



AD-A271 042



Item Calibration: Medium-of-Administration Effect on Computerized Adaptive Scores

**Rebecca D. Hetter
Bruce M. Bloxom
Daniel O. Segall**



93-25378



93 10 20 12 9

Item Calibration: Medium-of-Administration Effect on Computerized Adaptive Scores

Rebecca D. Hetter

Bruce M. Bloxom
Defense Manpower Data Center

Daniel O. Segall

Reviewed and approved by
W. A. Sands

Released by
John D. McAfee
Captain, U.S.Navy
Commanding Officer
and
Richard C. Sorenson
Technical Director (Acting)

Approved for public release;
distribution is unlimited.

Navy Personnel Research and Development Center
San Diego, California 92152-7250

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE September 1993		3. REPORT TYPE AND DATE COVERED Final—March 1987-December 1989	
4. TITLE AND SUBTITLE Item Calibration: Medium-of-Administration Effect on Computerized Adaptive Scores				5. FUNDING NUMBERS Program Element: 0604703N Work Units: R1822-MH001 R1822-MH001A Program Element: O&M,N Reimbursable	
6. AUTHOR(S) Rebecca D. Hetter, Bruce M. Bloxom, Daniel O. Segall					
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Navy Personnel Research and Development Center San Diego, California 92152-7250				8. PERFORMING ORGANIZATION REPORT NUMBER NPRDC-TR-93-9	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Bureau of Naval Personnel (PERS-23) Navy Department Washington, DC 20370-5000				10. SPONSORING/MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES Functional Area: Personnel Systems Product Line: Computerized Testing Effort: Computer Adaptive Testing (CAT)					
12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.				12b. DISTRIBUTION CODE A	
13. ABSTRACT (Maximum 200 words) An important question in the development of item pools for computerized adaptive tests (CATs) is whether data for calibrating items should be collected by a paper-and-pencil (P&P) or a computer administration of the items. This study evaluated the effect on adaptive scores of using a P&P calibration. The correspondence between adaptive scores obtained with computer-administered items and a P&P calibration with adaptive scores obtained with computer-administered items and a computer calibration was evaluated. Forty items from each of four Armed Services Vocational Aptitude Battery (ASVAB) content areas (general science, arithmetic reasoning, work knowledge, and shop information) were administered by computer to one group of Navy recruits and by P&P to a second group. These data were used to obtain computer-based and P&P-based calibrations of the items. Each calibration was then used to estimate item response theory adaptive scores for a third group of recruits who received the items by computer. The effect of medium of administration was assessed by comparative regression, correlation, and reliability analyses of the scores using the alternative calibrations. Results indicate that, although statistically significant medium effects were found on some content areas, medium of administration did not affect the reliability of the adaptive scores. Although these findings support the use of the P&P parameters of the current CAT-ASVAB item pool, it is recommended that further analyses be performed to elucidate the significant effects.					
14. SUBJECT TERMS Item calibration, computerized adaptive testing, computerized ability tests, computer-administered tests, calibration mode, ability tests				15. NUMBER OF PAGES 36	
				16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED	18. SECURITY CLASSIFICATION OF THIS PAGE UNCLASSIFIED	19. SECURITY CLASSIFICATION OF ABSTRACT UNCLASSIFIED	20. LIMITATION OF ABSTRACT UNLIMITED		

Foreword

A joint-service effort is in progress to develop a computerized adaptive testing (CAT) system and to evaluate its potential for use in the military entrance processing stations as a replacement for the paper-and-pencil Armed Services Vocational Aptitude Battery (ASVAB). The Department of the Navy has been designated as lead service for CAT system development and the Navy Personnel Research and Development Center has been designated as lead laboratory.

This research was funded under CAT-ASVAB Program Element 0604703N, Work Units R1822-MH001 and R1822-MH001A, and reimbursable Navy funding (O&M,N), sponsored by the Bureau of Naval Personnel (PERS-23).

This research was part of the overall evaluation of CAT-ASVAB. This report presents the results of an evaluation of the effect on adaptive scores of item calibration medium of administration. The data were collected by RGI, Inc., pursuant to contract N66001-86-C-0217. Results are directed toward technical, professional, and contractor personnel involved in implementing CAT.

JOHN D. McAFEE
Captain, U.S. Navy
Commanding Officer

RICHARD C. SORENSON
Technical Director (Acting)

Accession For	
NTIS GRA&I	☒
DTIC TAB	☐
Unannounced	☐
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

Summary

Problem

The Navy Personnel Research and Development Center is conducting research to design and evaluate a computerized adaptive test (CAT) as a potential replacement for the paper-and-pencil (P&P) Armed Services Vocational Aptitude Battery (ASVAB). In support of this effort, the Accelerated CAT-ASVAB Program (ACAP) is evaluating item pools specifically developed for computerized adaptive testing.

An important question in the development of item pools for computerized adaptive tests is whether data for calibrating items should be collected by a P&P or a computer administration of the items. If P&P administrations do not yield precise enough calibrations, items must be administered by computer for calibration just as they are during testing. Since the CAT-ASVAB item pools have been calibrated using P&P administrations, this is an issue of interest for ACAP research.

Objective

The objective of this study was to evaluate the effect on adaptive scores of using a P&P calibration. Specifically, to what extent do adaptive scores obtained with computer-administered items and a P&P calibration correspond to adaptive scores obtained with computer-administered items and a computer calibration?

Method

Forty items from each of four ASVAB content areas—general science (GS), arithmetic reasoning (AR), word knowledge (WK), and shop information (SI)—were administered by computer to one group of Navy recruits and by P&P to a second group. These data were used to obtain computer-based and P&P-based calibrations of the items. Each calibration was then used to estimate item response theory adaptive scores for a third group of recruits who had received the items by computer. The effect of medium of administration was assessed by comparative analyses of the scores using the alternative calibrations.

Testing was conducted at the Recruit Training Center in San Diego, CA. ASVAB scores of record were obtained for nearly all of the recruits and were used to assess whether the groups were comparable in ability levels.

Results and Discussion

Results of the reliability analyses indicate that random errors due to calibration have equivalent variance across different media. These results suggest that the use of item parameters obtained in a P&P calibration will not affect the reliability of CAT-ASVAB test scores, an important concern of the ACAP program.

Results of the regression and correlation analyses show statistically significant medium-of-administration effects. The regression results showed effects on AR, WK, and SI; and the correlation results showed effects on GS and WK.

Conclusions

Results of regression, correlation, and reliability analyses conducted to evaluate calibration medium-of-administration effect on adaptive scores indicate that, although statistically significant medium effects were found on some content areas, these effects did not affect the reliability of the CAT-ASVAB scores.

Recommendations

Although these findings support the use of the P&P parameters of the current CAT-ASVAB item pool, further hypothesis testing with an expanded reliability model is recommended to elucidate the significant effects. In addition, analyses of individual item parameters may be necessary for understanding these effects.

Contents

	Page
Introduction	1
Background	1
Objective	1
Method	1
Subjects	1
Items	2
Item Calibrations	2
Scores	3
Adaptive Scores	3
Nonadaptive Scores	3
ASVAB Scores	4
Results and Discussion	4
Calibration Samples	4
AFQT Comparisons	4
Number-Right Scores	5
Regression Analysis	7
Correlation Analysis	8
Reliability Analysis	8
Statistical Model	9
LISREL Models	11
Conclusions	12
Recommendation	12
References	13
Appendix—Regressions, Correlations, and Scatter Plots	A-0
Distribution List	

List of Tables

1.	Medium-of-Administration Subtest Order and Time Limits	2
2.	Calibration Design	3
3.	Computation of Theta Scores.....	3
4.	Analysis of Variance: AFQT by Calibration Group.....	4
5.	ASVAB vs. Group 1 (Computer): Number-Right Score Means, Standard Deviations, and Intercorrelations	5
6.	ASVAB vs. Group 2 (P&P): Number-Right Score Means, Standard Deviations, and Intercorrelations	6
7.	ASVAB vs. Group 3 (Computer): Number-Right Score Means, Standard Deviations, and Intercorrelations	7
8.	Test for Equality of Covariances: $COV(T1, T3) = (COV(T2, T3))$	8
9.	<i>t</i> -Test of the Difference Between Dependent Correlations	9
10.	Correlation and Variance Matrices in the Model	10
11.	Test for Equality of Reliabilities	12

Introduction

Background

The Navy Personnel Research and Development Center is conducting research to design and evaluate a computerized adaptive test (CAT) as a potential replacement for the paper-and-pencil (P&P) Armed Services Vocational Aptitude Battery (ASVAB). In support of this effort, the Accelerated CAT-ASVAB Program (ACAP) is evaluating item pools specifically developed for computerized adaptive testing.

An important question in the development of item pools for computerized adaptive tests is whether data for calibrating items should be collected by a P&P or a computer administration of the items. Although research shows that computerized adaptive tests with P&P item calibrations can have validities comparable to conventional P&P tests (Moreno, Segall, & Kieckhafer, 1985, pp. 29-33), how much less than optimal these computerized adaptive tests might be is unknown.

The concern about medium of administration in item calibration is that item parameters for some types of items (e.g., items with long paragraphs or with graphics) may differ between computer and P&P administrations. This could result in less-than-optimal item selection and score estimation in adaptive tests. If P&P administrations do not yield precise enough calibrations, items must be administered by computer during calibration just as they are during testing.

Objective

The objective of this study was to evaluate the effect on adaptive scores of using a P&P calibration. Specifically, to what extent do adaptive scores obtained with computer-administered items and a P&P calibration correspond to adaptive scores obtained with computer-administered items and a computer calibration?

Method

Fixed blocks of items were administered by computer to one group of examinees and by P&P to a second group. These data were used to obtain computer-based and P&P-based calibrations of the items. Each calibration was then used to estimate item response theory (IRT) adaptive scores (thetas) for a third group of examinees who had received the items by computer. The effect of medium of administration was assessed by comparative analyses of the thetas using the alternative calibrations.

Subjects

The subjects were Navy recruits who were randomly assigned to one of three groups. Data were collected for 2,955 examinees with 989 in Group 1 (computer), 978 in Group 2 (P&P), and 988 in Group 3 (computer). These sample sizes provide enough data for independent calibrations, since simulation results obtained by Hulin, Drasgow, and Jarsons (1983, pp.101-110) suggest that substantially larger samples produce little improvement in the precision of item characteristic curves and scores, given the number of items (40) used in these calibrations.

Testing was conducted at a Recruit Training Center in San Diego, CA. ASVAB scores of record were obtained for nearly all of the recruits and were used to assess whether the groups were comparable in ability levels.

Items

The items were taken from pools specifically developed in support of CAT-ASVAB by Prestwood, Vale, Massey, and Welsh (1985). Forty items from each of four ASVAB content areas (general science, arithmetic reasoning, word knowledge, and shop information) were administered by computer to Groups 1 and 3, and by P&P to Group 2. The items were conventionally administered in ascending order of difficulty, using the difficulties obtained by Prestwood et al. (1985). The three groups received the same items with the same instructions and practice problems, in the same order and with the same time limits. Although only 4 of the 11 CAT-ASVAB content areas were included in this study, the medium-of-administration (MOA) subtests were administered in the same order as in the CAT-ASVAB. Time limits were prorated from 95% completion times for the same content areas in ACAP, with 10% added to allow for a higher completion rate. Subtest order and time limits are shown in Table 1.

Table 1
Medium-of-Administration Subtest
Order and Time Limits

Subtest	Time (Minutes)
General Science (GS)	19
Arithmetic Reasoning (AR)	63
Word Knowledge (WK)	16
Shop Information (SI)	17
Total	115

The 40 items included 34 high-usage items (usage obtained from ACAP simulation studies) and six "seeds" (not-scored items administered for the purpose of gathering data for on-line calibration research). The booklet format was the same as that used in the original P&P calibration by Prestwood et al. (1985), and the computer format was the same as that used in ACAP. Practice problems and instructions were also as in ACAP.

Item Calibrations

IRT parameter estimates based on the three-parameter logistic model (Birnbaum, 1968) were obtained in separate calibrations for each of the two computer groups (calibrations C1 and C3) and for the P&P group (calibration C2). The data sets on which the calibrations are based are labelled U1, U3, and U2, correspondingly. The calibrations were performed with LOGIST6 (Wingersky, Barton, & Lord, 1982), a computer program that uses a joint maximum-likelihood approach. The design with the corresponding notation is summarized in Table 2.

Table 2**Calibration Design**

Group No.	Medium	Data Set/ Item Responses	Item Parameters/ Calibrations
1	Computer	U1	C1
2	P&P	U2	C2
3	Computer	U3	C3

Note. P&P = paper and pencil.

Scores

For each recruit in Group 3, three theta scores were computed: T1, T2, and T3 (see Table 3). All three scores were based on U3 responses. T1 scores were calculated from the computer-administration item parameters (C1). T2 scores were based on the P&P-administration parameters (C2), and T3 scores were calculated from the third parameter set (C3), also based on a computer-administration. As described below, T1 and T2 were adaptive scores, using only 10 of the 40 responses from a given examinee, while T3 theta was nonadaptive, using all 40 responses.

Table 3**Computation of Theta Scores**

Calibration Parameters	Response Set	Scoring Method	Test Length	Theta
C1 (Group 1, computer)	U3	Adaptive	10 items	T1
C2 (Group 2, P&P)	U3	Adaptive	10 items	T2
C3 (Group 3, computer)	U3	Nonadaptive	40 items	T3

Note. P&P = paper and pencil.

Adaptive Scores

To compute the adaptive thetas (T1 and T2), 10-item adaptive tests were simulated using actual examinee responses. Owen's Bayesian scoring (Owen, 1975) was used throughout the test to update the ability estimate, and a Bayesian modal estimate was computed at the end of the test to obtain the final score. Items were selected from information tables on the basis of maximum information. (An information table consists of lists of items by ability level. Within each list, all the items in the pool—40 in this case—are arranged in descending order of the values of their information functions computed at that ability level. The information tables used in this study were computed for 37 ability levels equally spaced along the $[-2.25, +2.25]$ interval).

Nonadaptive Scores

The nonadaptive thetas (T3) included all 40 items in the test. Final thetas were computed using the Bayesian modal estimate (since all the items go into the score, it is not necessary to update the ability estimate after each item).

Table 3 summarizes the method used for computing the theta scores used in the analyses.

ASVAB Scores

ASVAB subtest scores for the four content areas of interest and the Armed Forces Qualification Test (AFQT) were obtained from the records for most of the examinees. The subtests were General Science (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), and Auto Shop (AS). Notice that the ASVAB's Auto Shop subtest covers two content areas: auto information and shop information, whereas in the CAT-ASVAB each area constitutes a separate subtest. Since only shop information was administered in this study, ASVAB-AS was compared to MOA-SI.

Results and Discussion

Calibration Samples

Two subjects in Group 3 (computer) had fewer than 10 valid responses on subtests WK and SI, and LOGIST omitted them from the calibrations. These subjects were eliminated from all subsequent analyses of WK and SI, Group 3 (computer). Final sample sizes were 989 for Group 1 (computer), 978 for the Group 2 (P&P), 988 for GS & AR in Group 3 (computer), and 986 for WK & SI in Group 3 (computer).

AFQT Comparisons

To determine whether the three calibration groups were comparable in examinee ability, a one-way analysis of variance of AFQT by calibration group was computed. Results (Table 4) clearly indicate that there are no AFQT differences among the three groups. Sample sizes are slightly smaller in Tables 4 through 7 because AFQT scores were not available for some examinees.

Table 4

Analysis of Variance: AFQT by Calibration Group

Source	df	Sum of Squares	Mean Squares	F-Ratio	F-Prob.
Between Groups	2	29.8	14.9199	0.035	0.9656
Within Groups	2923	1247241.6	426.6990		
Total	2925	1247271.4			

Group No.	Count	Mean	Standard Deviation	Standard Error
1	985	55.5655	21.0157	0.6696
2	963	55.3946	20.4010	0.6574
3	978	55.3221	20.5418	0.6569
Total	2926	55.4279	20.6499	0.3818

Number-Right Scores

Tables 5, 6, and 7 show means, standard deviations, and intercorrelations of number-right scores for same-name ASVAB and MOA subtests by calibration group.

Table 5

**ASVAB vs. Group 1 (Computer): Number-Right Score
Means, Standard Deviations, and Intercorrelations
(N = 985)**

Subtest	P&P ASVAB					Group 1 (Computer)			
	GS	AR	WK	AS	AFQT	GS	AR	WK	SI
Correlation Matrix									
P&P ASVAB									
GS	1.00								
AR	0.57	1.00							
WK	0.73	0.52	1.00						
AS	0.50	0.42	0.48	1.00					
AFQT	0.69	0.85	0.78	0.45	1.00				
Group 1 (Computer)									
GS	0.81	0.58	0.73	0.53	0.71	1.00			
AR	0.49	0.73	0.46	0.35	0.71	0.57	1.00		
WK	0.69	0.47	0.77	0.43	0.67	0.72	0.46	1.00	
SI	0.54	0.44	0.49	0.78	0.46	0.60	0.41	0.48	1.00
Means and Standard Deviations									
Min	4.00	7.00	.00	0.00	21.00	8.00	6.00	6.00	6.00
Max	25.00	30.00	35.00	25.00	99.00	39.00	40.00	39.00	39.00
Mean	17.55	20.22	26.85	16.14	55.57	25.66	20.06	22.82	22.62
SD	4.44	5.80	5.41	5.16	21.01	6.23	5.96	5.13	6.88

Note. ASVAB = Armed Services Vocational Aptitude Battery, P&P = paper and pencil, GS = General Science, AR = Arithmetic Reasoning, WK = Word Knowledge, AS = Auto Shop, AFQT = Armed Forces Qualification Test, SI = Shop Information.

Table 6

**ASVAB vs. Group 2 (P&P): Number-Right Score
Means, Standard Deviations, and Intercorrelations
(N = 963)**

Subtest	P&P ASVAB					Group 2 (P&P)			
	GS	AR	WK	AS	AFQT	GS	AR	WK	SI
Correlation Matrix									
P&P ASVAB									
GS	1.00								
AR	0.50	1.00							
WK	0.74	0.48	1.00						
AS	0.52	0.39	0.46	1.00					
AFQT	0.69	0.82	0.78	0.42	1.00				
Group 2 (P&P)									
GS	0.80	0.54	0.76	0.51	0.72	1.00			
AR	0.45	0.72	0.44	0.30	0.68	0.54	1.00		
WK	0.69	0.45	0.79	0.38	0.68	0.74	0.44	1.00	
SI	0.53	0.41	0.50	0.77	0.46	0.56	0.36	0.43	1.00
Means and Standard Deviations									
Min	5.00	5.00	10.00	3.00	17.00	8.00	6.00	5.00	3.00
Max	25.00	30.00	35.00	25.00	99.00	39.00	39.00	40.00	40.00
Mean	17.44	20.41	26.68	16.16	55.39	25.33	20.22	22.74	22.96
SD	4.38	5.52	5.32	5.02	20.40	6.18	5.81	5.28	6.84

Note. See Table 5 for definitions of acronyms.

Table 7

**ASVAB vs. Group 3 (Computer): Number-Right Score
Means, Standard Deviations, and Intercorrelations
(N = 978)**

Subtest	P&P ASVAB					Group 3 (Computer)			
	GS	AR	WK	AS	AFQT	GS	AR	WK	SI
Correlation Matrix									
P&P ASVAB									
GS	1.00								
AR	0.48	1.00							
WK	0.74	0.44	1.00						
AS	0.47	0.35	0.45	1.00					
AFQT	0.66	0.83	0.75	0.40	1.00				
Group 3 (Computer)									
GS	0.81	0.51	0.75	0.52	0.69	1.00			
AR	0.46	0.74	0.42	0.31	0.70	0.53	1.00		
WK	0.70	0.40	0.79	0.38	0.66	0.73	0.45	1.00	
SI	0.52	0.38	0.49	0.76	0.45	0.60	0.36	0.48	1.00
Means and Standard Deviations									
Min	4.00	5.00	6.00	0.00	20.00	6.00	5.00	5.00	5.00
Max	25.00	30.00	35.00	25.00	99.00	40.00	40.00	40.00	39.00
Mean	17.42	20.14	26.84	16.14	55.32	25.70	19.70	22.66	22.56
SD	4.47	5.67	5.48	4.93	20.54	6.17	5.94	5.22	7.03

Note. See Table 5 for definition of acronyms.

Regression Analysis

This analysis was designed to test whether 10-item adaptive ability scores computed using computer or P&P calibrated items are equivalent. For each of the four content areas, the appendix presents the regressions and scatter plots of T1 on T3, and of T2 on T3, where T1, T2, and T3 are as defined in Table 3. Then, a LISREL-based (Joreskog & Sorbom, 1986) analysis was designed to test for the equality of the regression lines of T1 on T3 and of T2 on T3. The testing was sequential, first for equality of slopes and then for equality of intercepts (if the slopes are different, testing for equality of intercepts is not required).

To test for equality of slopes, a LISREL run that tested for equality of covariances was performed for each of the four content areas. The model specification for LISREL was as follows:

Φ = covariance matrix, free
 Λ_χ = identity matrix
 $\theta_{\hat{\epsilon}}$ = error matrix, zero
 $\text{COV}(T1, T3) = \text{COV}(T2, T3)$

Results are presented in Table 8. Significant differences were obtained for all content areas except GS.

Table 8

**Test for Equality of Covariances:
 $\text{COV}(T1, T3) = \text{COV}(T2, T3)$**

Subtest	Chi Sq	df	p	GoF	Adj GoF	RMSR
GS	.02	1	.876	.999	.995	.001
AR	10.92*	1	.001	.644	-1.137	.011
WK	21.38*	1	.000	.495	-2.028	.014
SI	12.13*	1	.000	.790	-0.261	.016

Notes. 1. p = probability, GoF = Goodness of Fit, RMSR = Root Mean Square Residuals.

2. See Table 5 for definitions of other acronyms.

* $p < .05$.

Correlation Analysis

After the results of the regression analyses were obtained, it was decided to use directional hypotheses in an attempt to explain the differences found. For each of the four content areas, the Pearson correlations, $r(T1 \text{ on } T3)$ and $r(T2 \text{ on } T3)$, were obtained, and t -tests of the difference between dependent correlations (Cohen & Cohen, 1975, p. 53) were computed. Table 9 presents these results.

For subtests GS and WK, when correlations were computed between 10-item adaptive and 40-item nonadaptive thetas, $r(T1, T3)$ (the computer-computer correlation) was significantly higher than $r(T2, T3)$ (the P&P-computer correlation). This is consistent with a hypothesis that thetas based on the same calibration medium of administration are more similar than thetas based on different media of administration. However, without further analysis, it is not clear why these results were obtained for GS and WK and not for AR and SI.

Reliability Analysis

Since the results from the regression and correlation analyses are conflicting and difficult to interpret, further analyses were required. A design was developed to assess the effect of calibration medium on test reliabilities. The model and the LISREL specifications are described below. These tests assess overall effect across the four content areas simultaneously; if a significant effect is found, further analyses would be required to attribute the error to specific subtests.

Table 9

t-Test of the Difference Between Dependent Correlations
H₀: r₁₃ > r₂₃

Subtest	N ^a	r(T1,T2)	r(T1,T3)	r(T2,T3)	df	t
GS	988	0.9703	0.9608	0.9552	985	2.768**
AR	988	0.9813	0.9586	0.9587	985	-0.060
WK	986	0.9798	0.9665	0.9632	983	2.114*
SI	986	0.9564	0.9508	0.9507	983	0.039

Notes. 1. r = Pearson correlation coefficient; T1 = Calibration Group 1 (computer), 10-item adaptive theta; T2 = Calibration Group 2 (P&P), 10-item adaptive theta; T3 = Calibration Group 3 (computer), 40-item nonadaptive theta.

2. See Table 5 for definition of other acronyms.

^aAll responses from Group 3 (computer) (U3).

** $p < .01$ (one tailed).

* $p < .05$ (one tailed).

Statistical Model

Assume that the observed theta, $\hat{\theta}$, values have three components: the true ability level θ , measurement error ϵ , and random error due to calibration δ . Then,

$$\hat{\theta} = \lambda (\theta + \epsilon) + \delta$$

$$\hat{\theta} = \lambda \xi + \delta$$

where $\xi = \theta + \epsilon$, the true ability plus the error of measurement, and λ is a scale factor. Then, the basic measurement model can be described by the following eight equations:

Equation	Subtest Score	Responses	Item Parameters
$\hat{\theta}_1 = \lambda_1 \xi_1 + \delta_1$	T1-GS	U3	Computer
$\hat{\theta}_2 = \lambda_2 \xi_2 + \delta_2$	T1-AR	U3	Computer
$\hat{\theta}_3 = \lambda_3 \xi_3 + \delta_3$	T1-WK	U3	Computer
$\hat{\theta}_4 = \lambda_4 \xi_4 + \delta_4$	T1-SI	U3	Computer
$\hat{\theta}_5 = \lambda_5 \xi_5 + \delta_5$	T2-GS	U3	P&P
$\hat{\theta}_6 = \lambda_6 \xi_6 + \delta_6$	T2-AR	U3	P&P
$\hat{\theta}_7 = \lambda_7 \xi_7 + \delta_7$	T2-WK	U3	P&P
$\hat{\theta}_8 = \lambda_8 \xi_8 + \delta_8$	T2-SI	U3	P&P

Selecting the best-fitting model consists of: (1) estimating the model in which certain parameters are set to be equal, (2) estimating a less constrained model, and (3) assessing the statistical significance of the improvement in fit going from the more constrained model to the less constrained model. If the more constrained model fits the data as well as the less constrained model (i.e., within sampling error limits), then one may conclude that the constraints do not seriously erode the fit of the model.

In this case, one model is specified such that the calibration errors of the pseudo-true test scores are constrained to be equal for the computer-based and the P&P item parameters; another model is specified such that these calibration errors are free to vary between the two media of administration. If the constrained model provides just as good a fit as the free model, then constraining calibration errors to be equal across item-parameter sets does not erode the fit of the model to the data, and one may conclude that the calibration errors of the ability scores are equal for computer and P&P item-parameters.

According to the model, the variance-covariance matrix Σ among the observed scores has the form:

$$\Sigma = \Lambda_x \Phi \Lambda'_x + \theta_\delta$$

where Λ_x is a diagonal matrix with standard deviations in the diagonal, θ_δ is a diagonal matrix of variances attributable to calibration error, and Φ is the attenuated correlation matrix among the ability values ξ . Notice that the matrix Φ is attenuated from only one source of error—the source attributable to the calibration; Φ is *not* attenuated with respect to measurement error. The fixed and estimated parameters of this model are displayed in Table 10.

Table 10
Correlation and Variance Matrices in the Model

Subtest Score	Correlation Matrix Φ							
	$\hat{\theta}_1$ (T1-GS)	$\hat{\theta}_2$ (T1-AR)	$\hat{\theta}_3$ (T1-WK)	$\hat{\theta}_4$ (T1-SI)	$\hat{\theta}_5$ (T2-GS)	$\hat{\theta}_6$ (T2-AR)	$\hat{\theta}_7$ (T2-WK)	$\hat{\theta}_8$ (T2-SI)
$\hat{\theta}_1$ (T1-GS)	1.0							
$\hat{\theta}_2$ (T1-AR)	$r(\hat{\theta}_2, \hat{\theta}_1)$	1.0						
$\hat{\theta}_3$ (T1-WK)	$r(\hat{\theta}_3, \hat{\theta}_1)$	$r(\hat{\theta}_3, \hat{\theta}_2)$	1.0					
$\hat{\theta}_4$ (T1-SI)	$r(\hat{\theta}_4, \hat{\theta}_1)$	$r(\hat{\theta}_4, \hat{\theta}_2)$	$r(\hat{\theta}_4, \hat{\theta}_3)$	1.0				
$\hat{\theta}_5$ (T2-GS)	1.0	$r(\hat{\theta}_2, \hat{\theta}_1)$	$r(\hat{\theta}_3, \hat{\theta}_1)$	$r(\hat{\theta}_4, \hat{\theta}_1)$	1.0			
$\hat{\theta}_6$ (T2-AR)	$r(\hat{\theta}_2, \hat{\theta}_1)$	1.0	$r(\hat{\theta}_3, \hat{\theta}_2)$	$r(\hat{\theta}_4, \hat{\theta}_2)$	$r(\hat{\theta}_2, \hat{\theta}_1)$	1.0		
$\hat{\theta}_7$ (T2-WK)	$r(\hat{\theta}_3, \hat{\theta}_1)$	$r(\hat{\theta}_3, \hat{\theta}_2)$	1.0	$r(\hat{\theta}_4, \hat{\theta}_3)$	$r(\hat{\theta}_3, \hat{\theta}_1)$	$r(\hat{\theta}_3, \hat{\theta}_2)$	1.0	
$\hat{\theta}_8$ (T2-SI)	$r(\hat{\theta}_4, \hat{\theta}_1)$	$r(\hat{\theta}_4, \hat{\theta}_2)$	$r(\hat{\theta}_4, \hat{\theta}_3)$	1.0	$r(\hat{\theta}_4, \hat{\theta}_1)$	$r(\hat{\theta}_4, \hat{\theta}_2)$	$r(\hat{\theta}_4, \hat{\theta}_3)$	1.0
Variance Matrix $\hat{\theta}_\delta$								
	$r(\hat{\theta}_1, \hat{\theta}_1)$	$r(\hat{\theta}_2, \hat{\theta}_2)$	$r(\hat{\theta}_3, \hat{\theta}_3)$	$r(\hat{\theta}_4, \hat{\theta}_4)$	$r(\hat{\theta}_5, \hat{\theta}_5)$	$r(\hat{\theta}_6, \hat{\theta}_6)$	$r(\hat{\theta}_7, \hat{\theta}_7)$	$r(\hat{\theta}_8, \hat{\theta}_8)$

Notice in Table 10 that the correlation of a subtest with itself (across media) is equal to one, and that the correlations between same-name subtests are assumed to be equal both across and within calibration media. For example, the correlation between P&P scores T2-AR and T2-GS should be represented by $r(\hat{\theta}_6, \hat{\theta}_5)$; however, it is represented by $r(\hat{\theta}_2, \hat{\theta}_1)$ because under the model all the correlations between GS and AR are assumed to be equal; that is,

$$r(\hat{\theta}_2, \hat{\theta}_1) = r(\hat{\theta}_6, \hat{\theta}_1) = r(\hat{\theta}_5, \hat{\theta}_2) = r(\hat{\theta}_6, \hat{\theta}_5) .$$

LISREL Models

To test the model fit, two models were specified and corresponding LISREL runs were performed. In Model 1, the variances of errors due to calibration (θ_8) were free to vary between the two media of administration. In Model 2, the variances of errors due to calibration were constrained to be equal for same-name subtests.

The Φ constraints in Table 10 were imposed for both Model 1 and Model 2.

The LISREL output yields a chi-square statistic that is a measure of how much Σ differs from S ; that is, how well the model fits the data. The difference in the chi-squares from the two models is also a chi-square with df equal to the difference in df from the two models. If this difference is not significant, then the data satisfy/fit the model independently of the calibration errors; that is, errors due to calibration across media for same-name subtests are equal.

The LISREL specifications for Model 1 were:

1. Lambda-X = Diagonal Matrix, Free.
2. PHI = Symmetrical Matrix, Free.
3. Theta-Delta = Diagonal Matrix, Free.

The LISREL specifications for Model 2 were:

1. Lambda-X = Diagonal Matrix, Free.
2. PHI = Symmetrical Matrix, Free.
3. Theta-Delta (TD) = Diagonal Matrix with Constraints.
4. TD Constraint No. 1: All off-diagonal θ_8 fixed at zero.
5. TD Constraint No. 2: θ_8 (computer) = θ_8 (P&P); that is,

$$\begin{aligned} \theta_8(1,1) &= \theta_8(5,5) \\ \theta_8(2,2) &= \theta_8(6,6) \\ \theta_8(3,3) &= \theta_8(7,7) \\ \theta_8(4,4) &= \theta_8(8,8). \end{aligned}$$

Table 11 presents goodness-of-fit statistics for these models. The likelihood ratio chi-square value of the model in Model 1 was 14.07 with 14 df . The result is not statistically significant, indicating that Model 1 adequately explains the observed covariance matrices. Results for Model 2 show a chi-square value of 19.57 with 18 df , which is also not statistically significant.

The difference in chi-squares between Model 1 and Model 2 is distributed as a chi-square with df equal to the difference in df from Model 1 and Model 2. This value (5.50 with 4 df) was not significant, indicating that allowing the error term to be free does not change the fit of the model.

Table 11

Test for Equality of Reliabilities

Model		Chi Sq	df	p	GoF	Adj Gof	RMSR
No.	Specification						
1	$\theta_{\delta} = \text{free}$	14.07	14	n.s.	0.996	0.991	0.002
2	$\theta_{\delta} (\text{Computer}) = \theta_{\delta} (\text{P\&P})$	19.57	18	n.s.	0.995	0.990	0.003
3	Model 2 - Model 1	5.50	4	n.s.			

Note. p = probability, GoF = Goodness of Fit., RMSR = Root Mean Square Residuals, n.s. = not significant.

Conclusions

Results of the regression, correlation, and reliability analyses indicate that, although statistically significant medium effects were found for some content areas, these effects did not affect the reliability of adaptive scores.

Results of the reliability analyses indicate that random errors due to calibration have equivalent variance across different media. This suggests that the use of item parameters obtained in a P&P calibration will not affect the reliability of CAT-ASVAB test scores, an important concern of the ACAP program.

The regression and correlation results showed significant medium-of-administration effects. The regression results showed effects on AR, WK, and SI, and the correlation results showed effects on GS and WK. Because these results can be attributed to the effects of calibration medium on the scale factor λ in the reliability analysis, further hypothesis testing with the reliability model is necessary to elucidate the scale effects. Analyses of individual item parameters may also be necessary for understanding these effects. In addition, the results may be clarified by alternative treatment of not-reached items when the thetas (T_1 , T_2 , and T_3), are being computed.

Recommendation

Although these findings support the use of the P&P parameters of the current CAT-ASVAB item pool, further hypothesis testing with an expanded reliability model is recommended to elucidate the significant effects. In addition, analyses of individual item parameters may be necessary for understanding these effects.

References

- Birnbaum, A. (1968). Some latent-trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Applications to psychological measurement*. Homewood, IL: Dow Jones-Irwin.
- Joreskog, K. G., & Sorbom, D. (1986). *LISREL VI: Analysis of linear structural relationships by maximum likelihood and least square methods*. Mooresville, IN: Scientific Software, Inc.
- Moreno, K. E., Segall, D. O., & Kieckhafer, W. F. (1985). A validity study of the computerized adaptive testing version of the Armed Services Vocational Aptitude Battery. *Proceedings of the 27th Annual Conference of the Military Testing Association*. San Diego, CA: Navy Personnel Research and Development Center.
- Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive testing. *Journal of the American Statistical Association*, 70, 351-356.
- Prestwood, J. S., Vale, C. D., Massey, R. H., & Welsh, J. R. (1985). *Armed Services Vocational Aptitude Battery: Development of an adaptive item pool* (AFHRL-TR-85-19). Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.
- Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide*. Princeton, NJ: Educational Testing Service.

Appendix
Regressions, Correlations, and Scatter Plots

GENERAL SCIENCE

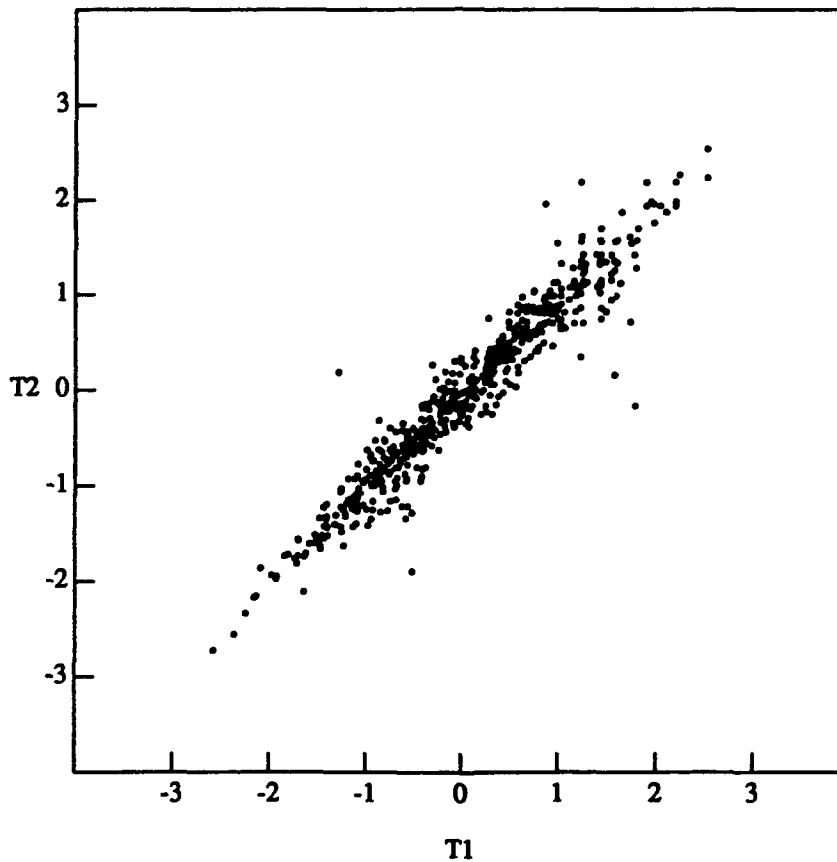
T1 = ADAPTIVE THETA (C1, U3), 10-items
T2 = ADAPTIVE THETA (C2, U3), 10-items
T3 = NON-ADAPTIVE THETA (C3, U3), 40-items

Analysis for 988 points of 2 variables:

Variable	T2	T1
Min	-2.5830	-2.6970
Max	2.5110	2.5660
Sum	66.6930	23.7240
Mean	0.0675	0.0240
SD	0.8631	0.8570

Correlation Matrix:

T2	1.0000	
T1	0.9703	1.0000
Variable	T2	T1



GENERAL SCIENCE

T1 = ADAPTIVE THETA (C1, U3), 10-items
T2 = ADAPTIVE THETA (C2, U3), 10-items
T3 = NON-ADAPTIVE THETA (C3, U3), 40-items

Analysis for 988 points of 2 variables:

Variable	T1	T3
Min	-2.6970	-3.0300
Max	2.5660	2.5710
Sum	23.7240	2.4860
Mean	0.0240	0.0025
SD	0.8570	0.9232

Correlation Matrix:

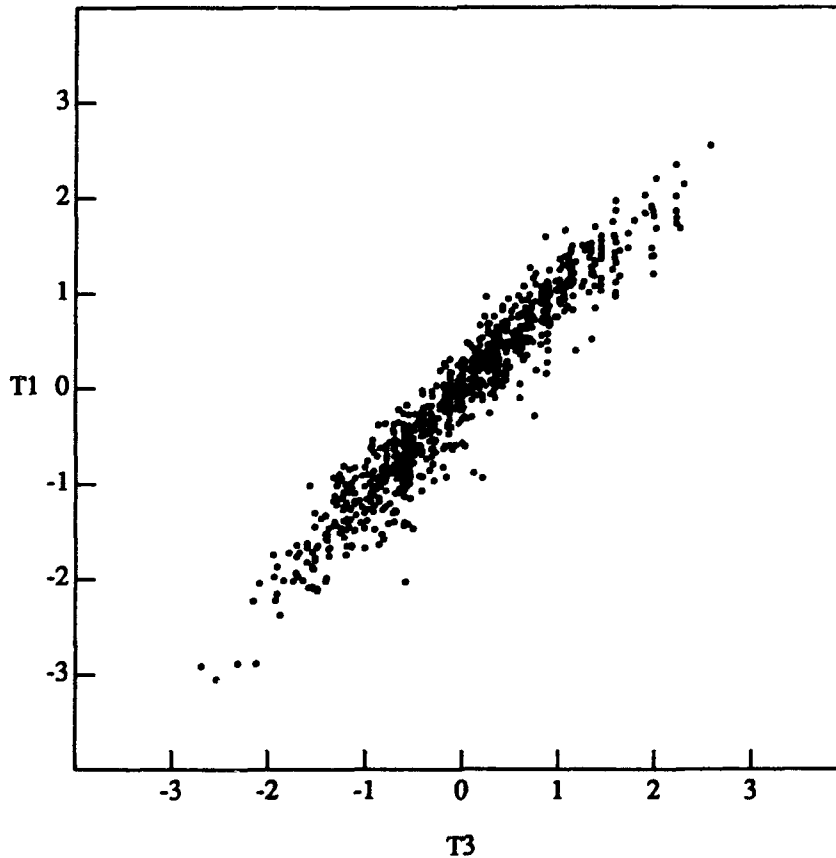
T1	1.0000	
T3	0.9608	1.0000
Variable	T1	T3

Regression Equation for T1:

$$T1 = 0.8919 T3 + 0.0217679$$

Significance test for prediction of T1

Multi-R	R-Squared	F(1,986)	prob (F)
0.9608	0.9232	11856.9617	0.0000



GENERAL SCIENCE

T1 = ADAPTIVE THETA (C1, U3), 10-items
T2 = ADAPTIVE THETA (C2, U3), 10-items
T3 = NON-ADAPTIVE THETA (C3, U3), 40-items

Analysis for 988 points of 2 variables:

Variable	T2	T3
Min	-2.5830	-3.0300
Max	2.5110	2.5710
Sum	66.6930	2.4860
Mean	0.0675	0.0025
SD	0.8631	0.9232

Correlation Matrix:

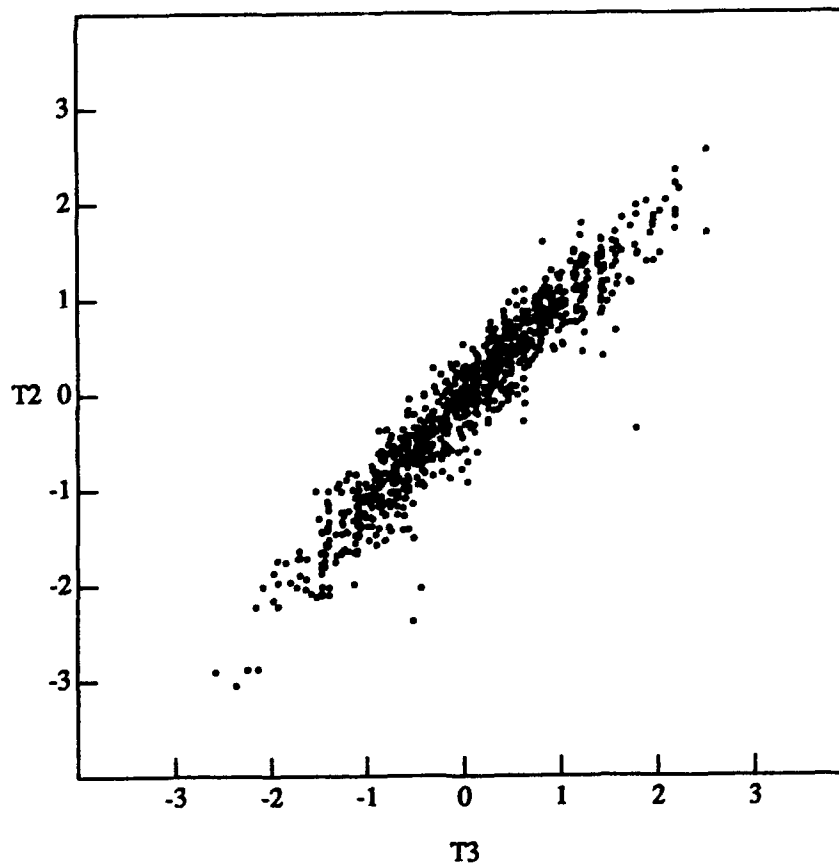
T2	1.0000	
T3	0.9552	1.0000
Variable	T2	T3

Regression Equation for T2:

$$T2 = 0.8931 T3 + 0.0652559$$

Significance test for prediction of T2

Mult-R	R-Squared	F(1,986)	prob (F)
0.9552	0.9124	10271.9555	0.0000



ARITHMETIC REASONING

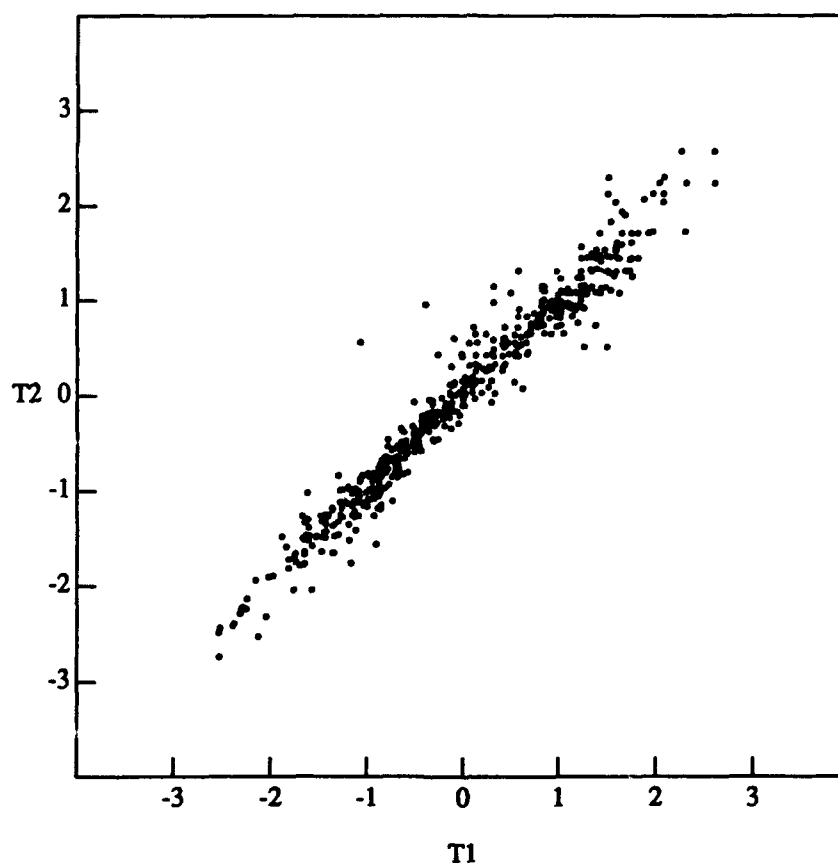
T1 = ADAPTIVE THETA (C1, U3), 10-items
T2 = ADAPTIVE THETA (C2, U3), 10-items
T3 = NON-ADAPTIVE THETA (C3, U3), 40-items

Analysis for 988 points of 2 variables:

Variable	T2	T1
Min	-2.5440	-2.7190
Max	2.5850	2.5890
Sum	-66.1150	-25.5640
Mean	-0.0669	-0.0259
SD	0.9463	0.9260

Correlation Matrix:

T2	1.0000	
T1	0.9813	1.0000
Variable	T2	T1



ARITHMETIC REASONING

T1 = ADAPTIVE THETA (C1, U3), 10-items
T2 = ADAPTIVE THETA (C2, U3), 10-items
T3 = NON-ADAPTIVE THETA (C3, U3), 40-items

Analysis for 988 points of 2 variables:

Variable	T1	T3
Min	-2.7190	-2.5370
Max	2.5890	2.9080
Sum	-25.5640	4.1980
Mean	-0.0259	0.0042
SD	0.9266	0.9449

Correlation Matrix:

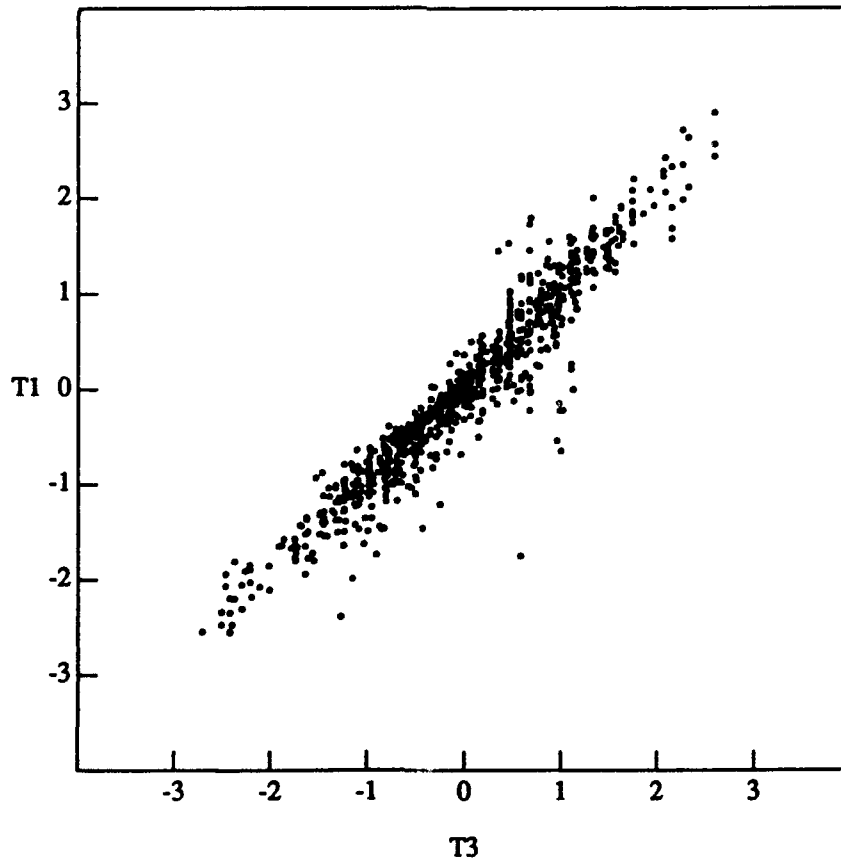
T1	1.0000	
T3	0.9586	1.0000
Variable	T1	T3

Regression Equation for T1:

$$T1 = 0.94 T3 + -0.0298686$$

Significance test for prediction of T1

Multi-R	R-Squared	F(1,986)	prob (F)
0.9586	0.9189	11174.2661	0.0000



ARITHMETIC REASONING

T1 = ADAPTIVE THETA (C1, U3), 10-items
T2 = ADAPTIVE THETA (C2, U3), 10-items
T3 = NON-ADAPTIVE THETA (C3, U3), 40-items

Analysis for 988 points of 2 variables:

Variable	T2	T3
Min	-2.5440	-2.5370
Max	2.5850	2.9080
Sum	-66.1150	4.1980
Mean	-0.0669	0.0042
SD	0.9463	0.9449

Correlation Matrix:

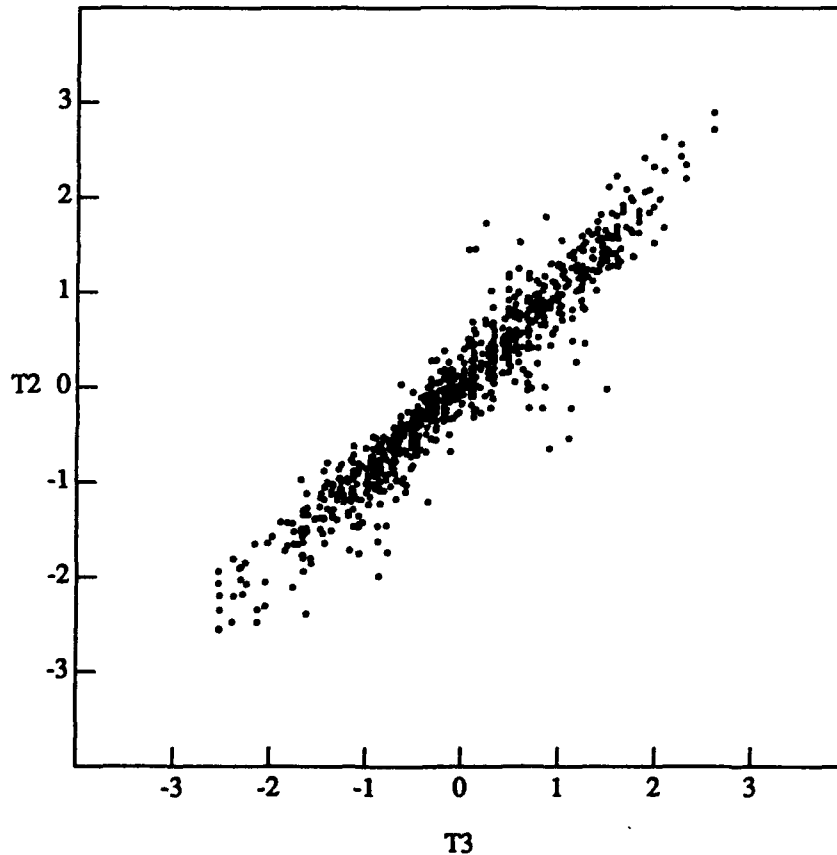
T2	1.0000	
T3	0.9587	1.0000
Variable	T2	T3

Regression Equation for T2:

$$T2 = 0.9602 T3 + -0.070998$$

Significance test for prediction of T2

Multi-R	R-Squared	F(1,986)	prob (F)
0.9587	0.9192	11216.0675	0.0000



WORD KNOWLEDGE

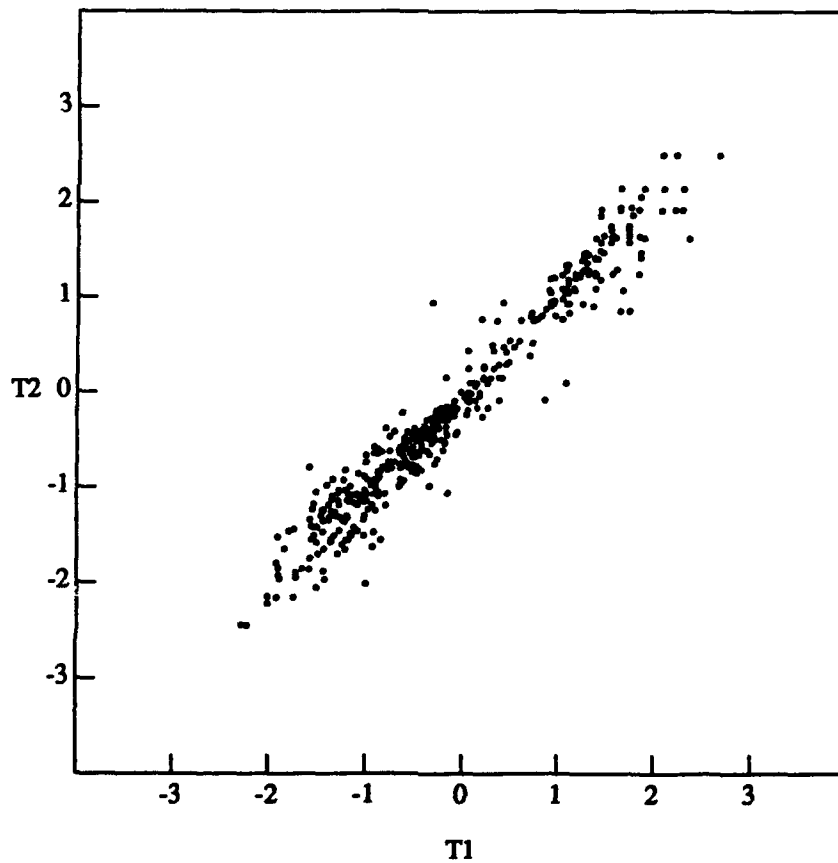
T1 = ADAPTIVE THETA (C1, U3), 10-items
T2 = ADAPTIVE THETA (C2, U3), 10-items
T3 = NON-ADAPTIVE THETA (C3, U3), 40-items

Analysis for 986 points of 2 variables:

Variable	T2	T1
Min	-2.2940	-2.4290
Max	2.6370	2.5080
Sum	33.4100	11.9230
Mean	0.0339	0.0121
SD	0.8531	0.8767

Correlation Matrix:

T2	1.0000	
T1	0.9798	1.0000
Variable	T2	T1



WORD KNOWLEDGE

T1 = ADAPTIVE THETA (C1, U3), 10-items
T2 = ADAPTIVE THETA (C2, U3), 10-items
T3 = NON-ADAPTIVE THETA (C3, U3), 40-items

Analysis for 986 points of 2 variables:

Variable	T1	T3
Min	-2.4290	-2.5000
Max	2.5080	2.9960
Sum	11.9230	5.2260
Mean	0.0121	0.0053
SD	0.8767	0.8920

Correlation Matrix:

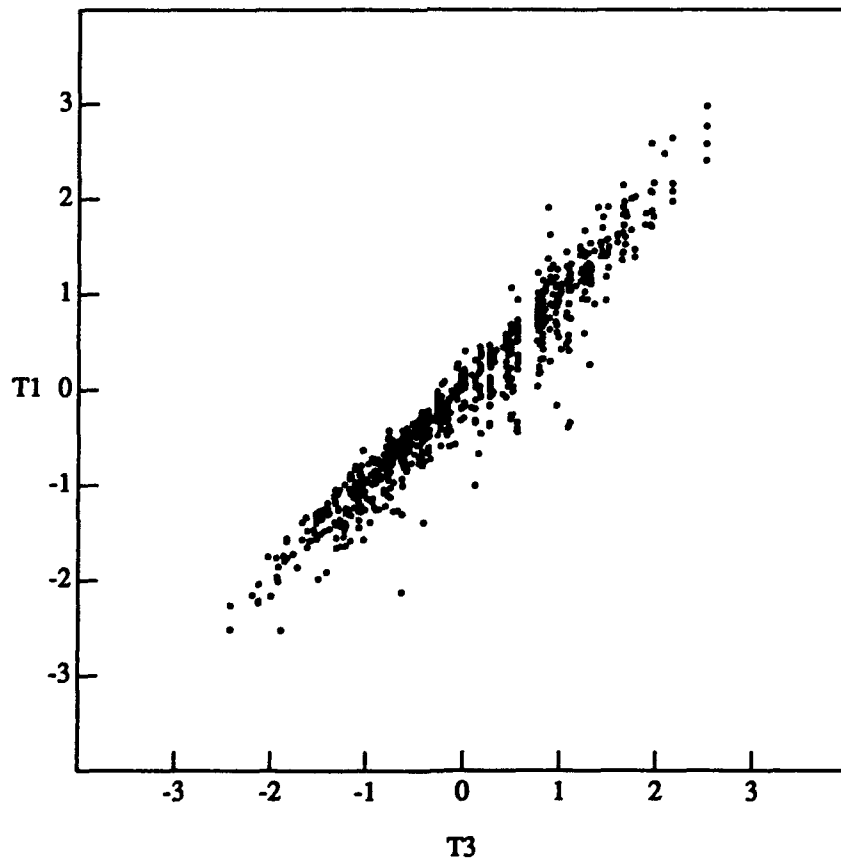
T1	1.0000	
T3	0.9665	1.0000
Variable	T1	T3

Regression Equation for T1:

$$T1 = 0.95 T3 + 0.00705733$$

Significance test for prediction of T1

Multi-R	R-Squared	F(1,984)	prob (F)
0.9665	0.9341	13958.1727	0.0000



WORD KNOWLEDGE

T1 = ADAPTIVE THETA (C1, U3), 10-items
T2 = ADAPTIVE THETA (C2, U3), 10-items
T3 = NON-ADAPTIVE THETA (C3, U3), 40-items

Analysis for 986 points of 2 variables:

Variable	T2	T3
Min	-2.2940	-2.5000
Max	2.6370	2.9960
Sum	33.4160	5.2260
Mean	0.0339	0.0053
SD	0.8531	0.8920

Correlation Matrix:

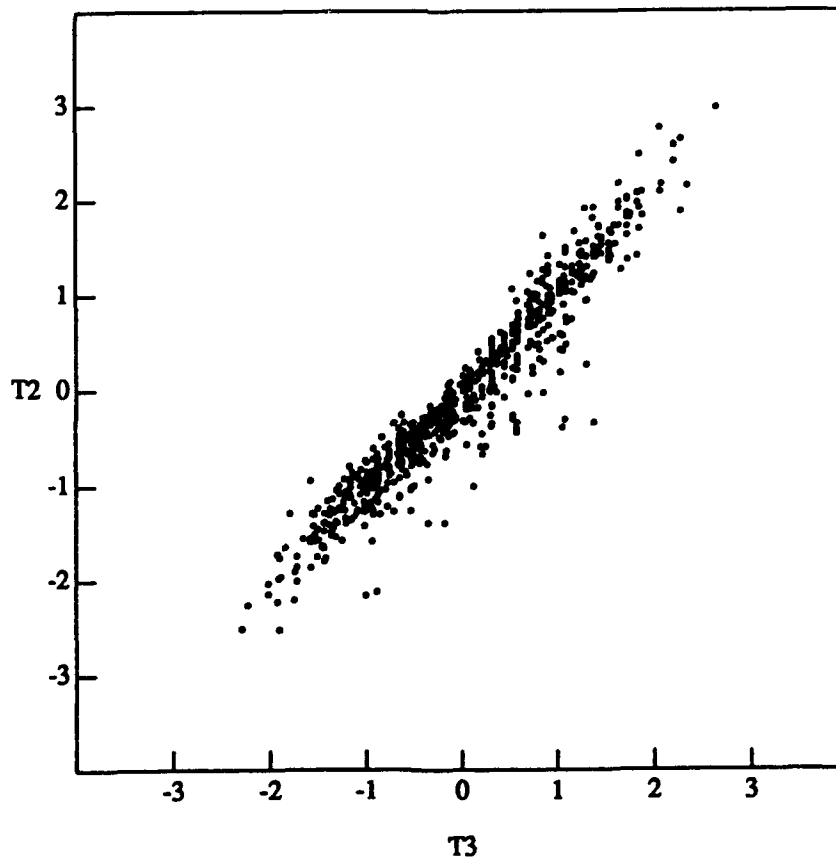
T2	1.0000	
T3	0.9632	1.0000
Variable	T2	T3

Regression Equation for T2:

$$T2 = 0.9211 T3 + 0.0290022$$

Significance test for prediction of T2

Mult-R	R-Squared	F(1,984)	prob (F)
0.9632	0.9277	12624.8208	0.0000



SHOP INFORMATION

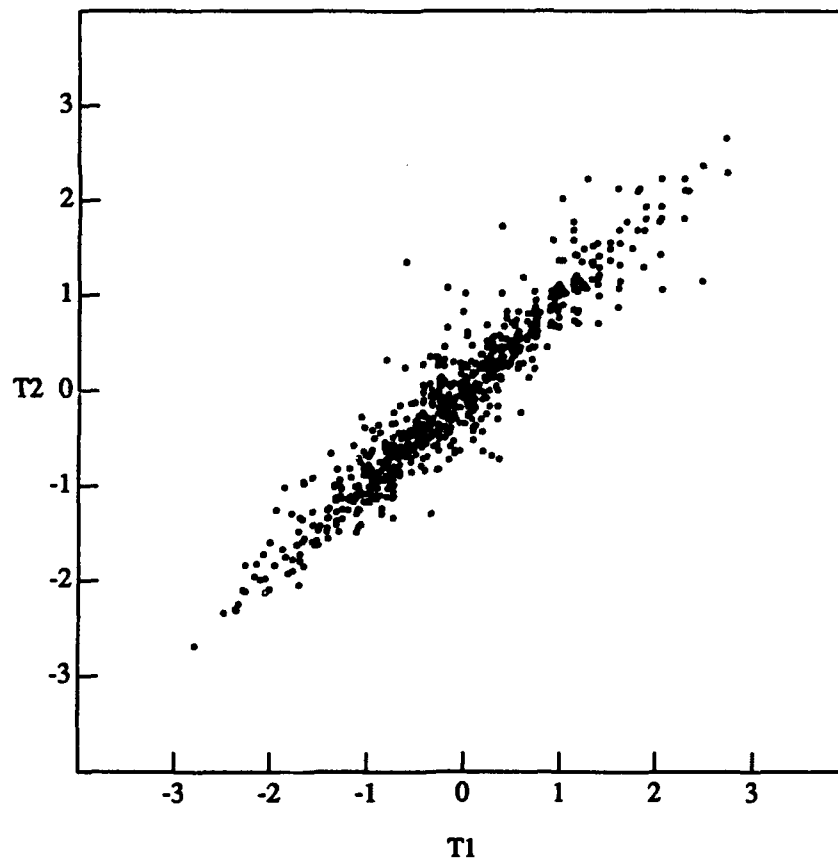
T1 = ADAPTIVE THETA (C1, U3), 10-items
T2 = ADAPTIVE THETA (C2, U3), 10-items
T3 = NON-ADAPTIVE THETA (C3, U3), 40-items

Analysis for 986 points of 2 variables:

Variable	T2	T1
Min	-2.7960	-2.6640
Max	2.7030	2.6750
Sum	12.1180	40.9270
Mean	0.0123	0.0415
SD	0.8962	0.8656

Correlation Matrix:

T2	1.0000	
T1	0.9564	1.0000
Variable	T2	T1



SHOP INFORMATION

T1 = ADAPTIVE THETA (C1, U3), 10-items
T2 = ADAPTIVE THETA (C2, U3), 10-items
T3 = NON-ADAPTIVE THETA (C3, U3), 40-items

Analysis for 986 points of 2 variables:

Variable	T1	T3
Min	-2.6640	-3.6840
Max	2.6750	2.5460
Sum	40.9270	-10.4760
Mean	0.0415	-0.0106
SD	0.8656	0.9146

Correlation Matrix:

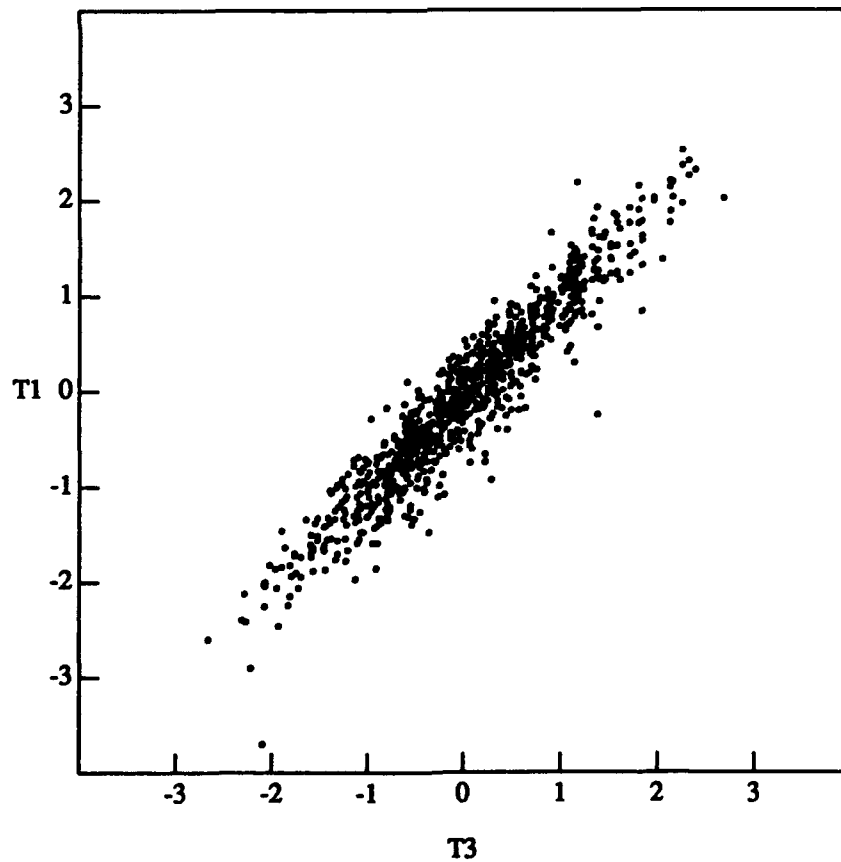
T1	1.0000	
T3	0.9508	1.0000
Variable	T1	T3

Regression Equation for T1:

$$T1 = 0.8999 T3 + 0.0510692$$

Significance test for prediction of T1

Mult-R	R-Squared	F(1,984)	prob (F)
0.9508	0.9041	9273.3592	0.0000



SHOP INFORMATION

T1 = ADAPTIVE THETA (C1, U3), 10-items
T2 = ADAPTIVE THETA (C2, U3), 10-items
T3 = NON-ADAPTIVE THETA (C3, U3), 40-items

Analysis for 986 points of 2 variables:

Variable	T2	T3
Min	-2.7960	-3.6840
Max	2.7030	2.5460
Sum	12.1180	-10.4760
Mean	0.0123	-0.0106
SD	0.8962	0.9146

Correlation Matrix:

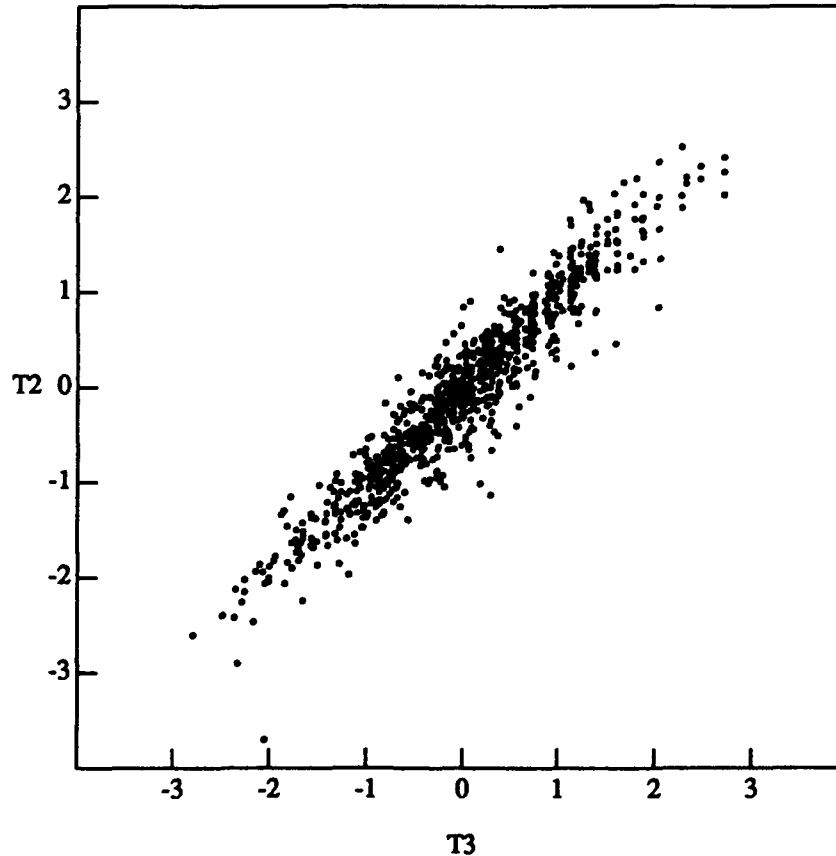
T2	1.0000	
T3	0.9507	1.0000
Variable	T2	T3

Regression Equation for T2:

$$T2 = 0.9316 T3 + 0.0221877$$

Significance test for prediction of T2

Multi-R	R-Squared	F(1,984)	prob (F)
0.9507	0.9039	9254.1786	0.0000



Distribution List

Bureau of Naval Personnel (PERS-23)

HQ USMEPCOM

Office of the Assistant Secretary of Defense (FM&P)

Defense Technical Information Center (DTIC) (4)