DOT/FAA/AM-99/16

Office of Aviation Medicine
Washington, D.C. 20591

Differential Prediction of FAA Academy Performance on the Basis of Race and Written Air Traffic Control Specialist Aptitude Test Scores

Dana Broach
William L. Farmer
Willie C. Young
Federal Aviation Administration
Civil Aeromedical Institute
Oklahoma City, Oklahoma 73125

19990601126

May 1999

Final Report

This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161.

U.S. Department of Transportation
Federal Aviation Administration

DTIC QUALITY INSPECTED 4

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents or use thereof.

Technical Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/AM-99/16		

4. Title and Subtitle	5. Report Date
Differential Prediction of FAA Academy Performance on the Basis of Race and Written Air Traffic Control Specialist Aptitude Test Scores	May 1999
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
Broach, D., Farmer, W. L., & Young, W.C.	
9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
FAA Civil Aeromedical Institute P.O. Box 25082 Oklahoma City, OK 73125-0082	11. Contract or Grant No.
12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Office of Aviation Medicine Federal Aviation Administration 800 Independence Avenue, S. W. Washington, DC 20591	

15. Supplementary Notes

This research was conducted under task AM-98-B-HRR-509.

16. Abstract

The written air traffic control specialist (ATCS) aptitude test battery was evaluated for evidence of predictive bias within the framework of the *Uniform Guidelines on Employee Selection Procedures* (29 CFR 1607) in a retrospective analysis. Step-down hierarchical regression analysis (Lautenschlager & Mendoza, 1986) was used to investigate differential prediction of performance in initial ATCS training at the Federal Aviation Administration (FAA) Academy in a sample of 282 African-American and 8,542 white first-time competitive entrants. Analysis based on correlations without corrections for restriction in range found significant differences in the intercepts, but not slopes, for African Americans and whites. Analysis based on correlations, corrected for explicit and implicit restriction in range, found significant differences in slopes and intercepts by race, suggesting that separate regression equations were appropriate to predict Academy performance for the groups. The two analyses indicated that the composite score on the written ATCS test battery exhibited predictive bias as defined by the *Uniform Guidelines on Employee Selection Procedures* (29 CFR 1607) and Cleary (1968). Specifically, the composite score TMC over-predicted the performance of African Americans in initial training at the FAA Academy. As a consequence of the over-prediction, significantly more of the African Americans that were accepted into training for the ATCS occupation on the basis of their aptitude test scores went on to fail training than would have been expected on the basis of the common or majority (white) regression line. An alternative explanation is considered that the observed differential prediction reflected criterion bias or other group differences in factors such as educational achievement and age. A path analytic approach is outlined for investigating the complex interactions between test score, the criterion, race, education, and age. Additional research on the consequences of over-prediction for African Americans in the FAA Academy is recommended in closing.

17. Key Words	18. Distribution Statement
ATCS selection, Test Bias, Adverse Impact, Fairness, Equal Employment Opportunity, Personnel Selection, Predictive Bias	Document is available to the public through the National Technical Information Service, Springfield, Virginia 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
UNCLASSIFIED	UNCLASSIFIED	28	

Form DOT F 1700.7 (8-72)

Reproduction of completed page authorized

Differential Prediction of FAA Academy Performance on the Basis of Race and Written Air Traffic Control Specialist Aptitude Test Scores

The Federal Aviation Administration (FAA) is committed to attracting, retaining, developing, and managing a productive and skilled work force that visibly reflects the nation's diversity (FAA, 1993, 1998). Achieving this goal will require substantial changes in the demographic profile of the Air Traffic Control Specialist (ATCS) workforce, the single largest (17,000) and most publicly visible occupational group in the agency. Air traffic control is a career field in which minority workers have been historically under-represented. From 1981 through February 1992, entry into the occupation was determined by applicant performance on a written aptitude test battery (Aul, 1991). This test battery emphasized the organization, definition, and manipulation of the perceptual field through verbal and numeric reasoning (Harris, 1986). Our purpose in this paper was to examine the fairness of the Office of Personnel Management (OPM) written ATCS aptitude test battery as the first step toward assessing to what degree, if any, the battery may have served as an "engine of exclusion" (Seymour, 1988) of minorities from the ATCS occupation.

By fairness, we are explicitly referring to the regression model of test bias (also referred to as the "Cleary model," "predictive bias," and "differential prediction") for which there is a reasonable professional consensus, as embodied in the current *Standards for Educational and Psychological Testing* ("*Standards*"; American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985) and the *Principles for the Validation and Use of Personnel Selection Procedures* ("*Principles*"; Society for Industrial and Organizational Psychology (SIOP), 1987, p. 18). As noted by Sackett (1996), it is important to differentiate between predictive bias as a technical characteristic of the use of a test score in a particular setting and fairness as a value judgment about the pattern of outcomes arising from use of the test score. Our focus in this paper is specifically on predictive bias as a technical characteristic of a composite score used for competitive selection of applicants into the ATCS occupation between 1981 and 1992.

Under the *Uniform Guidelines on Employee Selection Procedures* ("*Uniform Guidelines;*" 29 CFR 1607, Equal Employment Opportunity Commission, 1978), an investigation of predictive bias encompasses two issues. First, the impact on protected groups arising from use of a particular cut score on the predictor, must be evaluated. A selection rate for any protected group that is less than four-fifths (4/5 or 80%) of that of the majority group will "...generally be regarded by the Federal enforcement agencies as evidence of adverse impact" (29 CFR 1607. 14B.(8).(b)). Second, where use of a selection procedure results in adverse impact, the *Uniform Guidelines* require that the user of the test evaluate the degree to which differential predictions of future job performance are made from selection test scores by subgroup (29 CFR 1607.14.B.(8).(b)). A test exhibits predictive bias under the *Uniform Guidelines* if "... members of one race, sex, or ethnic group characteristically obtain lower scores on a selection procedure than members of another group, and the differences in scores are not reflected in differences in a measure of job performance" (29 CFR 1607.14.B.(8).(a)). In other words, a test demonstrates predictive bias "... if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup" (Cleary, 1968, p. 115).

A formal analysis of the controller aptitude test battery's impact on minority applicants was not technically feasible, as racial identifiers were not collected from ATCS job candidates at the time of testing. However, previous research reported significant differences in pass rates between whites and African Americans (Rock, Dailey, Ozur, Boone, & Pickrel, 1984a). In view of this previous research, an investigation of the relationship between test scores and job performance for evidence of differential prediction, in accordance with the *Uniform Guidelines* (29 CFR

1607.14.B.(8).(b)) and *Standards for Educational and Psychological Testing* (Standards 1.20, 1.21, 1.22, p. 17) was warranted. Our formal null hypothesis was no difference in the predictive validity of the test battery by race.

METHOD

Sample

Archival test data were available for a sample of 170,578 (42.2%) persons from the 403,997 job applicants who took the written ATCS aptitude test battery from 1981 through 1992. Between October 1985 and January 1992, 14,392 persons entered as students into initial controller training at the FAA Academy. Most (79.2%) of these new hires had competed under civil service regulations and were entering the Academy for the first time. The differential prediction analysis was based on the 8,824 students with full and complete racial identification, predictor, and criterion data. There were 8,542 (96.8%) whites and 282 (3.2%) African Americans in the research sample. Demographic information for the applicant sample, all 1986-1992 FAA Academy entrants, and the research sample are presented in Table 1.

Measures

Predictors. The written ATCS aptitude test battery was the initial hurdle in the ATCS selection process, and consisted of three tests: (a) the Multiplex Controller Aptitude Test (MCAT); (b) the Abstract Reasoning Test (ABSR); and (c) the Occupational Knowledge Test (OKT). The MCAT was a timed paper-and-pencil civil service test (OPM test No. 510) simulating activities required for control of air traffic. Aircraft locations and direction of flight were indicated with graphic symbols on a simplified, simulated radar display (Figure 1). An accompanying table provided relevant information required to answer the item, including aircraft altitudes, speeds, and planned routes of flight. MCAT test items required examinees to identify situations resulting in conflicts between aircraft, to solve time, speed, and distance problems and also to interpret the tabular and graphical information. The ABSR was a 50-item civil service examination (OPM test No. 157). To solve an item, examinees determined what relationships existed within sets of symbols or letters. The examinee then identified the next symbol or letter in the progression, or the element missing from the set. A sample ABSR item is presented in Figure 2. The OKT was an 80-item job knowledge test that contained items related to seven knowledge domains relevant to aviation generally, and to air traffic control phraseology and procedures, specifically. The OKT was developed as an alternative to self-reports of aviation and air traffic control experience. The OKT was found to be more predictive of performance in ATCS training than self-reports (Dailey & Pickrel, 1984; Lewis, 1978).

The development of the written ATCS aptitude test battery has been extensively described elsewhere (Brokaw, 1984; Collins, Boone, & VanDeventer, 1984; Manning, 1991; Sells, Dailey, & Pickrel, 1984). The test-retest correlation for the MCAT was estimated at .60 in a sample of 617 newly-hired controllers (Rock, Dailey, Ozur, Boone, & Pickrel, 1982, p. 59). Parallel form reliability, as computed on the same sample, ranged from .42 to .89 for various combinations of items (Rock et al., p. 103). Lilienthal and Pettyjohn (1981) examined internal consistency and item difficulties for ten versions of the MCAT. Cronbach's alpha for the ten versions ranged from .63 to .93; the alphas for 7 of the 10 versions were greater than .80. The available data suggest that the MCAT had acceptable reliability. In contrast, no item analyses, parallel form, test-retest, or internal consistency estimates of the ABSR test have been reported.

Scoring of the test battery was done initially by summing the MCAT and ABSR scores, as shown in Table 2. The resulting total weighted score (TWS) was then transformed to a score with a mean of 70 and maximum of 100, known as the Transmuted Composite Score (TMC). About half of all applicants were expected to score at or above the mean (Rock, Dailey, Ozur, Boone, & Pickrel, 1984b). Applicants with three years of general work experience, four years of college, or any combination of education and experience equivalent to three years of general experience, needed a TMC score of 75.1 to be considered for employment. Applicants with one year of graduate study, superior academic achievement, or specialized aviation or air traffic control experience, required a TMC score of 70 to qualify for employment consideration (Aul, 1991). Applicants not meeting these criteria were ineligible for consideration for employment as controllers.

Table 1
Demographic characteristics for applicant sample, all 1985-1992 FAA Academy entrants, and the research sample

Characteristic	Applicant sample[a] (N = 170,578)	FAA Academy entrants	
		FAA Academy entrants (N=14,392)	Research sample (N=8,824)
Race			
White		12,366	8,542
African American		819	282
Other[b]		811	
Missing		396	
Education			
< High School	404		
High School	28,147	1,576	969
Some college	82,414	7,750	4,928
Bachelor's degree	54,583	4,745	2,818
Advanced degree	3,934	176	109
Missing	1,096	145	
Age			
Mean		26.21	25.78
SD		4.90	2.86

Notes: [a]Racial identification and age data not available for applicant sample.
[b]Other includes American Indian/Alaskan Native, Asian/Pacific Islander, and Hispanic.

AIRCRAFT	ALTITUDE	SPEED	ROUTE
10	7000	480	AGKHC
20	7000	480	BGJE
30	7000	240	AGJE
40	6500	240	CHKJF
50	6500	240	DIKGB
60	8000	480	DIKJE
70	8000	480	FJKID

SAMPLE QUESTION

WHICH AIRCRAFT WILL CONFLICT?
A. 60 AND 70
B. 40 AND 70
C. 20 AND 30
D. NONE OF THESE

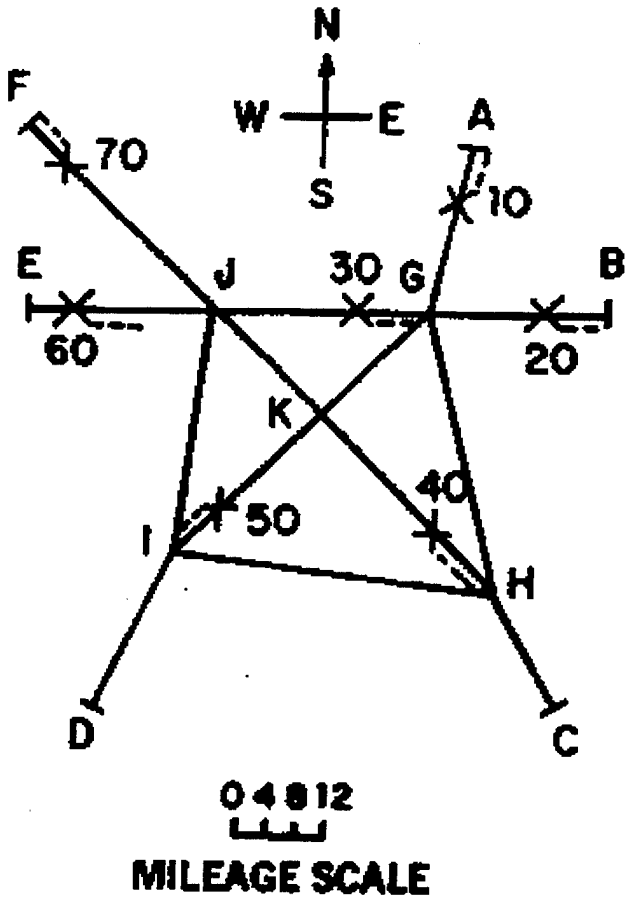

Figure 1: Example Multiplex Controller Aptitude Test (MCAT) item

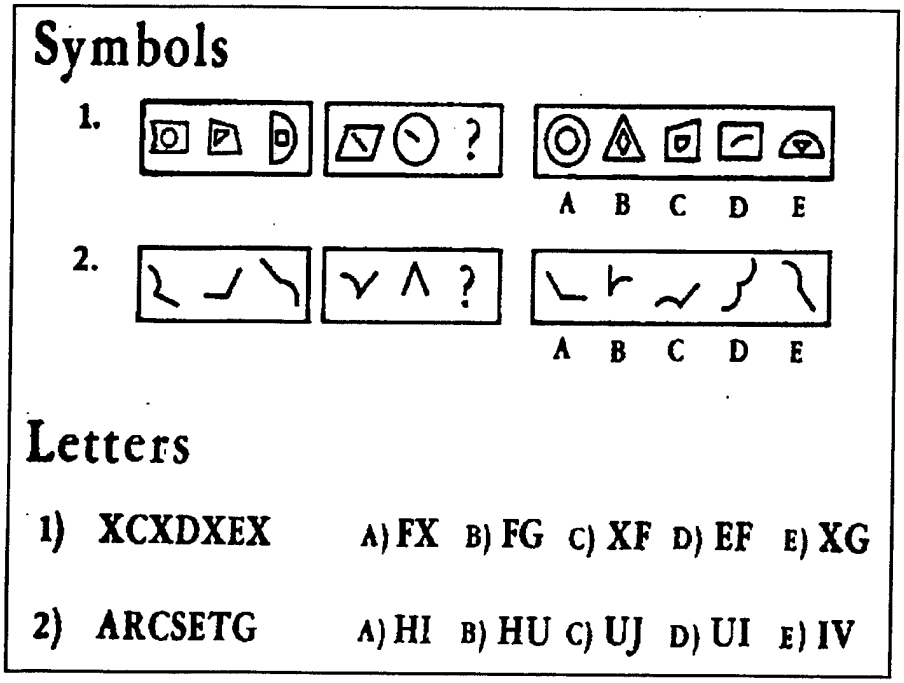

Figure 2: Example Abstract Reasoning Test (ABSR) item

Table 2
ATCS Aptitude test battery scoring

Test	OPM #	Scoring	Weight	N Items
MCAT	510	N Right	2	110
ABSR	157	N Right - (0.25*N Wrong)	1	50
TWS		[(2*MCAT) + ABSR][a]		

Notes: [a]The TWS raw score was then transformed by an OPM transmutation table to the TMC score with a mean of 70 and maximum of 100.

TMC was used in our differential prediction analyses as the measure of candidate aptitude, as it provided a measure of ability unadjusted for previous experience, occupational knowledge, and/or military service. Descriptive statistics for the predictor scores are presented in Table 3 for the applicant sample, all FAA Academy entrants, and the research sample. Minority status was represented by the independent variable RACE, coded as 0 = *Whites*, 1 = *African Americans*. The interaction term (TMC*RACE) was computed as the cross-product of RACE and TMC. Descriptive statistics for the predictor by race are presented in Table 4. The mean TMC for African Americans (89.77, SD = 5.53) was significantly less than the mean TMC for Whites (M = 91.67, SD = 4.77; $t(8,822) = 6.56, p \leq .001$). The distribution of TMC scores for Whites and African Americans is illustrated in Figure 3.

Criterion. The criterion in the differential prediction analysis was performance in the FAA Academy initial controller training program, known as the ATCS Non-radar Screen ("the Screen"). The Screen was originally established in response to recommendations made by the U.S. Congressional House Committee on Government Operations (U.S. Congress, 1976) to "...provide early and continued screening to insure the prompt elimination of unsuccessful trainees and relieve the regional facilities of much of this burden" (p. 13). The Screen was based upon a miniaturized training-testing-evaluation personnel selection model (Siegel, 1978, 1983; Siegel & Bergman, 1975) in which individuals with no prior knowledge of an occupation are trained and then assessed for their potential to succeed in the job. Thirteen assessments of performance, including six classroom tests, observations of performance in six laboratory simulations of non-radar air traffic control, and a final written examination, were made during the Screen (Della Rocco, 1998; Della Rocco, Manning, & Wing, 1990). The final summed composite score (SCREEN) was weighted 20% for the classroom tests, 60% for laboratory simulations, and 20% for the final examination. A minimum SCREEN score of 70 was required to pass. The final composite score was the criterion measure in this study. Descriptive statistics for SCREEN scores are also presented in Table 3 for all FAA Academy Screen entrants and for the research sample.

Procedure

Regression analysis. The classical, regression-based model of test bias was used as our analytic framework to evaluate the degree to which the written ATCS test battery differentially predicted performance in the Screen. Step-down hierarchical multiple regression analysis (Lautenschlager & Mendoza, 1986) was used to evaluate test bias. The step-down approach overcomes the shortcomings of the various step-up procedures (Bartlett, Bobko, Mosier, & Hannan, 1978; Cohen & Cohen, 1975) by accounting for the various changes in the sum of squared error term incrementally, while at the same time ensuring more statistical power than the other methods (Lautenschlager & Mendoza). Step-down analysis assumes the null hypothesis that a common regression line provides the best least-squares fit to the data. The alternative is that a full model including slope and intercept differences between groups is required to fit the data.

Table 3
Descriptive statistics for reference sample of job applicants, all 1985-1992 FAA Academy entrants, and the research sample

	Applicant sample (N=170,578)[a]				FAA Academy entrants (N=14,392)				Research sample (N=8,824)			
Variable[b]	*M*	*SD*	Min	Max	*M*	*SD*	Min	Max	*M*	*SD*	Min	Max
TMC	73.30	14.37	19.53	100.00	91.08	5.43	70.00	100.00	91.61	4.81	70.06	100.00
RACE					0.20	0.40	0.00	1.00	0.03	0.18	0.00	1.00
TMC*RACE					16.18	34.80	0.00	100.00	2.87	15.82	0.00	100.00
SCREEN					72.26	11.80	27.16	99.47	71.85	11.28	27.16	97.59

Notes: [a]Race data not available for applicant sample; Screen score not applicable for applicant sample.
[b]TMC = Transmuted Composite Score from written ATCS test battery; RACE = minority status (0 = *white*, 1 = *African American*); TMC*RACE = cross-product of TMC and RACE; SCREEN = final composite score in FAA Academy ATCS Non-radar Screen program

Table 4
Mean predictor (TMC) and criterion (SCREEN) score differences by race

Variable	Race	*N*	*M*	*SD*	*SE*	*t*	*df*
TMC	White	8,542	91.67	4.77	0.052	6.56***	8,822
	African American	282	89.77	5.53	0.329		
SCREEN	White	8,542	72.12	11.09	0.120	12.64***	8,822
	African American	282	63.57	13.73	0.818		

***$p \leq .001$

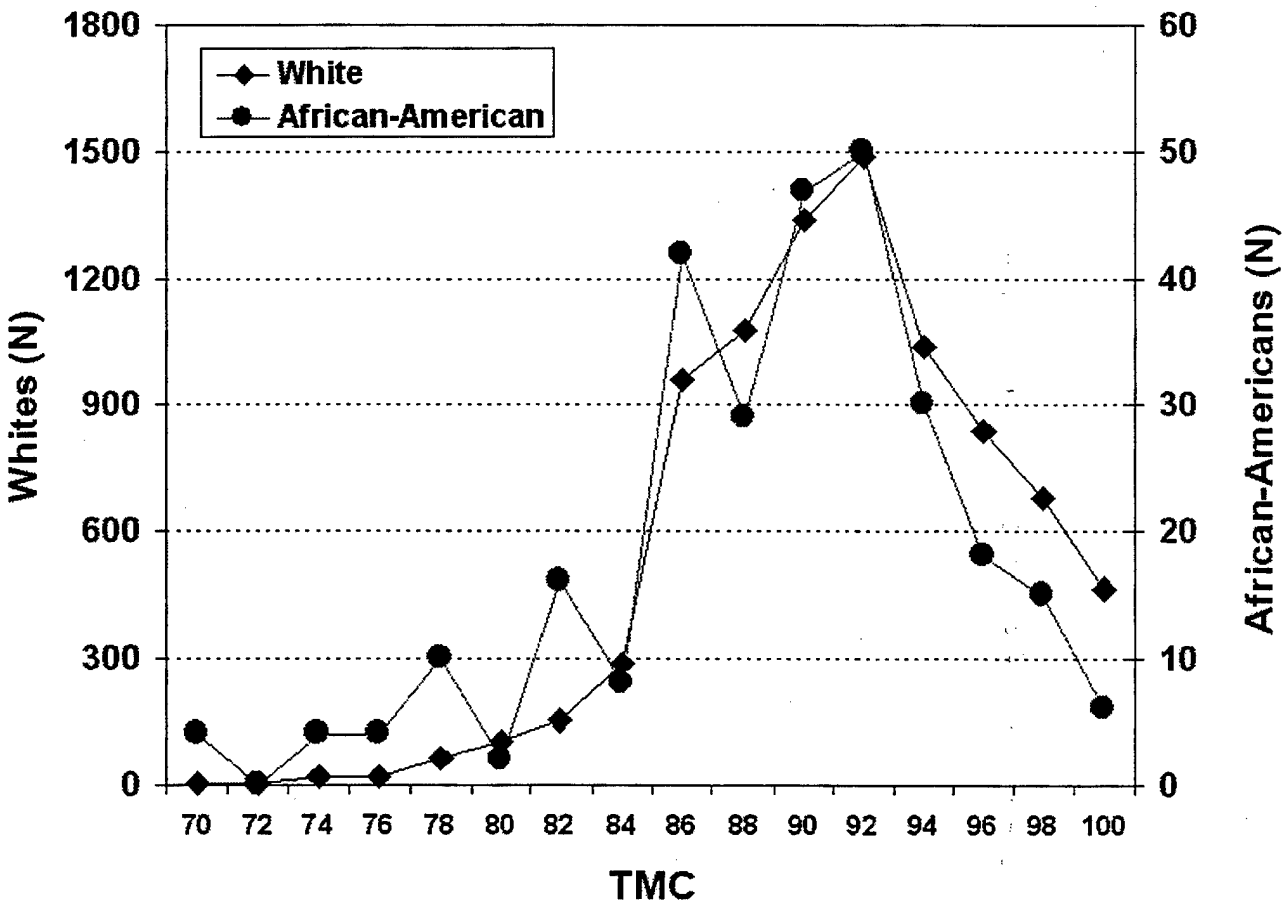

Figure 3: Predictor (TMC) score distribution by race

Our step-down analysis was conducted as follows, using the SPSS-X™ (SPSS Inc., 1989) regression procedure. First, SCREEN was regressed on TMC only (basic model). Second, the criterion was regressed on TMC, the dummy coded group membership variable (0 for whites, 1 for African Americans; in accordance with Pedhazur (1982, p. 274) and the cross-product of TMC and that dummy coded variable (full model). This full model was tested against the simple model of criterion and predictor test only for an incremental change in the R^2 (goodness-of-fit index). A significant change in R^2 suggested potential bias and dictated that further testing for slope and/or intercept differences for the groups be done.

Third, to test for slope differences between groups, SCREEN was regressed on TMC and the group membership variable (group model), and compared with the full model. A significant increment in the R^2 based on a comparison of the group to full model implied different slopes.

Fourth, if slope differences were found, then SCREEN was regressed on TMC and the cross-product of aptitude and group membership (cross-product model). The cross-product model was then compared with the group model; a significant change in R^2 indicated intercept, as well as slope differences, between groups. If no slope differences were found, then the cross-product model was compared with the basic model; a significant change in R^2 indicated only intercept differences between groups. The general logic and associated SPSS-X™ syntax for the step-down hierarchical regression analysis are illustrated in Figure 4.

Technical feasibility

Statistical considerations. Restriction in range, statistical power, and criterion bias are considerations that must be explicitly considered in determining the technical feasibility of an investigation of differential prediction under the *Uniform Guidelines* (29 **CFR** 1607.14.B. (8).(c) and (e); 29 **CFR** 1607.16.U). Both explicit and incidental restrictions in range are recurrent problems in ATCS selection research, as evidenced by the sample sizes and descriptive statistics in Table 3. Variance in TMC for the research sample was explicitly restricted in range due to selection on TMC. Therefore, correlations between TMC and the SCREEN criterion were corrected with respect to the reference sample of 170,578 applicants, using the formula presented by Ghiselli, Campbell, and Zedeck (1981, p. 299). Correlations between variables indicating minority status and the criterion were incidentally restricted in range. These minority status-criterion correlations, including the minority status-by-predictor cross-product to SCREEN correlation, were corrected with respect to the reference sample of 170,578 applicants using the Ghiselli, Campbell, and Zedeck (p. 304) formula for incidental range restriction. Separate differential prediction analyses were conducted based on sample and corrected correlations, as required by the *Uniform Guidelines* (29 **CFR** 1607.15.B. (8)).

The samples in this analysis were of sufficient size to provide enough statistical power to detect even small effects associated with group membership. The step-down approach pools majority and minority data in a single sample to test the null hypothesis that a common regression line provides the best fit to the data. We estimated the available statistical power using Cohen's (1988) regression power tables for three independent variables at an alpha of .05. The risk of a type II error (failing to find an effect, that in fact, was present) was very low, with a .995 probability of detecting even very small effect sizes ($f^2 \leq .01$, or $R = .14$) with a sample of more than 8,000 cases.

Criterion considerations. Finally, as noted by Sackett and Wilk (1994), Lautenschlager and Mendoza (1986), and the *Uniform Guidelines* (29 **CFR** 1607.16.U), the technical feasibility of an assessment of differential prediction depends upon the quality of the job-relevant criterion. If the criterion was systematically biased against minority members, for example, then the regression-based method could not be used to determine the presence or absence of differential prediction by subgroup. Descriptive statistics for the criterion are presented in Table 4; the distribution of criterion scores by minority status is shown in Figure 5. Observed mean criterion score differences in SCREEN by race were about .76 *SD* for the research sample, about twice the estimated differences of .3 to .4 *SD* by race reported by Ford, Kraiger, and Schectman (1986) for ratings-based criteria. In view of the large mean score difference on the composite SCREEN criterion score, we conducted a secondary analysis of the components of that score by race. The principal components of the SCREEN composite score were (a) the average percentage score on four 25-item multiple choice and one map knowledge tests, known as the academic block average score

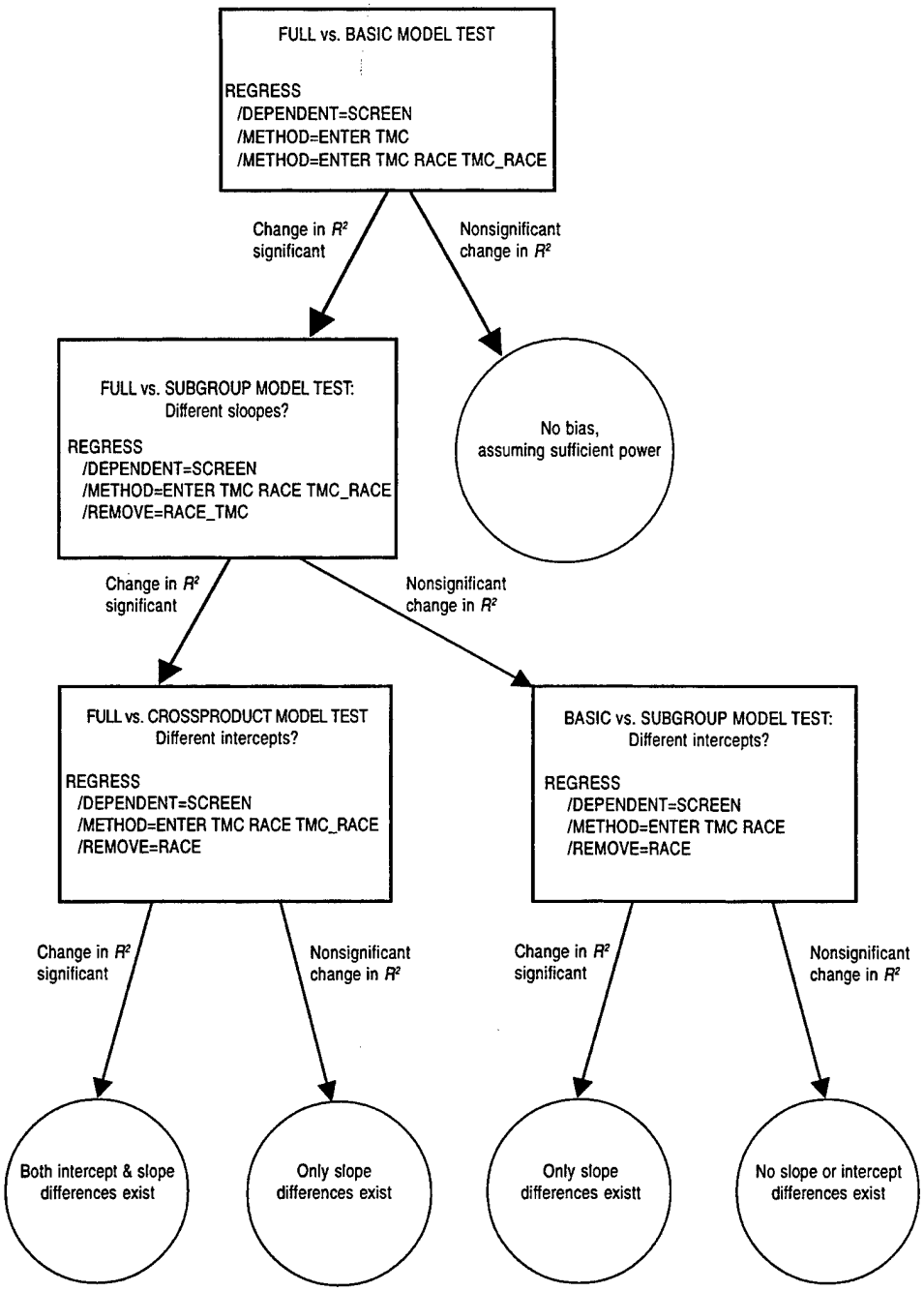

Figure 4: Step-down hierarchical regression analysis logic and SPSS syntax

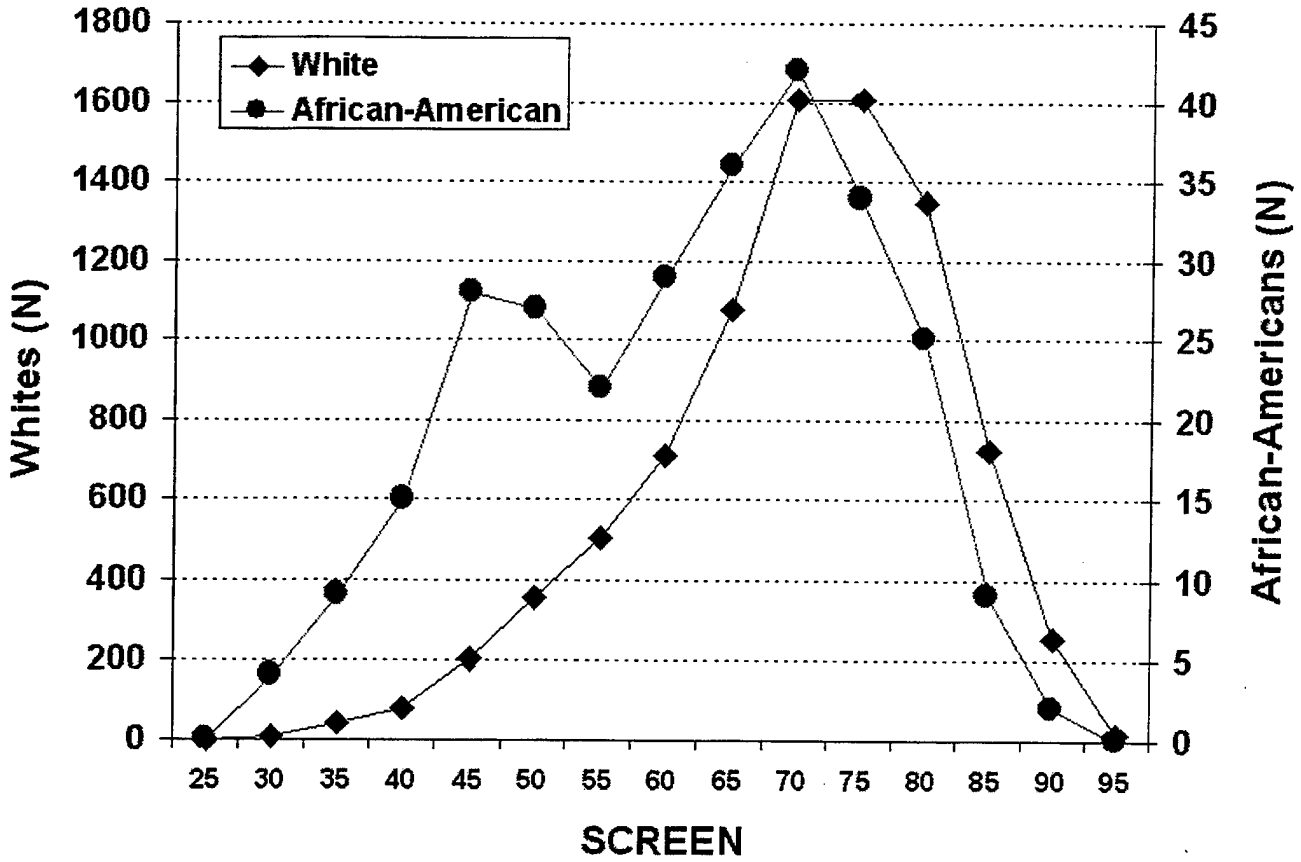

Figure 5: Criterion (SCREEN) score distribution by race

(BLOCKAVG), (b) the percentage score on a 50-item comprehensive multiple choice knowledge test known as the Comprehensive Phase Test (CPT), (c) the average score for the best five of six graded laboratory exercises (AVGLAB5), and (d) the percentage score on a 100-item comprehensive multiple choice Controller Skills Test (CST). Additional details regarding these measures can be found in Della Rocco (1998), Della Rocco, Manning, and Wing (1990), and Manning, Della Rocco, and Bryant (1989). The results of the analysis of component scores by race are presented in Table 5. African American students scored significantly less well than white students on the four criterion component scores. So it might be argued that the statistically significant mean differences on the overall criterion and its components indicated "systematic bias" in the criterion against minority members, and the differential prediction analysis was technically infeasible under the *Uniform Guidelines.*

The relevance of criteria are of particular concern when there are significant differences between groups on those criterion measures (29 CFR 1607.14.B.(2)). According to the *Uniform Guidelines* at 29 CFR 1607.14.B.(3), "Where performance in training is used as a criterion, success in training should be properly measured and the relevance of the training should be shown either through a comparison of the content of the training program with critical or important work behavior(s) of the job(s), or through a demonstration of the relationship between measures of performance in training and measures of job performance." The objective tests and laboratory simulations in the Screen were explicitly linked to specific air traffic control facts, definitions, and procedures found in the air traffic procedures manual. Many of the tasks taught in the Screen were comparable to the tasks performed by sector controllers in en route centers (Della Rocco, Manning, & Wing, 1990). In addition, the relationship of

Table 5
SCREEN criterion component scores by race

		African American			White			
Component	Description	*N*	*M*	*SD*	*N*	*M*	*SD*	*t*
BLOCKAVG	Academic block test average % score	282	88.70	11.24	8,542	93.40	6.87	-10.98***
CPT	Comprehensive phase test % score	282	87.97	9.53	8,542	90.43	7.40	-5.42***
LABAVG5	Average best 5 of 6 graded laboratory exercise score	282	54.68	16.05	8,542	64.31	13.83	-11.45***
CST	Controller skills test % score	282	65.59	17.99	8,542	76.10	13.73	-12.50***

***$p \leq .001$

performance in the Screen to subsequent job outcomes such as performance in radar training and attainment of FPL status has been repeatedly demonstrated (Broach, 1998; Broach & Manning, 1994; Della Rocco, Manning, & Wing, 1990; Manning, Della Rocco, & Bryant, 1989). These findings indicate that the Screen was relevant to the ATCS job as required by the *Uniform Guidelines*, and was, therefore, an appropriate criterion for this investigation of differential prediction on the basis of written ATCS aptitude test scores.

RESULTS

Without corrections for restriction in range

Computed correlations, without corrections for restriction in range, are presented in the lower left-hand corner of the matrix in Table 6. TMC was significantly correlated with minority status (RACE; r = -0.070, $p \leq .01$), the predictor-group cross-product (TMC*RACE; r = - .057, $p \leq .01$), and the final score in the FAA Academy Screen (r = .199, $p \leq .001$). RACE was negatively correlated with the criterion SCREEN score (r = -.133, $p \leq .001$). The results of the differential prediction analysis, using the step-down hierarchical regression analysis on the basis of the sample correlation matrix without corrections for restriction in range, are presented in Table 7.

The null hypothesis that a common regression line provided the best fit was rejected in the basic versus full model analysis, suggesting the presence of some degree of test bias. The increment in R^2 gained by using the full model (predictor, group membership, and cross-product) rather than the basic model (predictor only) was significant ($\Delta R^2 = .015$, $\Delta F = 68.34$, $p \leq .001$). Next, the null hypothesis of same slopes by race could not be rejected in the full versus subgroup model analysis. The subgroup model (predictor and group membership) did not explain any less variance than the full model (ΔR^2 = 0, ΔF= 2.96, *ns*). Following the analytic logic illustrated in Figure 4, the basic and subgroup models were next compared to determine if the intercepts were different for African-Americans and whites. The null hypothesis of same intercepts was rejected, with removal of RACE leading to a significant reduction in explained variance ($\Delta R^2 = -.014$, $\Delta F = 133.69$, $p \leq .001$). The regression of SCREEN on TMC for African Americans and whites is plotted in Figure 6. Overall, the results obtained with the uncorrected correlations did not indicate the need for separate regression equations for African Americans and whites as the slopes were the same for the two groups. However, as shown in Figure 6, criterion scores predicted from the white regression line would consistently over-predict criterion scores for African Americans.

Table 6
Sample (lower corner) and corrected (upper corner) correlation matrix for differential prediction analysis by race

	TMC	RACE	TMC*RACE	SCREEN
TMC		-0.070	-0.057	0.519
RACE	-0.070**		0.998	-0.208
TMC*RACE	-0.057**	0.9980***		-0.191
SCREEN	0.199**	-0.133***	-0.132***	

**$p < .01$, ***$p < .001$

Table 7
Results of step-down hierarchical regression analysis for test bias by race on basis of correlation matrix without corrections for restriction in range

Analysis	Model	R^2	ΔR^2	ΔF	F
Basic v. Full: Overall bias	TMC	0.040			364.12***
	TMC + RACE + TMC*RACE	0.054	0.015	68.34***	168.79***
Full v. Subgroup: Slopes	TMC + RACE + TMC*RACE	0.054			168.79***
	TMC + RACE	0.054	0.000	2.96	251.64***
Full v. Cross-product					
Basic v. Subgroup: Intercepts	TMC + RACE	0.054			251.64***
	TMC	0.040	-0.014	133.69***	364.12***

Notes: Full v. cross-product model comparison not conducted. See Figure 1 for logic and flow of step-down hierarchical regression analysis.

***$p < .001$

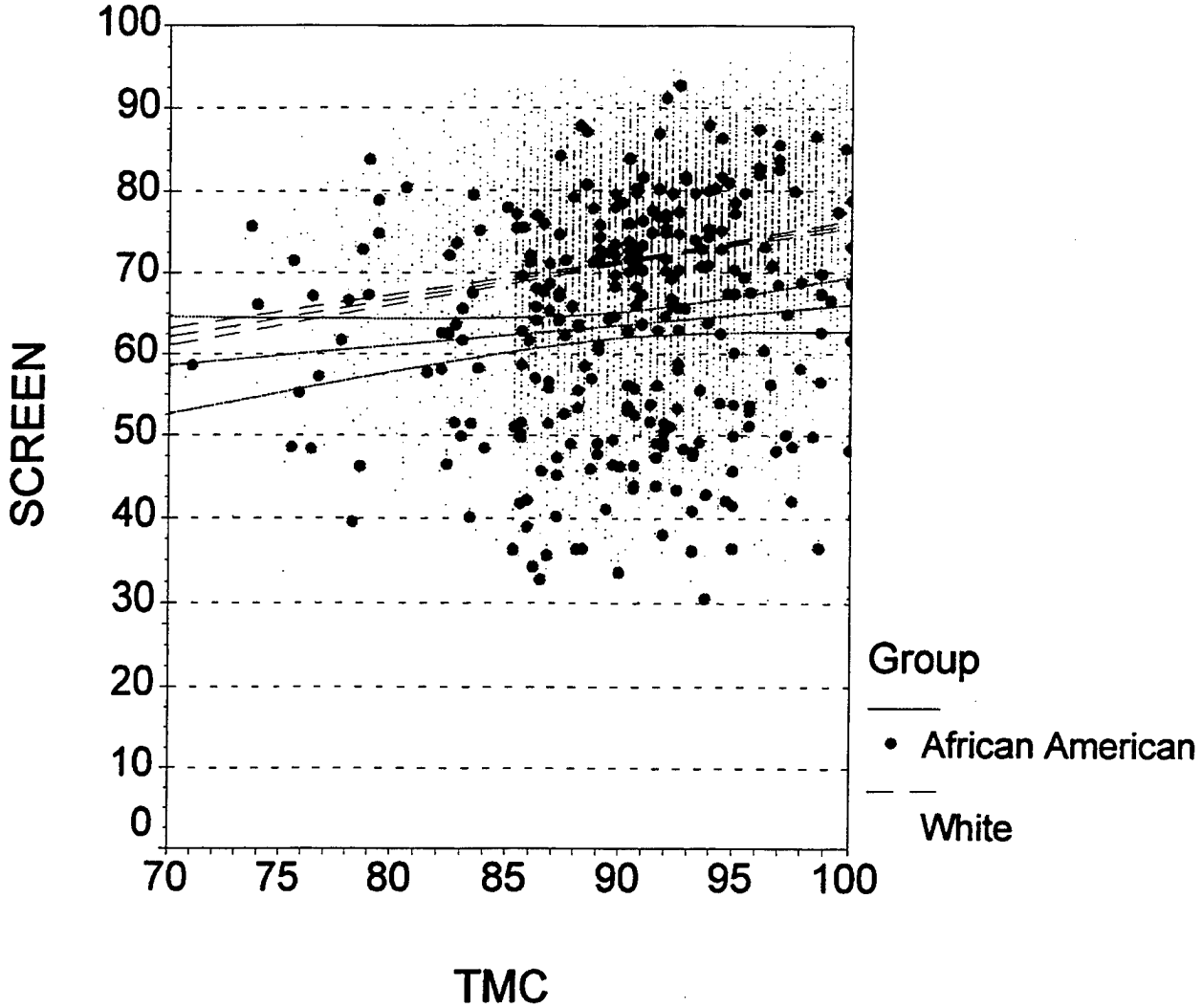

Figure 6: Regression of SCREEN on TMC for African Americans and whites, without corrections for restriction in range

Moreover, the regression line for whites is essentially the same regression line for the combined sample. The regression (without corrections for restriction in range) of SCREEN on TMC for the combined sample is

PREDICTED SCREEN = 29.04 + (0.467*TMC). (1)

The regression equation (without corrections for restriction in range) for whites only is

PREDICTED SCREEN = 30.28 + (0.456*TMC). (2)

The constant for the whites-only equation is within one standard error of the value computed on the basis of the combined sample; similarly, the unstandardized regression weight computed for whites only is within one standard error of the value computed for the combined sample. While no formal statistical test is possible as the samples (combined and whites only) are not independent, we argue that the combined group regression line is, in essence, identical to the regression line for whites. Determining the weights for the aptitude test components and cut score on the basis of a combined sample that is predominantly white, as was done for the ATCS aptitude test battery (Rock, Dailey, Ozur, Boone, & Pickrel, 1982, p. 193), in effect sets the weights and cut score on the basis of the white regression line. As a consequence, it is appropriate to assert that the performance of African Americans is over-predicted on the basis of the white regression line, and the test is biased under Cleary's definition.

With corrections for restriction in range

However, as shown in Table 3, predictor scores were severely restricted. Consequently, evidence based on those uncorrected correlations may be somewhat misleading as to the predictive bias of the predictor (29 CFR 1607.14.B.8.(c)). Analyses based on correlations corrected for explicit and implicit restriction in range may provide a better assessment of the predictive bias of the OPM test battery with respect to the applicant population. Corrected correlations are presented in the upper right-hand triangle of the matrix in Table 6. The estimated population correlation between TMC and performance in the Academy Screen increased from .199 to .519 with correction for explicit restriction in range. After correcting for incidental restriction in range, the correlation between RACE and SCREEN increased to -.208, as did the correlation between the cross-product (TMC_SCREEN) and SCREEN (-.191).

The results of the differential prediction analysis, using the step-down hierarchical regression analysis based on the corrected correlations, are presented in Table 8. The null hypothesis of a common regression line was rejected in the basic versus full model comparison, suggesting the presence of some degree of test bias. The increment in R^2 associated with the full model over the basic model was significant (ΔR^2 = .056, ΔF = 366.54, $p \leq .001$). Next, the null hypothesis of same slopes by race was rejected in the full versus subgroup model analysis. The subgroup model (predictor and group membership) explained less variance than the full model (ΔR^2 = -.026, ΔF = 345.85, $p \leq .001$). Following the logic illustrated in Figure 4, the full and cross-product models were next compared to determine if the intercepts were different for African American and white applicants. The null hypothesis of same intercepts was also rejected, with removal of RACE leading to a significant reduction in explained variance (ΔR^2 = -.030, ΔF = 391.29, $p \leq .001$). Overall, the results obtained with the corrected correlations indicated the need for separate regression equations for African American and white applicants, with different slopes and intercepts for the two groups. Therefore, correlations between TMC and SCREEN were computed for African Americans and whites separately, based on the research sample data. These correlations were then corrected for explicit restriction in range based on the applicant sample TMC *SD*. The corrected correlations were then submitted to regression analysis. The resulting equation for whites was

$$\text{PREDICTED SCREEN} = -37.85 + (1.200 * \text{TMC}), \quad (3)$$

compared with an equation for African Americans of

$$\text{PREDICTED SCREEN} = 7.63 + (0.623 * \text{TMC}. \quad (4)$$

These regression equations are plotted in Figure 7. This analysis based on corrected correlations leads to the same conclusion as the preceding analysis: TMC appeared to be biased using Cleary's (1968) definition of test bias. The performance of African Americans in the FAA Academy Non-radar ATCS Screen would be over-predicted from the white regression line. Given that the combined sample regression line is essentially identical to the white regression line, it is fair to conclude that the performance of African Americans in the Screen would be over-predicted by the combined sample regression line.

DISCUSSION

Pass rate differences

Evaluation of a selection test under the *Uniform Guidelines* includes (a) an assessment of differences in pass rates between groups arising from use of the test, and if differential pass rates are demonstrated, (b) an investigation of predictive bias associated with the test. Our study is silent as to any differences in pass rates for whites and African Americans arising from use of the written ATCS aptitude test battery, due to the lack of racial identifiers for applicants. However, previous research indicated the African American pass rate on the ATCS aptitude test battery was likely to differ significantly from that of whites. Previous research also concluded that the battery would likely exhibit differential prediction for African Americans (Rock, Dailey, Ozur, Boone, & Pickrel, 1982, p. 153). Therefore, an evaluation of the predictive bias for the battery was conducted, as required by the *Uniform Guidelines* and relevant professional selection testing standards and principles.

Table 8
Results of step-down hierarchical regression analysis for test bias by race on basis of correlation matrix corrected for restriction in range

Analysis	Model	R^2	ΔR^2	ΔF	F
Basic v. Full: Overall bias	TMC	0.519			3252.36***
	TMC + RACE + TMC*RACE	0.570	0.056	366.54***	1418.32***
Full v. Subgroup: Slopes	TMC + RACE + TMC*RACE	0.570			1418.32***
	TMC + RACE	0.547	-0.026	345.85***	1881.02***
Full v. Cross-product: Intercepts	TMC + RACE + TMC*RACE	0.570			1418.32***
	TMC + TMC*RACE	0.544	-0.030	391.29***	1849.98***
Basic v. Subgroup					

Notes: Analysis terminated with comparison of full v. cross-product models. See Figure 1 for logic and flow of step-down hierarchical regression analysis.

***$p < .001$

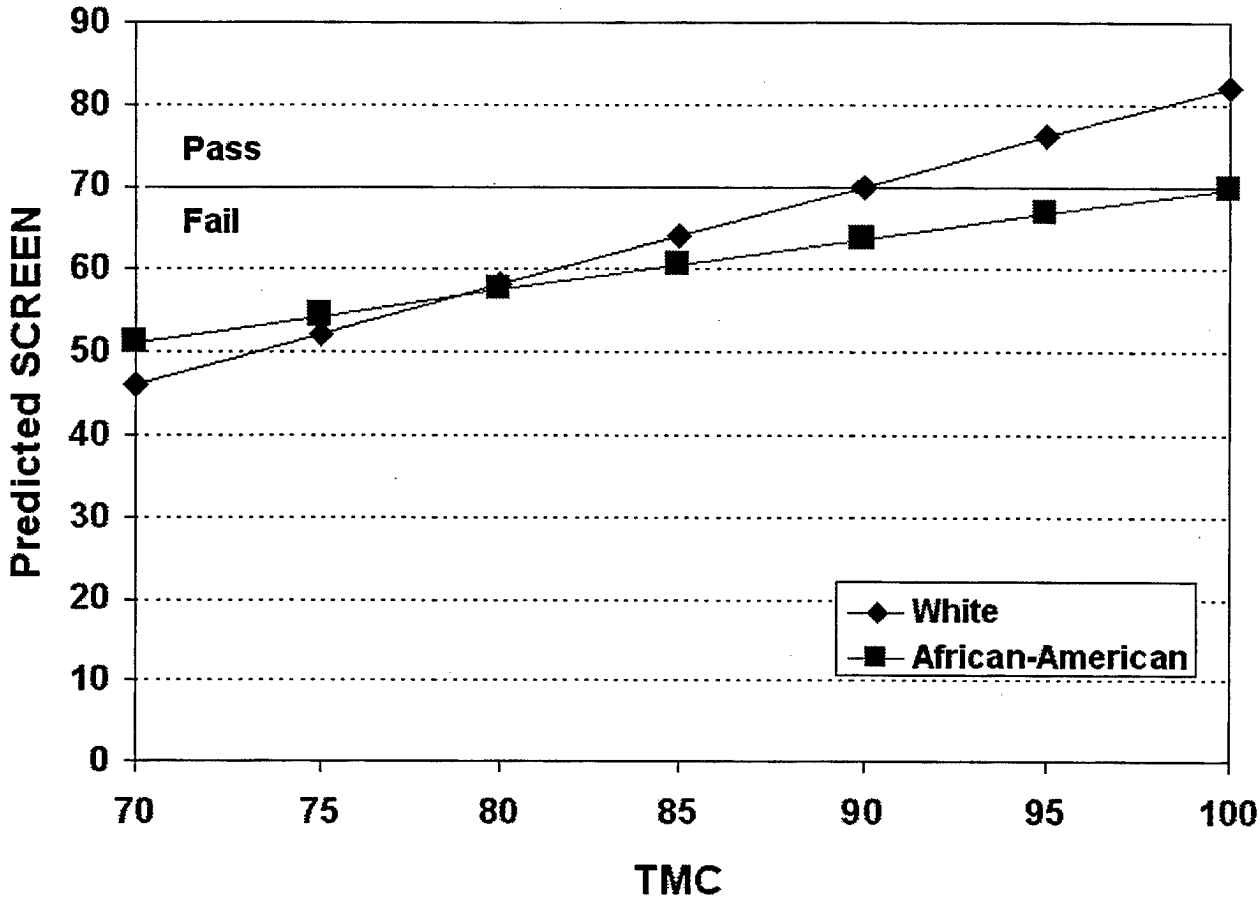

Figure 7: Regression of SCREEN on TMC for African Americans and whites, with corrections for restriction in range

Test bias

Our step-down hierarchical regression analyses found that the written ATCS aptitude test battery exhibited predictive bias, as defined by the *Uniform Guidelines* at 29 **CFR** 1607.14.B.(8).(a) and Cleary (1968). That is, the white regression line over-predicted the performance of the protected group. Moreover, the white regression line was indistinguishable from the combined sample regression. Therefore, we concluded that the common regression line, used as the operational basis for determining test weights and cut score (Rock, Dailey, Ozur, Boone, & Pickrel, 1982), over-predicted the performance of African Americans in the FAA Academy ATCS Non-radar Screen. Adverse impact and over-prediction of training and job performance has been reported for tests of cognitive ability similar to the ATCS test battery such as the General Aptitude Test Battery (GATB) (Hartigan & Wigdor, 1989; Hunter, 1983; Schmidt, 1988; Wigdor & Garner, 1982; Wigdor & Sackett, 1993) and the Armed Services Vocational Aptitude Battery (ASVAB; Dunbar & Novick, 1988). These outcomes do not occur because scores on tests such as the ATCS aptitude test battery mean something different for African Americans. The tests are not biased in that sense. Rather, the different pass rates and over-prediction of subsequent performance in initial training results from the interplay of two factors, as has been found in other selection settings (see Gordon, Lewis, & Quigley, 1988): the lower average scores for African Americans relative to whites, and the less than perfect validity of the test scores.

Impact of over-prediction

The impact of this apparent over-prediction on African Americans is illustrated by an analysis of the selection decisions that would have been made using a strict cut-off on TMC as the hiring criterion. In such an analysis, hire/no hire recommendations are crosstabulated with job outcomes to create a decision table. As previously noted, a TMC score of at least 90 was required to predict a passing score of 70 in the FAA Academy Screen, based on the common (white) regression line. Based on that statistical relationship, an operational decision rule of hiring applicants scoring at 90 or above was generally used by the FAA. The policy of preferring applicants with scores of 90 or above was made explicit in 1990 (Associate Administrator for Human Resources Management, 1990). TMC was therefore recoded into a dichotomous variable (0 = *TMC* < *90*; 1 = *TMC* ≥ *90*) to represent this operational hiring criterion. The dichotomized variable was then crosstabulated with FAA Academy ATCS Non-radar Screen pass/fail outcomes by race, as shown in Table 9.

Arrangement of the crosstabulation

By convention, two-by-two predictive tables such as Table 9 are arranged such that the rows are defined by *test* predictions (positive on top, negative on the bottom), and columns are defined by *criterion* results (negative on the left, positive on the right)(Gordon, Lewis, & Quigley, 1988). The top row for each group, therefore, consists of two cells: false positives (in selection, also known as incorrect acceptances), and true positives (correct acceptances). The bottom row for each group consists of two cells: true negatives (correct rejections), and false negatives (also known as incorrect rejections). False positives are those persons who had a TMC of 90 or greater (TMC ≥ 90) <u>and</u> who subsequently failed the Screen (SCREEN < 70). True positives are those cases with a TMC score of 90 or greater <u>and</u> who subsequently passed the Screen (SCREEN ≥ 70). True negatives are those individuals who had a TMC of less than 90 <u>and</u> also failed the Screen (SCREEN < 70). Finally, false negatives are those persons with TMC scores of less than 90 <u>and</u> who passed the Screen (SCREEN ≥ 70).

Analysis of crosstabulation cells

Both temporal ordering and direction determine which variable should be used as the base for calculation of the percentages showing the effect in a crosstabulation (Davis, 1971; Zeisel, 1957). In this case, the written ATCS test battery preceded the Screen, often by several months (Aul, 1991). The putative causal role belongs to the predictor test; the analytic goal is to estimate performance on the criterion (SCREEN) from the predictor score (TMC). Therefore, the percentages of true and false positives and true and false negatives in a fourfold classification table such as Table 9 should be calculated horizontally, using *the selection test recommendation* (hire, no hire) as the base for such calculations (Gordon, Lewis, & Quigley). The focus of a fourfold classification table is the accuracy of predictions made on the basis of test scores about future job performance as represented by the decision errors. Decision errors provide the data for evaluating the impact of over- and under-prediction of subsequent training or job performance (Gordon, Lewis, & Quigley).

Decision errors: Incorrect rejections (false negatives). The incorrect rejection rate for each group is defined as the ratio of (a) the number of persons with TMC scores of less than 90 who passed the Screen (the lower right cell for each group in Table 9) to (b) the marginal row total for persons with TMC scores of less than 90 for each group. The actual job performance for persons in this cell for each group was under-predicted by their predictor test score. There were 39 African Americans with TMC scores of less than 90 who passed the FAA Academy, out of 128 total with TMC scores of less than 90, for an incorrect rejection (false negative) rate of 30.5%. As shown in Table 9, there were 1,684 whites with TMC scores of less than 90 who passed the Screen out of 3,046 whites with TMC scores of less than 90, for an incorrect rejection rate of 55.3% (1,684/3,046). If the "TMC ≥ 90" hiring rule had been used, the proportion of rejections that would have been incorrect for African Americans was significantly less than the proportion for whites ($Z = -5.52$, $p \leq .001$).

Decision errors: Incorrect acceptances (false positives). The incorrect acceptance or false positive rate for each group is defined explicitly as the ratio of (a) the number of persons with TMC scores of 90 or greater who failed the Screen (b) to the marginal row total of persons with TMC scores of 90 or greater. There were 84 African Americans who failed the Screen out of a total of 154 with TMC scores greater than or equal to 90, for an incorrect acceptance rate of 54.5%. Of the 5,496 whites with TMC scores of 90 or greater, 1,768 failed the Screen, for an incorrect acceptance rate of 32.2%. If the "TMC ≥ 90" hiring

Table 9
Crosstabulation of dichotomized predictor scores (TMC <90, ≥90) and FAA Academy Screen outcomes (SCREEN <70, ≥70) by race

African Americans

		Fail (SCREEN < 70)	Pass (SCREEN ≥ 70)	Row Totals
Hire (TMC ≥ 90)	*n*	84	70	154
	Row %	54.5%	45.5%	
	Column %	48.6%	64.2%	54.6%
	Total %	29.8%	24.8%	(Hire)
		False Positive	*True Positive*	
No Hire (TMC <90)	*n*	89	39	128
	Row %	69.5%	30.5%	
	Column %	51.4%	35.8%	45.4%
	Total %	31.6%	13.8%	(No Hire)
		True Negative	*False Negative*	
	Column Totals	173	109	282
		61.3%	38.7%	
		(Fail Screen)	(Pass Screen)	

Whites

Hire (TMC ≥ 90)	*n*	1,768	3,728	5,496
	Row %	32.2%	67.8%	
	Column %	56.5%	68.9%	64.3%
	Total %	20.7%	43.6%	(Hire)
		False Positive	*True Positive*	
No Hire (TMC <90)	*n*	1,362	1,684	3,046
	Row %	44.7%	55.3%	
	Column %	43.5%	31.1%	35.7%
	Total %	15.9%	19.7%	(No Hire)
		True Negative	*False Negative*	
	Column Totals	3,130	5,412	8,542
		36.6%	63.4%	
		(Fail Screen)	(Pass Screen)	

rule had been used, the proportion of "recommended hire" decisions that would have been incorrect (false positives) for African Americans was significantly greater than the proportion for whites ($Z = 5.81, p \leq .001$). In other words, it is likely that significantly more African-American than white candidates would have been incorrectly accepted on the basis of their test scores using a TMC score of 90 or greater as the hiring rule.

Interpretation of decision errors. Use of the "TMC ≥ 90" rule would have resulted in significantly different proportions of decision errors for whites and African Americans. Specifically, significantly more of the African Americans with aptitude scores greater than 90 went on to fail training than would have been expected on the basis of the common (white) regression line. That is, the performance of African Americans in the Screen was over-predicted by TMC. Moreover, the burden of incorrect rejections (false negatives) on the basis of ATCS aptitude test scores did not fall disproportionately on minority candidates, in contrast to results from previous research on the GATB (Hartigan & Wigdor, 1989; Wigdor & Sackett, 1993).

Interpretive issues

Three issues might be raised as objections or concerns to our interpretation of the results. First, the statistical effects detected in the differential prediction analyses were generally small and detectable only with very large samples. Those small effects were more pronounced with corrections for restriction in range. One could argue, therefore, that the results were artifactual (cf. Hunter, Schmidt, & Rauschenberger, 1983), and had little practical significance. We would counter by noting that the FAA controller selection process was a large scale selection system. Small effect sizes have significant practical effects in large-scale selection systems such as that for controllers (Schroeder, Broach, & Young, 1993). Corrected correlations may also provide more accurate estimates of test validity, particularly in large samples and under stringent selection ratios (Bobko, 1983; Millsap, 1988), such as encountered by the FAA. Uncorrected coefficients appear to be downwardly biased estimates of the true population validity coefficients (Lee, Miller, & Graham, 1982). Therefore, differential prediction analyses based on corrected correlations provide less biased estimates of population effects. Moreover, study factors such as disparate sample sizes between groups and the small moderating effect may have reduced the overall statistical power of the analysis (Aguinis & Stone-Romero, 1997). Yet statistically a significant moderator effect for race on the validity of the written ATCS aptitude test battery was detected. Finally, we believe that these effects cannot be lightly dismissed, in view of the very real practical consequence for the FAA ATCS selection program: a higher proportion of African Americans failed than would have been expected on the basis of aptitude test scores.

Second, one might argue that the observed differential prediction of SCREEN on the basis of TMC for African Americans and whites might be attributable to bias in the criterion measure. The mean criterion scores for the groups were significantly different (Table 3), and the proportion of African Americans passing the Screen (38.7%) was significantly lower than the white proportion (63.4%; $Z = -8.34, p \leq .001$). The pass rate ratio at the Screen was .61, indicating that African American trainees passed the Screen at 61% of the white pass rate. Yet a difference in average criterion scores between groups does not establish bias; "... the presence or absence of bias cannot be detected from knowledge of criterion scores alone" (SIOP, 1987, p. 10). Bias, in this sense, is the extent to which a criterion includes unwanted systematic variance. Unwanted systematic variance might be introduced into supervisory ratings, for example, by rater errors such as halo, leniency, stereotyping, and ratee-by-rater race interactions (Ford, Kraiger, & Schechtman, 1986; Kraiger & Ford, 1985). But, as noted by Ford and his co-authors in their review of the literature, the effects found for rating errors and rater-by-ratee interactions do not account for all of the variability between groups, and do not preclude the possibility that actual performance differences between groups exist. The degree to which unwanted systematic variance might have been introduced into the measures comprising the SCREEN composite criterion score has not been formally assessed; therefore, the possibility of criterion bias cannot be dismissed. Further investigations of the process by which the criterion measures were generated and the degree to which possible biasing factors account for group differences are required.

Third, unmeasured variables may have been confounded with the predictor, resulting in a defective study design (Anastasi, 1988; see Standard 1.22, p. 17, American Educational Research Association,

American Psychological Association, & National Council on Measurement in Education, 1985). One might suspect, for example, that education and scores on the written aptitude test might be confounded in view of the generally positive correlation between such tests and educational attainment: The group with lower scores on an aptitude test battery might have lower overall educational levels than the other group with higher scores. Overall, African Americans in the research sample had slightly lower educational levels than whites: 22.7% of African Americans reported completing a baccalaureate degree, compared with 32.2% of whites (Z = -3.37, $p \leq .001$). The correlation between education and TMC was .089 ($p \leq .01$). While low, this correlation is about the same magnitude as the uncorrected correlations between TMC and SCREEN reported in Table 6. Thus, a plausible alternative explanation for our results might be the differing educational levels for the groups. Another unmeasured variable that might influence our results might be age at entry into the FAA Academy. Age has been found to be related to Screen outcomes in previous research (Collins, Nye, & Manning, 1990; VanDeventer, 1983).

Introduction of these additional variables, and their multiple interactions with race, aptitude, and each other, however, presents a difficult analytic problem. Procedures for conducting a hierarchical, step-down regression analysis with multiple variables, and interpretation of the results from such an analysis, have not been defined. However, statistical procedures for testing the fit of a measurement and structural model for different groups or populations are relatively well established. An alternative analytic approach would be to conduct a structural equations analysis of the relationships between aptitude (TMC), education, age at entry, and performance in the FAA Academy Screen. For example, the model could be developed on the basis of half of the white sample, and cross-validated on the other half. The degree that the cross-validated white model fit the data for African Americans could then be formally tested. Such an approach would also allow a very focused test of the equality of the TMC – SCREEN path parameter for African Americans and whites. Moreover, such a strong analytic method would allow for the more precise specification of method, predictor, and criterion measures and latent constructs required to assess criterion as well as predictor bias in criterion-related validation (Schmitt, Pulakos, Nason, & Whitney, 1996).

CONCLUSION

In conclusion, on the one hand, the ATCS written aptitude test battery in use by the FAA through February 1992 may have operated as an "engine of exclusion" (Seymour, 1988) in terms of differential selection rates. Historical data suggest that the test battery may have excluded more African Americans than whites; however, a definitive selection rate analysis was not technically feasible in this study due to the lack of racial identifiers for applicants. On the other hand, the analyses indicated that the written ATCS aptitude test battery, in accordance with the definition used by Cleary (1968), the *Uniform Guidelines*, and relevant professional testing standards and principles, appears to have been biased in that the performance of African Americans was over-predicted by the common (white) regression of aptitude scores on subsequent performance in training. That is, use of the recommended cut-off of 90 on the composite predictor score TMC would have resulted in a greater incorrect acceptance (false positive) rate for African Americans than whites. These results appear to be consistent with the general findings that ability tests do not under-predict the performance of minorities (Linn, 1994; Schmidt, 1988). We recommend three additional research studies as next steps. First, we recommend a careful and detailed assessment of how, and to what degree, inappropriate systematic variance might have been introduced into the 13 scores representing the degree to which students mastered required air traffic control skills and knowledges in the FAA Academy screening program. The second step is to incorporate those findings on criterion bias with the predictor and other exogenous measures such as education and age into a structural equations model as the basis for testing hypotheses about differential prediction in the selection of controllers. The final step we recommend is to examine the practical and organizational consequences of operational use of the majority (white) regression line as the basis for historical selection decisions, in view of the apparent over-prediction of the performance of African Americans in the FAA Academy Screen.

REFERENCES

Aguinis, H., & Stone-Romero, E.F. (1997). Methodological artifacts in moderated multiple regression and their effects on statistical power. *Journal of Applied Psychology, 82*, 192 - 206.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* (5th ed.). Washington, DC: American Psychological Association.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Associate Administrator for Human Resources Management. (1990, August). *Report on the FAA employment manager's conference, June 25-29, 1990, Arlington, VA.* Washington, DC: Federal Aviation Administration Staffing Policy Division.

Aul, J.C. (1991). Employing air traffic controllers. In H. Wing & C.A. Manning (Eds.), *Selection of air traffic controllers: Complexity, requirements, and public interest* (pp. 7-12). (DOT/FAA/AM-91/9). Washington, DC: Federal Aviation Administration Office of Aviation Medicine. (NTIS No. ADA238267)

Bartlett, C.J., Bobko, P., Mosier, S.B., & Hannan, R. (1978). Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis. *Personnel Psychology, 31*, 233-241.

Bobko, P. (1983). An analysis of correlations corrected for attenuation and range restriction. *Journal of Applied Psychology, 68*, 584-589.

Broach, D. (1998). Air traffic control specialist aptitude testing, 1981-1992. In D.Broach (Ed.), *Recovery of the FAA air traffic control specialist workforce, 1981-1992* (pp. 7-16). Washington, DC: Federal Aviation Administration Office of Aviation Medicine.

Broach, D., & Manning, C.A. (1994). *Validity of the air traffic control specialist non-radar screen as a predictor of performance in radar-based air traffic control training.* (DOT/FAA/AM-94/9). Washington, DC: Federal Aviation Administration Office of Aviation Medicine. (NTIS No. ADA279745)

Brokaw, L.D. (1984). Early research on controller selection. In S.B. Sells, J.T. Dailey, & E.W. Pickrel (Eds.), *Selection of air traffic controllers* (pp. 39-78). (DOT/FAA/AM-84/2). Washington, DC: Federal Aviation Administration Office of Aviation Medicine. (NTIS No. ADA147765)

Cleary, T.A. (1968). Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement, 5*, 115-124.

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

Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Collins, W.E., Boone, J.O., & VanDeventer, A.D. (1984). The selection of air traffic control specialist: Contributions by the Civil Aeromedical Institute. In S.B. Sells, J.T. Dailey, & E.W. Pickrel (Eds.), *Selection of air traffic controllers* (pp. 79-112). (DOT/FAA/AM-84/2). Washington, DC: Federal Aviation Administration Office of Aviation Medicine. (NTIS No. ADA147765)

Collins, W.E., Nye, L.G., & Manning, C.A. (1990). *Studies of poststrike air traffic control specialist trainees: III. Changes in demographic characteristics of Academy entrants and biodemographic predictors of success in air traffic controller selection and Academy screening.* (DOT/FAA/AM-90/4). Washington, DC: Federal Aviation Administration Office of Aviation Medicine. (NTIS No. ADA223480)

Dailey, J.T., & Pickrel, E.W. (1984). Development of the air traffic controller Occupational Knowledge Test. In S.B. Sells, J. T. Dailey, & E.W. Pickrel (Eds.), *Selection of air traffic controllers* (pp. 299-322). (DOT/FAA/AM-84/2). Washington, DC: Federal Aviation Administration Office of Aviation Medicine. (NTIS No. ADA147765)

Davis, J.A. (1971). *Elementary survey analysis.* Englewood Cliffs, NJ: Prentice-Hall.

Della Rocco, P.S. (1998). Air traffic control specialist screening programs and strike recovery. In D. Broach (Ed.), *Recovery of the FAA air traffic control specialist workforce, 1981-1992* (pp. 17-22). Washington, DC: Federal Aviation Administration Office of Aviation Medicine.

Della Rocco, P.S., Manning, C.A., & Wing, H. (1990). *Selection of controllers for automated systems: Applications from current research.* (DOT/FAA/AM-90/13). Washington, DC: Federal Aviation Administration Office of Aviation Medicine. (NTIS No. ADA230058)

Dunbar, S.B., & Novick, M.R. (1988). On predicting success in training for men and women: Examples from Marine Corps clerical specialties. *Journal of Applied Psychology, 73,* 545-550.

Equal Employment Opportunity Commission (1978). *Uniform Guidelines on Employee Selection Procedures.* 29 CFR 1607.

Federal Aviation Administration. (1993). *FAA diversity plan.* Washington, DC: Federal Aviation Administration Office of the Administrator for Human Resources Management.

Federal Aviation Administration. (1998). *1998 Federal Aviation Administration strategic plan.* Washington, DC: Federal Aviation Administration Office of the Administrator.

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

Ghiselli, E., Campbell, J., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences.* San Francisco: Freeman.

Gordon, R.A., Lewis, M.A., & Quigley, A.M. (1988). Can we count on muddling through the *g* crisis in employment? *Journal of Vocational Behavior, 33,* 424-452.

Harris, P. (1986). *A construct validity study of the Federal Aviation Administration Multiplex Controller Aptitude Test.* Washington, DC: U.S. Office of Personnel Management Office of Staffing Policy.

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

Hunter, J.E. (1983). *Fairness of the General Aptitude Test Battery: Ability differences and their impact on minority hiring rates.* (United States Employment Service Test Research Report No. 46). Washington, DC: U.S. Department of Labor Employment and Training Administration.

Hunter, J.E., Schmidt, F.L., & Rauschenberger, J. (1984). Methodological and statistical issues in the study of bias in mental testing. In C.R. Reynolds & R.T. Brown (Eds.), *Perspectives on bias in mental testing.* New York: Plenum.

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

Lautenschlager, G.J., & Mendoza, J.L. (1986). A step-down hierarchical multiple regression analysis for examining hypotheses about test bias in prediction. *Applied Psychological Measurement, 10,* 133-139.

Lee, R., Miller, K., & Graham, W. (1982). Corrections for restriction of range and attenuation in criterion-related validation studies. *Journal of Applied Psychology, 67,* 637-639.

Lewis, M.A. (1978). Objective assessment of prior air traffic control related experience through the use of the Occupational Knowledge Test. *Aviation, Space and Environmental Medicine, 49,* 1155-1159.

Lilienthal, M.G., & Pettyjohn, F.S. (1981). *Multiplex Controller Aptitude Test and Occupational Knowledge Test: Selection tools for air traffic controllers.* (NAMRL Special Report 82-1). Pensacola, FL: Naval Aerospace Medical Research Laboratory. (NTIS No. ADA118803).

Linn, R.L. (1994). Fair test use: Research and policy. In M.G. Rumsey, C.B. Walker, & J.H. Harris (Eds.), *Personnel selection and classification* (pp. 363 - 375). Hillsdale, NJ: Erlbaum.

Manning, C.A. (1991). Procedures for selection of air traffic control specialists. In H. Wing & C.A. Manning (Eds.), *Selection of air traffic controllers: Complexity, requirements, and public interest* (pp. 13-22). (DOT/FAA/AM-91/9). Washington, DC: Federal Aviation Administration Office of Aviation Medicine. (NTIS No. ADA238267)

Manning, C.A., Della Rocco, P.S., & Bryant, K. (1989). *Prediction of success in air traffic control field training as a function of selection and screening performance.* (DOT/FAA/AM-89/6). Washington, DC: Federal Aviation Administration Office of Aviation Medicine. (NTIS No. ADA209327)

Millsap, R.E. (1988). Sampling variance in the correlation coefficient under range restriction: A monte carlo study. *Journal of Applied Psychology, 74,* 456-461.

Pedhazur, E.J. (1982). *Multiple regression in behavioral research* (2nd Ed.). New York: CBS College Publishing.

Rock, D.B., Dailey, J.T., Ozur, H., Boone, J.O., & Pickrel, E.W. (1982). *Selection of applicants for the air traffic controller occupation.* (DOT/FAA/AM-82/11). Washington, DC: Federal Aviation Administration Office of Aviation Medicine. (NTIS No. ADA122795/8)

Rock, D.B., Dailey, J.T., Ozur, H., Boone, J.O., & Pickrel, E.W. (1984a). Conformity of the new experimental test battery to the Uniform Guidelines on Employee Selection Requirements. In S.B. Sells, J.T. Dailey, & E.W. Pickrel (Eds.), *Selection of air traffic controllers* (pp. 503-542). (DOT/FAA/AM-84/2). Washington, DC: Federal Aviation Administration Office of Aviation Medicine. (NTIS No. ADA147765)

Rock, D.B., Dailey, J.T., Ozur, H., Boone, J.O., & Pickrel, E.W. (1984b). Validity and utility of the ATC experimental tests battery. Study of Academy trainees, 1978. In S.B. Sells, J.T. Dailey, & E.W. Pickrel (Eds.), *Selection of air traffic controllers* (pp. 459-502). (DOT/FAA/AM-84/2). Washington, DC: Federal Aviation Administration Office of Aviation Medicine. (NTIS No. ADA147765)

Sackett, P.R. (1996). Interpreting the ban on minority group score adjustment in preemployment testing. In R.S. Barrett (Ed.), *Fair employment strategies in human resource management* (pp. 246-256). Westport, CT: Quorum Books.

Sackett, P.R., & Wilk, S.L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist, 49,* 929-954.

Schmidt, F.L. (1988). The problem of group differences in ability test scores in employment selection. *Journal of Vocational Behavior, 33,* 272-292.

Schmitt, N., Pulakos, E.D., Nason, E., & Whitney, D.J. (1996). Likability and similarity as potential sources of predictor-related criterion bias in validation research. *Organizational Behavior and Human Decision Processes, 68,* 272-286.

Schroeder, D.J., Broach, D., & Young, W.C. (1993). *Contributions of personality to the prediction of success in initial air traffic control specialist training.* (DOT/FAA/AM-93/4). Washington, DC: Federal Aviation Administration Office of Aviation Medicine. (NTIS No. ADA264699)

Sells, S.B., Dailey, J.T., & Pickrel, E.W. (Eds.) (1984), *Selection of air traffic controllers.* (DOT/FAA/AM-84/2). Washington, DC: Federal Aviation Administration Office of Aviation Medicine. (NTIS No. ADA147765)

Seymour, R.T. (1988). Why plaintiffs' counsel challenge tests, and how they can successfully challenge the theory of "validity generalization." *Journal of Vocational Behavior, 33,* 331-364.

Siegel, A.I. (1978). Miniature job training and evaluation as a selection/classification device. *Human Factors, 20,* 189-200.

Siegel, A.I. (1983). The miniature job training and evaluation approach: Additional findings. *Personnel Psychology, 36,* 41-56.

Siegel, A.I., & Bergman, B.A. (1975). A job learning approach to performance prediction. *Personnel Psychology, 28,* 325-339.

Society for Industrial and Organizational Psychology. (1987). *Principles for the validation and use of employee selection procedures.* (3rd ed.). College Park, MD: Author.

SPSS, Inc. (1989). *SPSS-X User's Guide.* Chicago, IL: Author.

U.S. Congress. (January 20, 1976). *House Committee on Government Operations: Recommendations on air traffic control training.* Washington, DC: Government Printing Office.

VanDeventer, A.D. (1983). Biographical profiles of successful and unsuccessful air traffic control specialist trainees. In A.D. VanDeventer, D.K. Taylor, W.E. Collins, & J.O. Boone (Eds.), *Three studies of biographical factors associated with success in air traffic control specialist screening/training at the FAA Academy* (pp. 1-5). (DOT/FAA/AM-83/6). Washington, DC: Federal Aviation Administration Office of Aviation Medicine. (NTIS No. ADA128784/6)

Wigdor, A.K., & Garner, W.R. (Eds.) (1982). *Ability testing: Uses, consequences, and controversies: Part 1. Report of the committee.* Washington, DC: National Academy Press.

Wigdor, A.K., & Sackett, P.R. (1993). Employment testing and public policy: The case of General Aptitude Test Battery. In H. Schuler, J.L. Farr, & M. Smith (Eds.), *Personnel selection and assessment: Individual and organizational perspectives* (pp. 183-204). Hillsdale, NJ: Erlbaum.

Zeisel, H. (1957). *Say it with figures.* (4th Ed.). New York: Harper & Row.