DOT/FAA/AM-96/13

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

Office of Aviation Medicine
Washington, D.C. 20591

Willie C. Young
Dana Broach
William L. Farmer

Civil Aeromedical Institute
Federal Aviation Administration
Oklahoma City, Oklahoma 73125

April 1996

Final Report

U.S. Department of Transportation
Federal Aviation Administration

Technical Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/AM-96/13		

4. Title and Subtitle	5. Report Date
Differential prediction of FAA Academy performance on the basis of gender and written Air Traffic Control Specialist aptitude test scores	April 1996
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Young, W.C., M.Ed.; Broach, D., Ph.D.; and Farmer, W.L., M.S.	

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
FAA Civil Aeromedical Institute P.O. Box 25082 Oklahoma City, OK 73125	
	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 ws conducted under task AM-92-B-HRR-128

16. Abstract

The technical fairness of the written air traffic control specialist (ATCS) aptitude test battery was investigated within the framework of the Uniform Guidelines on Employee Selection Procedures (29 CFR 1607). First, the adverse impact of using a composite score on the test battery was evaluated on an archival sample of 170,578 job applicants. Mean score differences by gender of 0.35 SD in the favor of men were found, suggesting that adverse impact on women could be expected from use of test scores in selection. Analysis of selection rates by gender found that women were classified as eligible for employment consideration on the basis of composite test scores at a lower rate (38.5%) than men (50.4%). 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 Academy in a sample of 9,552 first-time competitive entrants. Analysis based on correlations corrected for explicit and implicit restriction in range found significant differences in slopes and intercepts by gender, suggesting that separate regression equations were appropriate to predict Academy performance for the genders. The practical significance of the statistically small effect sizes is considered within the framework of current equal employment opportunity legislation and case law in the discussion of these results. The alternative explanation that these results reflect gender differences in job-related abilities, rather than test bias, is considered.

17. Key Words	18. Distribution Statement
ATCS selection; Test bias; Hierarchical regression; Adverse impact, Fairness; Equal employment opportunity; Personnel selection	Document is available to the public through the National Technical Information Service, Springfield, VA 22161

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

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

DIFFERENTIAL PREDICTION OF FAA ACADEMY PERFORMANCE ON THE BASIS OF GENDER AND WRITTEN AIR TRAFFIC CONTROL SPECIALIST APTITUDE TEST SCORES

The Federal Aviation Administration (FAA), in its 1993 Diversity Plan, made a commitment to attract, retain, develop, and manage a diverse work force that visibly reflected the American population at large by the year 2000. Achieving this goal will require substantial changes in the demographic profile of the Air Traffic Control Specialist (ATCS) occupation, the single largest (17,000) and most publicly visible occupational group in the agency. Air traffic control is a career field in which female workers have been historically under-represented relative to the American population at large. Entry into the occupation has been determined since 1981 by applicant performance on a written aptitude test battery administered by the US Office of Personnel Management (OPM; Aul, 1991). This test battery emphasized the organization, definition, and manipulation of the perceptual field through verbal and numeric reasoning (Harris, 1986). Yet, it is exactly such a test battery of cognitive abilities that may have been an inadvertent device for the exclusion of women from this traditionally male occupation.

Our purpose in this paper was to examine the technical fairness of the written ATCS aptitude test battery as the first step toward assessing to what degree, if any, that the battery may have served as an "engine of exclusion" (Seymour, 1988) of women from the ATCS occupation. By technical fairness, we are referring to the regression model of test bias for which there is a reasonable professional consensus, as embodied in the 1985 *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education), rather than a socially constructed standard regarding test use (Sackett & Wilk, 1994; Gottfredson, 1994). Technical fairness in this sense, and under the *Uniform Guidelines on Employee Selection Procedures* (29 CFR 1607), 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" (29CFR 1607.4.D). Second, where use of a selection procedure results in adverse impact, the *Uniform Guidelines* require that the test user 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)). This study investigated the technical fairness of the ATCS written aptitude test battery toward women from two perspectives: adverse impact and differential prediction.

ADVERSE IMPACT ANALYSIS

Previous research on written ATCS selection tests suggested that mean score differences by gender were insignificant (Rock, Dailey, Ozur, Boone, & Pickrel, 1984a, pp. 476) and that, overall, "the evidence for adverse impact against women based on this sample was marginal, at best" (Rock, Dailey, Ozur, Boone, & Pickrel, 1984b, pp. 507). This conclusion was based primarily on results of their 1984 study in which 57% of men (n = 3835) passed the screen in comparison to 45% of women (n = 1473). The adverse impact ratio in this case was 0.78 rather than the 0.80 required under the "four-fifths rule of thumb."

In the present study, we hypothesized that the composite of scores earned on the written ATCS aptitude test battery, as used by OPM to determine eligibility for employment, had no adverse impact on women applicants. The composite of test scores would be considered technically fair if there was no adverse impact arising from its use as a personnel selection device.

METHOD

Sample

The adverse impact analysis was based on determinations of eligibility for employment made by OPM for job applicants on the basis of a composite of two written test scores. Over 200,000 job applicants have taken the OPM written ATCS test battery since 1981. Records for 170,578 applicants with complete test scores were available in the data base. These records, as provided by OPM, included test raw scores, gender, education, and a determination of eligibility for employment based on test scores; racial identification data were not available. Demographic characteristics for this reference population of applicants are presented in Table 1. Gender (SEX), as indicated by OPM values, was recoded as 0 for males, and 1 for females.

Selection test scores

The selection test score used by OPM to determine eligibility for employment was a composite of scores earned on two written ATCS aptitude tests: the Multiplex Controller Aptitude Test (MCAT), and the Abstract Reasoning Test (ABSR). The development, psychometric characteristics, and validity of these written ATCS aptitude tests have been extensively described elsewhere (Brokaw, 1984; Collins, Boone, & VanDeventer, 1984; Manning, 1991; Sells, Dailey, & Pickrel, 1984). Scoring of the tests was done initially by summing the MCAT (weighted 2) and ABSR (weighted 1) scores. The resulting weighted scores were then transformed via an OPM test score transmutation table into the Transmuted Composite Score (TMC). About half of all applicants were expected to score at or above the mean on this composite (Rock, Dailey, Ozur, Boone, & Pickrel, 1984a). Applicants with 3 years of general experience, 4 years of college, or any combination of education and experience equating to 3 years of general experience and without prior aviation experience, were required to earn a TMC of at least 75.1 to be eligible for employment. Applicants with specific air traffic control-related aviation experience, or 4 years of college plus 1 year of graduate study, were eligible for employment if they earned a TMC of at least 70. In other words, a cut score of 75.1 or 70 on TMC, depending on applicant background, was used to determine eligibility for employment. Applicants not meeting these criteria were ineligible for employment as controllers. The determination of eligibility for employment was made by OPM. Codes indicating that an applicant had either failed the test ('IA') or scored too low for consideration ('IS'), based on TMC, were recoded as test failures. All other ineligibility codes were recoded as "other ineligible," and codes indicating eligibility were recoded as "eligible" for employment. The adverse impact analysis was based on this eligibility variable.

Procedure

The adverse impact analysis was conducted in two steps. First, TMC distributions were analyzed by gender; a *t*-test was used to evaluate mean score differences. Second, selection rates on the basis of eligibility codes by gender were evaluated. The proportion of applicants determined to be eligible on the basis of their test scores was compared to the proportion ruled as ineligible on the basis of test scores; applicants determined to be ineligible on any other basis (e.g., age, salary requirements, experience, or education) were excluded from the analysis. Fisher's Z test was used to statistically compare selection rates.

RESULTS

Group differences

Analysis of group differences in predictor scores by gender are presented in Table 2. Males earned significantly higher mean TMC scores (M = 74.44, SD = 14.17) than females (M = 69.32, SD = 14.37; $t(170{,}576) = 61.75$, $p \leq .001$). The distribution of TMC by sex in the research sample is illustrated in Figure 1. The standardized effect size (d) for gender on TMC scores is 0.35 SD, corresponding to a small to medium effect size (Cohen, 1988). This contrasts with previous research suggesting that mean differences in TMC were insignificant (Rock, Dailey, Ozur, Boone, & Pickrel, 1984b). Mean score differences might be expected to translate into differences in selection rates by gender. In the ATCS selection process, only candidates scoring at or above the average TMC were eligible for employment.

Table 1

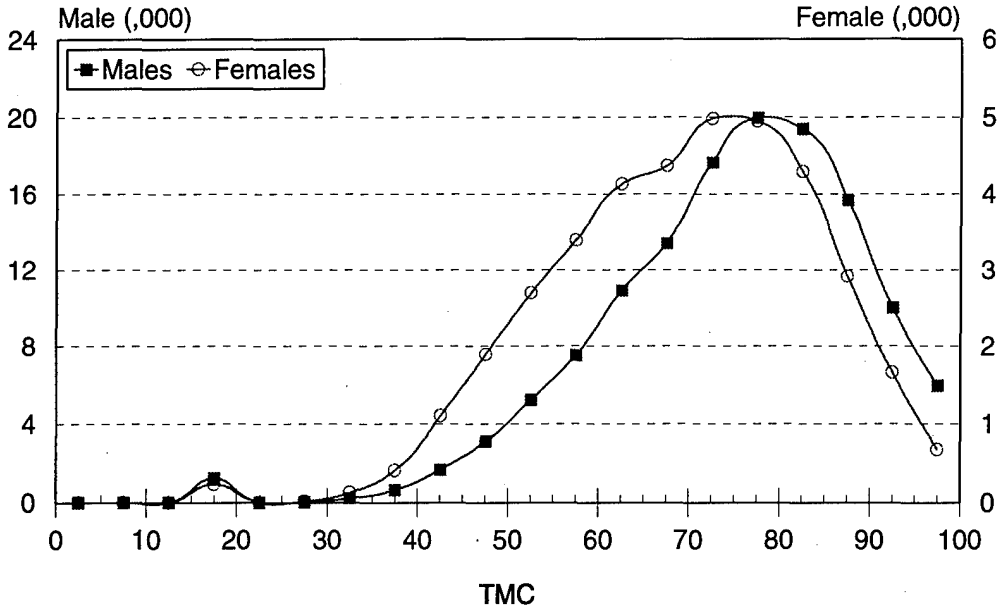
Demographic characteristics for reference population, all 1985-1992 FAA Academy entrants, and research sample

Characteristic	Reference Population (N = 170,578)	FAA Academy entrants	
		Academy entrants (N=14,392)	Sample (N=9,552)
Sex			
Male	132,708	11,460	7,935
Female	37,870	2,932	1,617
Race			
Asian/Pac-Island		91	59
American Indian		195	131
African American		819	283
Hispanic		525	293
White		12,366	8,555
Missing Data		396	231
Education			
LT High School	404		
High School	28,147	1,576	1,046
Some college	82,414	7,750	5,351
Bachelor's degree	54,583	4,745	3,033
Advanced degree	3,934	176	116
Missing Data	1,096	145	6
Age			
Mean		26.01	25.77
SD		2.99	2.85

Notes: Racial identification and age data not available for reference population of all applicants.

Figure 1

Predictor (TMC) score distribution by gender for reference population of applicants (N = 170,578)



Selection ratio

The results of the adverse impact analysis by gender, based on OPM eligibility codes, are presented in Table 3. Approximately half (52.04%) of the 132,708 men were considered eligible for employment on the basis of their aptitude test scores. With a majority selection rate of about 50% and mean differences of .35 *SD*, we anticipated a selection rate for women of about 35 to 38%, based on Sackett and Wilk (1994). In fact, 38.46% of the 37,870 women were eligible for employment on the basis of the aptitude test scores. The proportion of women determined to be eligible for employment was significantly less than the proportion of men (Z = -46.63, $p \leq .001$); the ratio of female to male selection rates was .74. Using the 4/5ths rule of thumb of the *Uniform Guidelines*, it appeared that use of scores on the written ATCS aptitude test battery to determine eligibility for employment resulted in statistically significant adverse impact against female applicants.

DIFFERENTIAL PREDICTION ANALYSIS

Given the finding that there appeared to be adverse impact against women, the *Uniform Guidelines* (29 CFR 1607.14.B.(8).(b)) and *Standards for Educational and Psychological Testing* (Standard 1.20, p. 17) required an investigation of the relationship between test scores and job performance for evidence of differential prediction by subgroup. We hypothesized that there was no difference in the predictive validity of the test battery by gender.

METHOD

Sample

The differential prediction analysis was based on a sample of persons actually hired by the FAA on the basis of their aptitude test scores. Between October 1985 and January 1992, a total of 14,392 ATCS candidates entered the FAA Academy. The majority (11,405) had competed under civil service regulations for hire and were entering the Academy for the first time. Complete gender, racial identification, predictor, and criterion data were available for the research sample of 8,842 male and female students. There were 7,332 (82.9%) men and 1,510 (17.1%) women in the sample. Demographic information for all Academy entrants and the research sample is presented in Table 1. As with the reference applicant population, gender (SEX) was coded as 0 for males, and 1 for females.

Table 2

Mean predictor and criterion score differences by gender

Variable	Group	*N*	*M*	*SD*	*SE*	*t*	*df*
TMC	Males	132,708	74.44	14.17	0.039	61.75***	170,576
	Females	37,870	69.32	14.37	0.074		
SCREEN	Males	10,252	72.75	11.70	0.116	9.51***	12,754
	Females	2,504	70.26	12.00	0.240		

***$p \leq .001$

Table 3

Adverse impact analysis for reference pop. by gender for based on OPM eligibility codes

	Sex		
OPM Eligibility	Males	Females	Row totals
Eligible	69,056 (52.04%)	14,564 (38.46%)	84,070
Failed test	49,902 (37.60%)	20,077 (53.02%)	69,979
Other ineligible	13,750 (10.36%)	3,229 (8.53%)	16,979
Column totals	132,708	37,870	170,578

Measures

Predictors. 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 and/or military service. Descriptive statistics for the predictor scores are presented in Table 4 for the reference population of applicants, all Academy entrants, and the research sample.

Criterion. The criterion in the differential prediction analysis was performance in the FAA Academy initial ATCS training program, known as the ATCS Nonradar Screen ("the Screen"). Training may be used as a criterion measure where success in training is "properly measured," and the relevance of the training can be demonstrated through comparison of training content to critical or important job behaviors or by showing that training measures are related to subsequent measures of job performance (29 **CFR** 1607.14.B.(3)). The Screen was originally established in response to recommendations made by the US Congressional House Committee on Government Operations (US 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. Performance in the Screen has been shown to predict subsequent performance in radar-based training 1 to 2 years after entry into the occupation (Broach & Manning, 1994) as well as completion of the rigorous on-the-job training sequence and certification as a qualified "full performance level" controller (Della Rocco, Manning, & Wing, 1990; Manning, Della Rocco, & Bryant, 1989).

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, Manning, & Wing, 1990). The final summed composite score (SCREEN) was weighted 20% for academics, 60% for laboratory simulations, and 20% for the final examination. A minimum SCREEN score of 70 was required to pass the Academy program. This final composite score was the criterion measure in this study. Descriptive statistics for SCREEN scores are also presented in Table 4 for all Academy entrants and for the research sample.

Procedure

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. A 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 provide a significantly better fit to the data.

Our step-down analysis was conducted as follows, using the SPSS (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, 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 dummy-coded variable indicating group membership (group model), and compared to 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

Table 4

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

	Reference population (N=170,578)				All Academy entrants (N=14,392)				Research sample (N=9,552)			
Variable	*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	91.55	5.03	70.00	100.00
SEX	0.22	0.42			0.20	0.40			0.17	0.38		
TMC_SEX	15.39	29.60	0.00	100.00	16.18	34.80	0.00	100.00	15.52	34.43	0.00	100.00
SCREEN					72.26	11.80	27.16	99.47	71.68	11.36	27.16	97.59

Notes: Screen score not applicable for reference population of job applicants.

of aptitude and group membership (cross-product model). The cross-product model was then compared to 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 to the basic model; a significant change in R^2 indicated only intercept differences between groups. The general logic and associated SPSS syntax for the step-down hierarchical regression analysis are illustrated in Figure 2.

Technical feasibility

Restriction in range, statistical power, and criterion bias are considerations in evaluating the technical feasibility of a test fairness investigation that must be explicitly considered under the *Uniform Guidelines* (29 **CFR** 1607.14.B.(8).(c) and (e); 29 **CFR** 1607.16.U). Both explicit and incidental restriction in range are recurrent problems in ATCS selection research, as evidenced by the sample sizes and descriptive statistics in Table 4. Variance in TMC for the research sample was explicitly restricted in range due to selection. Therefore, correlations between TMC and the SCREEN criterion were corrected with respect to the reference population of 170,578 applicants, using the formula presented by Ghiselli, Campbell, and Zedeck (1981, p. 299). Correlations between variables indicating gender and the criterion were incidentally restricted in range. These gender-criterion correlations, including the gender-by-predictor crossproduct to SCREEN correlation, were corrected with respect to the reference population of 170,578 applicants using the Ghiselli, Campbell, and Zedeck (1981, p. 304) formula for incidental range restriction. Finally, values for the population correlations between gender and the gender-predictor crossproduct were computed. The overall structure of the correlation matrix is described in Table 5; sample and corrected correlations are presented in Table 6. Sample correlations, without corrections for restriction in range, are presented in the lower left-hand corner, while corrected and population correlations are presented in the upper right-hand corner of the overall matrix. Separate differential prediction analyses were conducted on the basis of sample and corrected correlations, as required by the *Uniform Guidelines* (29 **CFR** 1607.15.B.(8)).

Sample sizes in these analyses were clearly of sufficient size to provide more than enough power to detect even small statistical effects. We estimated the

Table 5

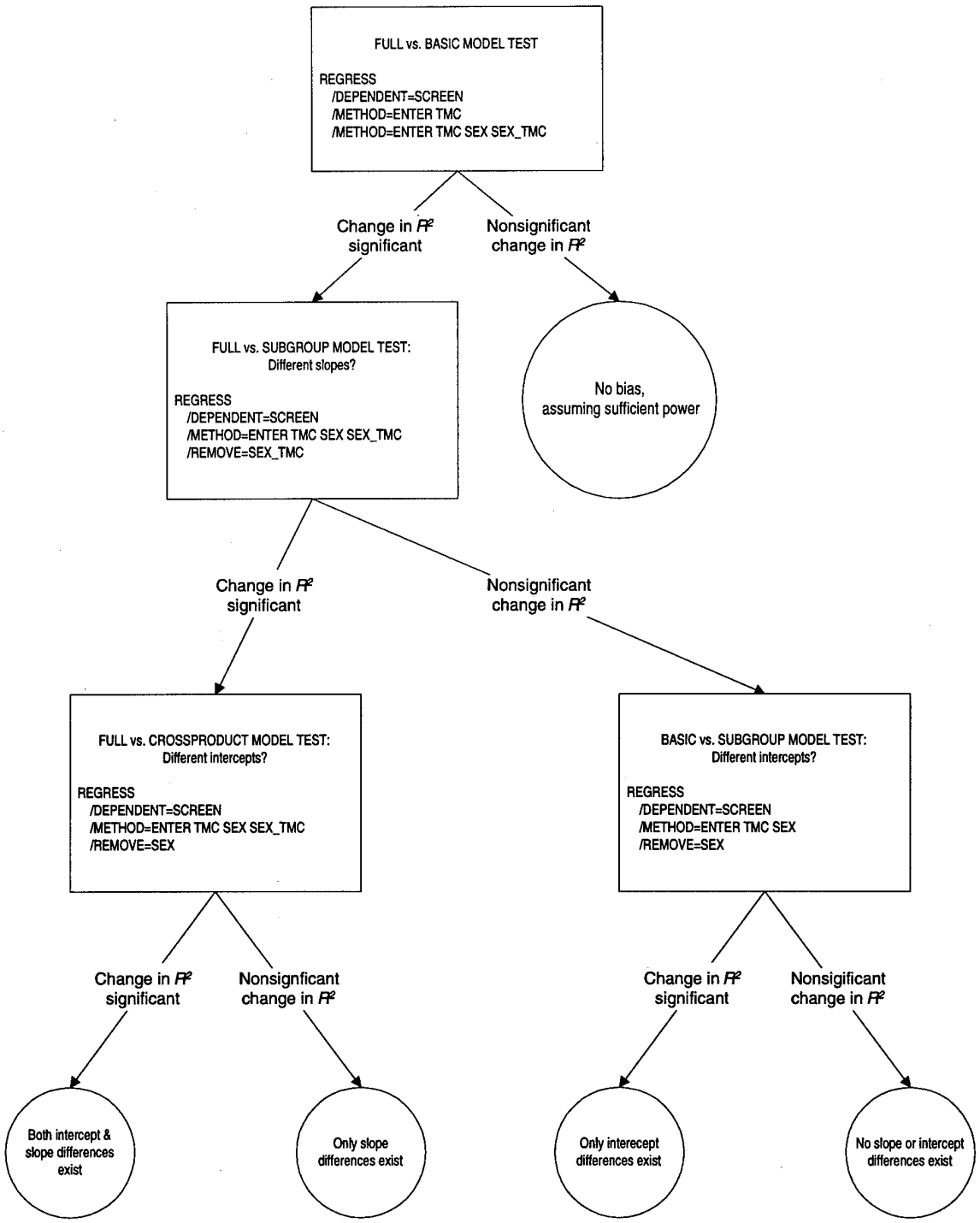
Correlation matrix structure for differential prediction analysis

	TMC	SEX	TMC_SEX	SCREEN
TMC		r_p	r_p	r_e
SEX	r_s		r_p	r_i
TMC_SEX	r_s	r_s		r_i
SCREEN	r_s	r_s	r_s	

Note: Sample correlation matrix structure shown below the diagonal, corrected matrix structure above the diagonal. r_s = sample correlation; r_p = population correlation, where population is reference population of all applicants; r_e = sample correlation corrected for explicit restriction in range, based on reference population of all applicants; r_i = sample correlation corrected for incidental restriction in range, based on reference population of all applicants

Figure 2

Step-down hierarchical regression analysis logic and SPSS syntax



available statistical power using Cohen's (1988) regression power tables (Table 9.3.1) for as many as 3 independent variables at an alpha of .01. The risk of a type II error (failing to find an effect that in fact was present) was very low, with a .98 probability of detecting even very small effect sizes ($f^2 \leq .01$) with a sample of more than 8,000 cases.

Finally, as noted by Sackett and Wilk (1994), Lautenschlager and Mendoza (1986), as well as by the *Uniform Guidelines* (29 CFR 1607.16.U), the feasibility of an assessment of technical fairness depends upon the quality of the job-relevant criterion: If the criterion was systematically biased against women, for example, then the regression-based method could not be used to determine the presence or absence of differential prediction by subgroup. The distribution of criterion scores by gender is shown in Figure 3. Observed mean score differences in SCREEN by gender were about 0.23 *SD* for the research sample. These results are somewhat less than the estimated differences of .3 to .4 *SD* by race reported by Ford, Kraiger, and Schectman (1986). There are several possibilities for these observed differences by gender: these mean criterion differences may represent some degree of "systematic bias" against women; the seeming bias may have been confounded, at least in part, with differences attributable to selection; these differences may reflect true distinctions in performance that were incidental to sex; and the apparent bias may be due to differences in information processing strategies used by the sexes. However, the limits of the data available for this study did not permit a definitive evaluation of these alternatives. Therefore, we cannot, with certainty, claim an unbiased criterion. Yet, in accordance with the *Uniform Guidelines*, Screen score was properly measured, and related to subsequent organizationally valued outcomes. It was also no more biased than measures used in previous published selection test fairness studies. Therefore, it was an appropriate criterion in this assessment of technical fairness under the *Uniform Guidelines.*

RESULTS

Without corrections for restriction in range

The adverse impact analysis suggested that use of TMC to determine eligibility for employment as an air traffic controller may have contributed to a situation of adverse impact against women. The focus of an evaluation of technical fairness, therefore, shifted to the degree to which the predictor score differentially predicted the criterion. Sample correlations, without corrections for restriction in range, are presented in the lower left-hand triangle of the matrix in Table 6. TMC was significantly correlated with final score in the Academy Screen ($r = .1844$, $p < .001$) and slightly with the predictor-group crossproduct ($r = .0296$, $p < .01$). Gender (SEX) was negatively correlated with the criterion SCREEN score ($r = -.0847$, $p < .001$), where gender was coded as 1 for females and 0 for males. The results of the differential prediction analysis using the step-down hierarchical regression analysis on the basis of the sample correlation matrix without any 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 first 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 crossproduct), rather than the basic model (predictor only), was significant ($\Delta R^2 = .008$, $\Delta F = 38.60$, $p \leq .001$). Next, the null hypothesis of same slopes by gender could not be rejected; the subgroup model (predictor and group membership) did not explain any less variance than the full model ($\Delta R^2 = 0$, $\Delta F = 1.02$, *ns*). Following the analytic logic illustrated in Figure 1, the basic and subgroup models were next compared to determine if the intercepts were different for men and women. The null hypothesis of same intercepts was rejected, with removal of SEX leading to a significant reduction in the amount of explained variance ($\Delta R^2 = -.008$, $\Delta F = 76.18$, $p \leq .001$). Overall, the results obtained with the uncorrected correlations indicated significant intercept differences, but no differences in slopes by gender.

Table 6

Sample and corrected correlation matrix for differential prediction analysis by gender

	TMC	SEX	TMC_SEX	SCREEN
TMC		-0.1479***	-0.0361***	0.4724
SEX	0.0106		0.9735***	-0.0661
TMC_SEX	0.0296**	0.9986***		-0.0408
SCREEN	0.1844***	-0.0877***	-0.0847***	

Note: As described in Table 5, correlations between SCREEN and TMC, SEX, and TMC_SEX in the upper right-hand corner are corrected, and therefore, no significance tests are reported

$^{**}p \leq .01$, $^{***}p \leq .001$

Figure 3

Criterion (SCREEN) score distribution by gender in research sample (N = 12,756)

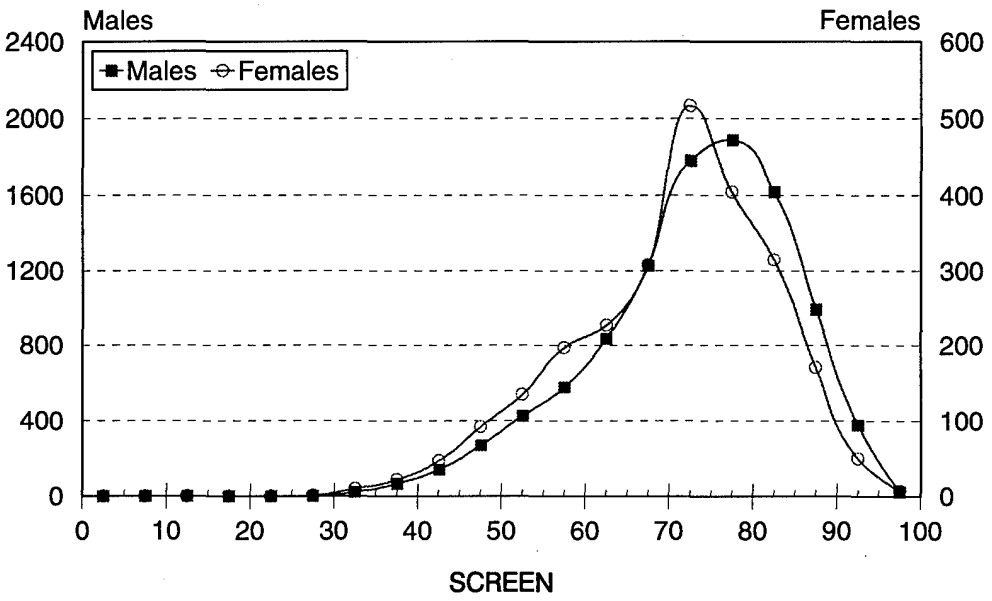


Table 7

Results of step-down hierarchical regression analysis,for test bias in research sample by gender on basis of correlation matrix without corrections for restriction in range

Analsysis	Model	R^2	ΔR^2	ΔF	F
Basic v. Full: Overall bias	TMC	0.037			339.94^{***}
	TMC + SEX + TMC_SEX	0.045	0.008	38.60^{***}	140.01^{***}
Full v. Group: Slopes	TMC + SEX + TMC_SEX	0.045			140.01^{***}
	TMC + SEX	0.045	-0.000	1.02	209.50^{***}
	TMC				
Full v. Crossproduct[a]	TMC + SEX + TMC_SEX	N/A			
	TMC + TMC_SEX				
Basic v. Group: Intercepts	TMC + SEX	0.045			209.50^{***}
	TMC	0.037	-0.008	76.18^{***}	339.94^{***}

Notes: [a]Full v. crossproduct model comparison not conducted. See Figure 1 for logic and flow of step-down hierarchical regression analysis.

$^{***}p \leq .001$

With corrections for restriction in range

However, as shown in Table 2, the sample range of scores on the predictor was severely restricted; as a consequence, evidence based on those uncorrected correlations may be somewhat misleading as to the fairness 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 fairness of the OPM test battery with respect to the applicant population. Population and 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 .1844 to .4724 with correction for explicit restriction in range. After correcting for incidental restriction in range, the correlation between gender and SCREEN decreased to -.0661, as did the correlation between the crossproduct and SCREEN (-.0408). The results of the differential prediction analysis, using the step-down hierarchical regression analysis on the basis of the corrected correlations, are presented in Table 8. The null hypothesis of a common regression line was rejected, 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 (R^2 = .0187, F = 117.76, $p \leq .001$). Next, the null hypothesis of same slopes by gender was rejected; the subgroup model (predictor and group membership) explained less variance than the full model (R^2 = -.0187, F = 235.34, $p \leq .001$). Following the analytic logic illustrated in Figure 1, the full and crossproduct models were next compared to determine if the intercepts were different for men and women. The null hypothesis of same intercepts was also rejected, with removal of SEX leading to a significant reduction in the amount of explained variance (R^2 = -.0181, F = 228.41, $p \leq .001$). Overall, the results obtained with the corrected correlations indicated the need for separate regression equations for men and women for predictions based on both raw and standardized predictor scores. Therefore, correlations between TMC and SCREEN were computed for men and women separately, corrected for explicit restriction in range based on the standard deviation of the aptitude by sex (TMC_SEX) interaction term, and submitted to regression analysis. The equation for men was:

SCREEN' = -24.8680 + (1.0597 * TMC)

compared to and equation for women of

SCREEN' = -23.1254 + (1.0102 * TMC)

where SCREEN' is the predicted score in the FAA Academy ATCS Nonradar Screen. The regression equations are plotted in Figure 4.

DISCUSSION

Overall, the analyses reported in this study indicated that the written ATCS aptitude test battery did not fulfill the technical fairness requirements outlined by the *Uniform Guidelines on Employee Selection Procedures.* The results of the adverse impact analysis indicated that use of the weighted composite of MCAT and ABSR scores as a qualification criterion resulted in the exclusion of greater proportions of women than men from further consideration for employment. Moreover, the adverse impact could be attributed to a specific practice (*Antonio v. Ward's Cove Packing Co.*, 1989; *EEOC v. Greyhound Lines*, 1980; *Pouncy v. Prudential Insurance Co.*, 1982) and was statistically significant (*Hazelwood School District v. United States*, 1977). Similarly, there appeared to be subtly different relationships for the sexes between aptitude score and subsequent performance at the Academy, after correcting the sample data correlations for restriction in range. The corrected majority regression line slightly overpredicted the performance of the minority group, as shown in Figure 4. Schmidt (1988) suggested that this is a common finding in differential prediction analyses. For example, Dunbar and Novick (1988) reported similar results for predictions of training success from the Armed Services Vocational Aptitude Battery (ASVAB) scores by gender.

Evidence for differential prediction such as we found in this study of the ATCS written aptitude test battery has been discounted on the basis of factors such as use of inappropriate statistical procedures and defects in study designs (Hunter, 1973). On one hand, the statistical effects detected in our differential

Table 8

Results of step-down hierarchical regression analysis for test bias in research sample by gender on basis of correlation matrix corrected for restriction in range

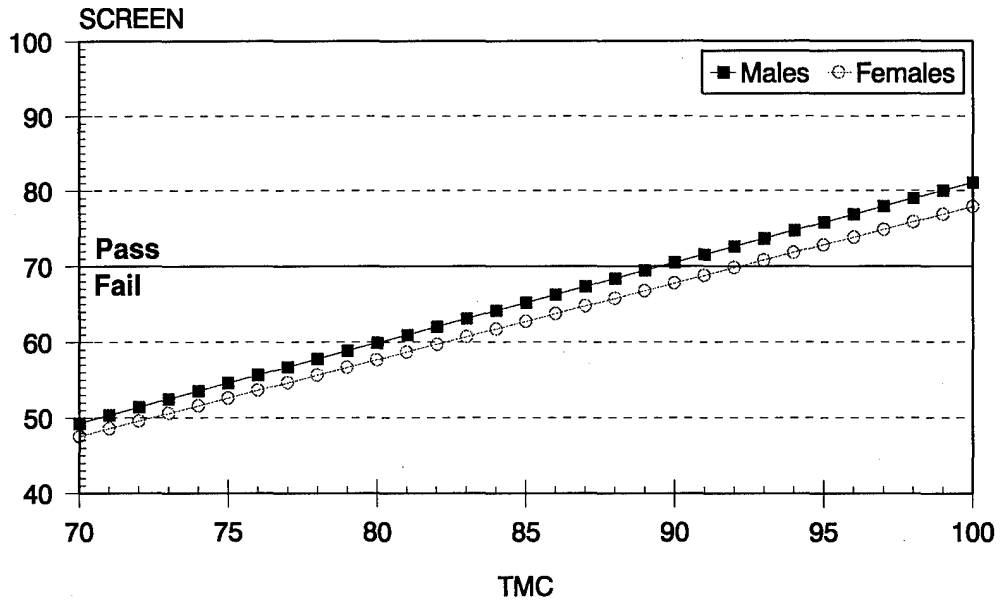
Analsysis	Model	R^2	ΔR^2	ΔF	F
Basic v. Full: Overall bias	TMC	0.2232			2743.42***
	TMC + SEX + TMC_SEX	0.2419	0.0187	117.76***	1015.34***
Full v. Group: Slopes	TMC + SEX + TMC_SEX	0.2419			1015.34***
	TMC + SEX	0.2232	-0.0187	235.34***	1371.68***
Full v. Crossproduct: Intercepts	TMC + SEX + TMC_SEX	0.2419			1015.34***
	TMC + TMC_SEX	0.2237	-0.0181	228.41***	1376.04***
Basic v. Group[a]	TMC	N/A			
	TMC + SEX				

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

***$p \leq .001$

Figure 4

Regression functions by gender in research sample



prediction analyses were generally small and detectable only with very large samples, after corrections for restriction in range. One might argue that, as a consequence, the results were artifacts of an inappropriate analysis of corrected correlations, and have little practical significance. We would counter by noting that corrected correlations may, in fact, provide more accurate estimates of test validity, particularly in large samples and under stringent selection ratios (Bobko, 1983; Millsap, 1988). 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 that provide less biased estimates of true population values are likely to provide similarly less biased estimates of population effects, and are not artifactual. Moreover, we believe that these effects cannot be lightly dismissed in view of the very real practical consequences for the ATCS selection program. One practical consequence of a mean score difference on TMC of 0.35 *SD* was adverse impact on women, as defined under the *Uniform Guidelines*. Moreover, the practical consequence of the apparent differential prediction in the population was that women may have effectively needed a higher TMC than men to have an equal likelihood of passing the FAA Academy. The implications of differential prediction relative to the consequences of overprediction will be investigated in greater detail in another study.

On the other hand, unmeasured variables may have been confounded with the predictor, resulting in a defective study design (Anastasi, 1988). One might suspect, for example, that education and scores on the 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. However, a significantly greater proportion of women (39.3% of 34,479) than men (35.7% of 118,735) had achieved a baccalaureate degree or more in the reference population of 170,578 applicants ($Z = 12.22, p \leq .001$). A similar pattern was found for the sample of 9,552 FAA Academy entrants used in the differential prediction analysis. These data provide some evidence to suggest, pending more detailed analyses, that unmeasured variables such as education may not account for the observed differential prediction in this study.

There is an alternative explanation to conclusions of test bias or artifactual results due to statistical corrections or unmeasured variables. The results might

accurately reflect true differences in capabilities and performance by gender. One recent analysis of the ATCS job found that perceptual processes such as visualization and scanning, are important worker requirements (Nickels, Bobko, Blair, Sands, & Tartak, 1995). Another analysis suggested perceptual speed and reasoning with numerical information were relevant cognitive attributes to the controller job (Broach & Aul, in preparation). An analysis of the abilities required specifically for success in the Screen also pointed toward the visual-spatial domain of abilities (Gibb, Smith, Swindells, Tyson, Gieraltowski, Petschauer, & Haney, 1991). These studies indicate that there is at least some need to utilize abilities in the visual-spatial domain in the performance of ATCS tasks. The construct validity study of the OPM test battery conducted by Harris (1986) provided evidence that the MCAT, in particular, measured some aspects of this domain of job-relevant abilities with its emphasis on the definition and manipulation of the perceptual field and reasoning with verbal and numeric information. There appear to be subtle but persistent sex differences in the visual-spatial abilities domain (Halpern, 1986; Voyer, Voyer, & Bryden, 1995). Sex differences in the abilities measured by the ATCS written aptitude test battery might explain, in part, the mean score differences observed on the predictor TMC. Similarly, sex differences on visual-spatial abilities important to performance in the Screen might similarly account for the apparent differential prediction of Screen scores from aptitude scores. Current research being conducted under the FAA Separation and Control Hiring Assessment (SACHA) project (Bobko, Nickels, Blair, & Tartak, 1994; University Research Corporation, 1994) may provide further data elucidating the relationships between gender, visual-spatial abilities as measured by aptitude tests, and ATCS job performance.

REFERENCES

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.

Antonio v. Ward's Cove Packing Co., 490 US 642 (1989).

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.

Bartlett, C. J., and O'Leary, B. S. (1969). A differential prediction model to moderate the effects of heterogeneous groups in personnel selection and classification. *Personnel Psychologist*, 22, 1-18.

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 Psychologist*, 31, 233 - 241.

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

Bobko, P., Nickels, B. J., Blair, M. D., & Tartak, E. L. (1994, March). *Preliminary internal report on the current status of the SACHA model and task interconnections: Volume I.* (Interim report delivered under FAA contract DTFA01-91-C-00032). Washington, DC: Federal Aviation Administration Office of Personnel.

Broach, D., & Aul, J. C. (In preparation). *Analysis of the ATCS job using the Position Analysis Questionnaire (PAQ).* Manuscript in preparation.

Broach, D., & Manning, C. A. (1994). *Validity of the air traffic control specialist nonradar 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.

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.

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-91/9). 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.

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.

EEOC v. Greyhound Lines, 635 F.2d 188 (3CA) (1980).

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.

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

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

Gibb, G. D., Smith, M. L., Swindells, N., Tyson, D., Gieraltowski, M. J., Petschauer, K. J., & Haney, D. U. (1991). *The development of an experimental selection test battery for air traffic control specialists.* Daytona Beach, FL: Embry-Riddle Aeronautical University Airway Simulation Laboratory.

Gottfredson, L. S. (1994). The science and politics of race-norming. *American Psychologist*, 49, 955 - 963.

Halpern, D. F. (1986). *Sex differences in cognitive abilities.* Hillsdale, NJ: Erlbaum.

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.

Hazelwood School District v. United States, 433 US 299 (1977).

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.

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.

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.

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

Nickels, B. J., Bobko, P., Blair, M. D., Sands, W. A., & Tartak, E. L. (1995, January). *Separation and Control Hiring Assessment (SACHA) final job analysis report.* (CDRL-007A delivered under FAA contract DTFA01-91-C-00032). Washington, DC: Federal Aviation Administration Office of Personnel.

Pouncy v. Prudential Insurance Co., 668 F.2d 795 (5CA) (1982).

Rock, D. B., Dailey, J. T., Ozur, H., Boone, J. O., & Pickrel, E. W. (1984a). Validity and utility of the ATC experimental tests battery. Study of Academy trainees, 1982. 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.

Rock, D. B., Dailey, J. T., Ozur, H., Boone, J. O., & Pickrel, E. W. (1984b). 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.

Sackett, P. R., & Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in pre employment 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.

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.

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.

☆U.S. GOVERNMENT PRINTING OFFICE: 1996 - 760-082/20075