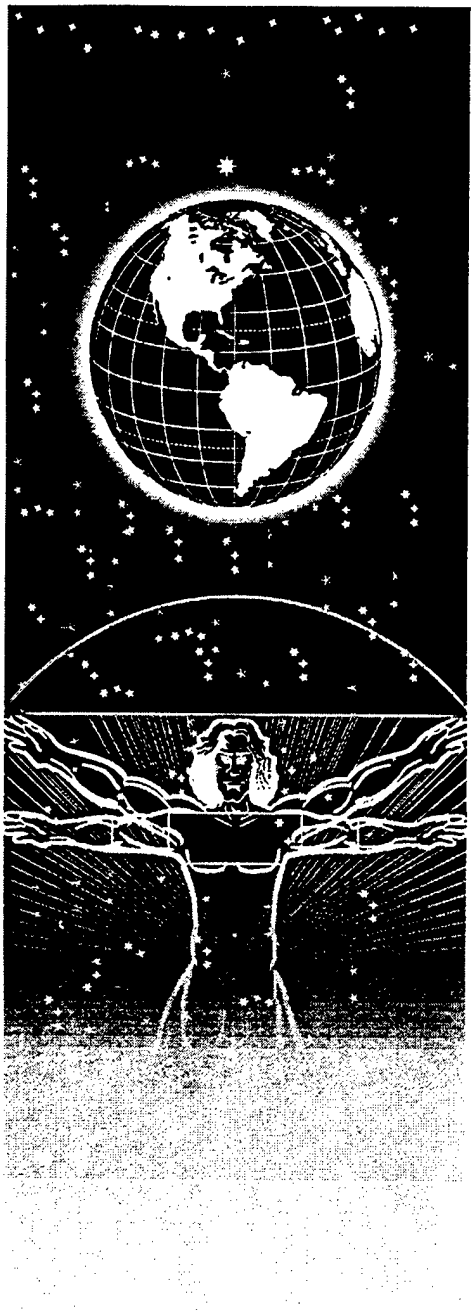
AL/HR-TP-1997-0016



UNITED STATES AIR FORCE RESEARCH LABORATORY

INTERCHANGEABILITY OF VERBAL AND QUANTITATIVE SCORES FOR PERSONNEL SELECTION: AN EXAMPLE

Malcolm J. Ree
Cognition and Performance Division

Thomas R. Carretta
Training Effectiveness Branch

HUMAN EFFECTIVENESS DIRECTORATE
WARFIGHTER TRAINING RESEARCH DIVISION
7909 Lindbergh Drive
Brooks AFB TX 78235-5352

September 1998

AIR FORCE MATERIEL COMMAND
AIR FORCE RESEARCH LABORATORY
HUMAN EFFECTIVENESS DIRECTORATE
WARFIGHTER TRAINING RESEARCH DIVISION
6001 South Power Road, Building 558
Mesa AZ 85206-0904

NOTICES

Publication of this paper does not constitute approval or disapproval of the ideas or findings. It is published in the interest of scientific and technical information (STINFO) exchange.

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder or any other person or corporation, or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Office of Public Affairs has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

THOMAS R. CARRETTA
Project Scientist

DEE H. ANDREWS
Technical Director

LYNN A CARROLL, Colonel, USAF
Chief, Warfighter Training Research Division

Please notify AFRL/HEOP, 2509 Kennedy Drive, Brooks AFB TX 78235-5118, if your address changes, or if you no longer want to receive our reports. You may write or call the STINFO Office at DSN 240-3877 or Commercial (210) 536-3877; or e-mail Shirley.Walker@platinum.brooks.af.mil.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 1998	Interim - September 1996 to February 1997

4. TITLE AND SUBTITLE

Interchangeability of Verbal and Quantitative Scores for Personnel Selection: An Example

5. FUNDING NUMBERS

PE - 62205F
PR - 1123
TA - B1
WU - 01

6. AUTHOR(S)

Malcolm J. Ree
Thomas R. Carretta

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Training Research Division, Training Effectiveness Branch
7909 Lindbergh Drive
Brooks AFB TX 78235-5352

8. PERFORMING ORGANIZATION

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Training Research Division
6001 South Power Road, Bldg 558
Mesa AZ 85206-0904

10. SPONSORING/MONITORING

AL/HR-TP-1997-0016

11. SUPPLEMENTARY NOTES

Air Force Research Laboratory Technical Monitor: Dr Thomas R. Carretta, (510) 536-3956

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Even though tests or composite tests may have the same name, that is no guarantee of similarity of construct. A sample of 7,940 military participants took both the Air Force Officer Qualifying Test (AFOQT) and the Scholastic Aptitude Test (SAT). The scores from the verbal and quantitative sections of the AFOQT were correlated with the verbal and quantitative scores from the SAT. Correlations were very high, approaching 1.0. An Eigenvalue analysis revealed one very large factor and several smaller factors. These analyses indicated a great similarity between the verbal and quantitative sections of the AFOQT and the SAT suggesting high interchangeability for these content areas. Additional study is necessary, such as extension of the similarity analyses to AFOQT Pilot and Navigator-Technical composites.

14. SUBJECT TERMS

Ability testing; AFOQT; Air Force Officer Qualifying Test; Correlation; Personnel measurement; Personnel selection; Regression; SAT; Scholastic Aptitude Test

15. NUMBER OF PAGES

11

16. PRICE CODE

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

CONTENTS

	Page
SUMMARY	1
INTRODUCTION	1
METHOD	1
Participants	1
Measures	2
Analyses	2
RESULTS AND DISCUSSION	2
REFERENCES	5

TABLES

Table No.		
1	Means, Standard Deviations and Correlations of the Scores	3
2	Eigenvalue Analysis of the Scores	3

PREFACE

This project was conducted under Work Unit 1123-B1-01, Pilot Selection and Classification Support, which is dedicated to research into the selection and classification of United States Air Force aircrew personnel. Work Unit Monitor was Dr Thomas R. Carretta. We thank N. Dorans and H. Wainer for their help in this effort.

Send correspondence and requests for reprints to the first author at AFRL/HEAI, 7909 Lindbergh Drive, Brooks Air Force Base, TX 78235-5352. Send electronic mail to REE@ALHRM.BROOKS.AF.MIL.

INTERCHANGEABILITY OF VERBAL AND QUANTITATIVE SCORES FOR PERSONNEL SELECTION: AN EXAMPLE

SUMMARY

Even though tests or composites of tests may have the same name, that is no guarantee of similarity of construct. A sample of 7,940 military participants took both the Air Force Officer Qualifying Test (AFOQT) and the Scholastic Aptitude Test (SAT). The scores from the verbal and quantitative sections of the AFOQT were correlated with the verbal and quantitative scores from the SAT. Correlations were very high, approaching 1.0. An Eigenvalue analysis revealed one very large factor and several smaller factors. These analyses indicated a great similarity between the verbal and quantitative sections of the AFOQT and the SAT suggesting high interchangeability for these content areas. Additional study is necessary, such as extension of the similarity analyses to AFOQT Pilot and Navigator-Technical composites.

INTRODUCTION

Testing job candidates is very expensive. Many organizations could be looking for cheaper ways of assessing ability of job applicants. For certain jobs, one way would be to allow the applicant to submit results of standardized tests taken for other purposes, such as college entrance exams or military qualification tests.

The United States Air Force (USAF), which tests about 15,000 officer applicants per year, provides an example of how this might be accomplished. The Air Force Officer Qualifying Test (AFOQT) is a multiple-aptitude battery measuring general cognitive ability and five lower-order factors: verbal, mathematics, spatial, perceptual speed, and pilot job knowledge (Carretta & Ree, 1996). The Scholastic Aptitude Test[1] (SAT) measures general cognitive ability and verbal and quantitative ability. Recently, within the U. S. Air Force, there has been interest in the potential interchangeability of these two tests for personnel selection. However, certain senior managers expressed the opinion that the two tests would not be the same because they come from different organizations. The purpose of this experiment was to examine the extent to which these tests measure the same constructs, despite differences in authorship, item construction rules, and test content taxonomy.

METHOD

Participants

The participants were 7,940 young men and women enrolled in the United States Air Force Reserve Officer Training Corps. All were either in college or about to start college when they were administered the AFOQT and applying to college when they were administered the SAT.

[1] Now called the Scholastic Achievement Test

Because the participants have been subject to prior selection, the scores create a range restricted sample (Ree, Carretta, Earles, & Albert, 1994; Thorndike, 1949).

Measures

The AFOQT is a 16-test, multiple-aptitude battery used for selection into military training. It is constructed and used by the U. S. Air Force. Its factor structure (Carretta & Ree, 1996) and its validity for pilot selection (Carretta & Ree, 1995) have been examined. Only the verbal and mathematics tests were used in this study. The other AFOQT tests were not used as they include specialized knowledge and abilities not claimed by the SAT.

The AFOQT verbal tests are Verbal Analogies (VA), Reading Comprehension (RC), and Word Knowledge (WK). The mathematics tests are Arithmetic Reasoning (AR), Data Interpretation (DI), and Mathematics Knowledge (MK). These tests are operationally used as two aggregated composites: Verbal (VA, RC, WK) and Quantitative (AR, DI, MK). The reliability of the composites of the AFOQT were provided by Carretta and Ree (1997). They give .88 as the reliability of the Verbal composite and .84 as the reliability of the Quantitative composite. These were adjusted to the normative reliability using the procedure given by Gulliksen (1950). The adjusted reliability of the Verbal composite was .90 and the adjusted reliability of the Quantitative composite was .86.

The SAT is a two-test battery developed under contract by the Educational Testing Service. It is used by many colleges and universities as part of their admission process. Its two parts are verbal and quantitative. The reliability of the SAT as provided by Donolon and Livingston (1984, pp. 33-34) was .93 and .92 for the verbal and quantitative scores, respectively.

Analyses

The scores of the participants on the two test batteries were correlated. Because the participants had been selected on the tests that were the subject of these analyses, they constituted a range-restricted sample. The correlations among the tests were corrected for range restriction using Lawley's (1943) method.

The corrected correlation matrix was subjected to an Eigenvalue analysis and further correlations. The AFOQT Verbal composite and the SAT verbal test were correlated as were the AFOQT Quantitative composite and SAT Quantitative test. The correlations within each battery were estimated. Finally, the correlations were corrected for attenuation due to unreliability. This provides the best estimate of the construct similarity.

RESULTS AND DISCUSSION

Table 1 provides the means, standard deviations, and correlations for the two batteries. These data are presented in both uncorrected and corrected-for-range-restriction form. In range-restricted form, all the means were above applicant values and the standard deviations were reduced. The two SAT means are elevated about one standard deviation and the two AFOQT

means about a half a standard deviation, reflecting their respective normative groups. The four standard deviations have been reduced to about 75% of the normative values. In both the uncorrected and the corrected-for-range-restriction correlation matrices, the values are all positive. As would be expected in a selection setting, the corrected correlations are higher than the uncorrected correlations.

Table 1. Means, Standard Deviations and Correlations of the Scores

	SAT-V	SAT-Q	AFOQT-V	AFOQT-Q	Mean	SD
	1.000	.700	.845	.641	425.000	110.000
	.538	1.000	.616	.842	475.000	120.000
	.761	.441	1.000	.647	38.792	27.367
	.470	.752	.508	1.000	41.366	26.185
Mean	531.220	05.460	61.690	65.880		
SD	82.440	89.250	22.540	21.390		

Note. Entries below the diagonal are observed and those above have been corrected for range restriction.

Because the corrected correlations are the superior statistical estimates we will limit our discussion to them. An Eigenvalue analysis of the corrected matrix of the four scores disclosed an unrotated first factor that accounted for 79% of the total variance. Each of the four scores loaded about .89 on this factor. A second unrotated factor accounted for 14% of the variance and weighted the two verbal scores negatively and the two quantitative scores positively. The third unrotated factor at 5%, weighted the SAT negatively and the AFOQT positively. The fourth factor was not interpretable. Table 2 presents the result of the Eigenvalue analysis.

Table 2. Eigenvalue Analysis of the Scores

Eigenvalues	% Accounted For
3.146	78.663
0.543	13.584
0.200	5.004
0.109	2.747

Loadings on the Unrotated Factors

Factor	1	2	3	4
SAT-V	.898	-.332	-.222	-.179
SAT-Q	.890	.358	-.220	.171
AFOQT-V	.875	-.403	.211	.160
AFOQT-Q	.881	.377	.239	-.149

The correlation of the two verbal scores was very high at .85. The same was true for the quantitative scores with a correlation of .84. After correction for attenuation, the correlations were .93 for verbal and .94 for quantitative suggesting a near identity of the constructs measured. Within-battery correlations were not quite as strong but still high. The correlation of the two SAT tests was .70 and the correlation of the two AFOQT composites was .65. Corrected for unreliability, these two values became .76 and .74--quite close.

These two test batteries were written by different groups of individuals, using different rules of item construction, and different content taxonomies. Despite this, the results suggest that the verbal sections measure the same construct, and the quantitative sections measure the same constructs. Further, both batteries are highly saturated with a first factor and each of the four scores loads about the same on this first factor. All these analyses point to a great similarity between the SAT and the verbal and quantitative sections of the AFOQT.

The results of this experiment are promising, however, additional analyses are required. This effort was limited to the AFOQT Verbal and Quantitative composites. The US Air Force uses Pilot and Navigator-Technical composites for classification into aircrew training specialties. Olea and Ree (1994) have demonstrated that job knowledge tests from the AFOQT such as Aviation Information and Instrument Comprehension add substantially to the prediction of pilot training success. The SAT includes no such content and, therefore, cannot be expected to be as valid for pilot training. Additionally, validation of the SAT for technical training courses should be undertaken.

Pending additional studies, it appears that the potential for interchangeability is high. Use of standardized tests could save millions of dollars in test construction and administration costs.

References

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

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

Carretta, T. R., & Ree, M. J. (1997). *The best retest score is the average: Findings and implications* (AL/HR-TP-1996-0021). Brooks AFB, TX: Manpower and Personnel Research Division, Armstrong Laboratory, Human Resources Directorate.

Donolon, T., & Livingston, S. (Ed.) (1984). Psychometric methods used in the admissions testing program in T. Donolon (Ed.). *College board handbook for the Scholastic Aptitude Test and Achievement tests*. NY: College Entrance Examination Board.

Gulliksen, H. (1950). *Theory of mental tests*. NY: Wiley.

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

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

Ree, M. J., Carretta, T. R., Earles, J. A., & Albert, W. (1994). Sign changes when correcting for range restriction: A note on Pearson's and Lawley's selection formulas. *Journal of Applied Psychology, 79*, 298-301.

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