, 6, 233-240.

Tatsuoka, M. M. (1988). *Multivariate analysis: Techniques for educational and psychological research*. New York: Macmillan.

Thorndike, R. L. (1949). *Personnel selection.* New York: Wiley.

Wing, H. (1980). Profiles of cognitve ability of different racial/ethnic and sex groups on a multiple abilities test battery. *Journal of Applied Psychology, 63,* 289-29.