4

CRM 88-259 / March 1989

DTIC FILE COPY

RESEARCH MEMORANDUM

AD-A210 126

ANALYSIS OF DATA QUALITY FOR THE INFANTRY PHASE OF THE MARINE CORPS JOB PERFORMANCE MEASUREMENT PROJECT

Paul W. Mayberry

DTIC
ELECTE
JUL 17 1989
S B D

A Division of CNA *Hudson Institute*

CENTER FOR NAVAL ANALYSES
4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

Work conducted under contract N00014-87-C-0001.

This Research Memorandum represents the best opinion of CNA at the time of issue.
It does not necessarily represent the opinion of the Department of the Navy.

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3. DISTRIBUTION / AVAILABILITY OF REPORT
Approved for Public Release; Distribution Unlimited

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
CRM 88-259

5. MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION
Center for Naval Analyses

6b. OFFICE SYMBOL *(If applicable)*
CNA

7a. NAME OF MONITORING ORGANIZATION
Commanding General, Marine Corps Combat Development Command

6c. ADDRESS (*City, State, and ZIP Code*)
4401 Ford Avenue
Alexandria, Virginia 22302-0268

7b. ADDRESS (*City, State, and ZIP Code*)
Warfighting Center
Quantico, Virginia 22134

8a. NAME OF FUNDING ORGANIZATION
Office of Naval Research

8b. OFFICE SYMBOL *(If applicable)*
ONR

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
N00014-87-C-0001

8c. ADDRESS (*City, State, and ZIP Code*)
800 North Quincy Street
Arlington, Virginia 22217

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
65153M	C0031		

11. TITLE *(Include Security Classification)*
Analysis of Data Quality for the Infantry Phase of the Marine Corps Job Performance Measurement Project

12. PERSONAL AUTHOR(S)
Paul W. Mayberry

13a. TYPE OF REPORT
Final

13b. TIME COVERED
FROM TO

14. DATE OF REPORT *(Year, Month, Day)*
March 1989

15. PAGE COUNT
68

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP
05	08	

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
Aptitude tests, ASVAB (armed services vocational aptitude battery), Data acquisition, Job analysis, Marine corps personnel, Performance (human), Quality, Statistical data, Statistical samples, Test scores, Validation

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

All large-scale data collection efforts must contend with the issue of data quality. This research memorandum examines the quality of data collected for the infantry portion of the Marine Corps Job Performance Measurement Project. Particular attention is focused on data inconsistencies and imputation of missing data.

see p V

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
□ UNCLASSIFIED / UNLIMITED ☒ SAME AS RPT. □ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
UNCLASSIFIED

22a. NAME OF RESPONSIBLE INDIVIDUAL
Colonel Preston

22b. TELEPHONE *(Include Area Code)*

22c. OFFICE SYMBOL

83 APR edition may be used until exhausted.
All other editions are obsolete.

CNA

CENTER FOR NAVAL ANALYSES

A Division of Hudson Institute 4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268 • (703) 824-2000

11 April 1989

MEMORANDUM FOR DISTRIBUTION LIST

Subj: Center for Naval Analyses Research Memorandum 88-259

Encl: (1) CNA Research Memorandum 88-259, *Analysis of Data Quality for the Infantry Phase of the Marine Corps Job Performance Measurement Project*, by Paul W. Mayberry, Mar 1989

1. Enclosure (1) is forwarded as a matter of possible interest.

2. All large-scale data collection efforts must contend with the issue of data quality. This research memorandum examines the quality of data collected for the infantry portion of the Marine Corps Job Performance Measurement project. Particular attention is focused on data inconsistencies and imputation of missing data.

Lewis R Cabe

Lewis R. Cabe
Director
Manpower and Training Program

Distribution List:
Reverse page

Subj: Center for Naval Analyses Research Memorandum 88-259

Distribution List

SNDL

45B	CG FIRST MARDIV
45B	CG SECOND MARDIV
A1	ASSTSECNAV MRA
A1	DASN MANPOWER (2 Copies)
A2A	CNR
A6	HQMC MPR & RA
	Attn: Code M
	Attn: Code MR
	Attn: Code MP
	Attn: Code MM
	Attn: Code MA (3 copies)
A6	CG MCRDAC, Washington
A6	HQMC AVN
FF38	USNA
	Attn: Nimitz Library
FF42	NAVPGSCOL
FF44	NAVWARCOL
FJA1	COMNAVMILPERSCOM
FJB1	COMNAVCRUITCOM
FKQ6D	NAVPERSRANDCEN
	Attn: Technical Director (Code 01)
	Attn: Technical Library
	Attn: Director, Manpower Systems (Code 61)
	Attn: Director, Personnel Systems (Code 62)
FT1	CNET
V12	MCCDC
	Attn: Director, Warfighting Center
	Attn: Warfighting Center, MAGTF Proponency and Requirements Branch (2 copies)
	Attn: Director, Training and Education Center
V12	CG MCRDAC, Quantico

OPNAV

OP-01
OP-11
OP-12
OP-13

OTHER

Defense Advisory Committee on Military Personnel Testing (8 copies)
Joint Service Job Performance Measurement Working Group (13 copies)

CRM 88-259 / March 1989

ANALYSIS OF DATA QUALITY FOR THE INFANTRY PHASE OF THE MARINE CORPS JOB PERFORMANCE MEASUREMENT PROJECT

Paul W. Mayberry

A Division of CNA Hudson Institute

CENTER FOR NAVAL ANALYSES

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268

ABSTRACT

All large-scale data collection efforts must contend with the issue of data quality. This research memorandum examines the quality of data collected for the infantry portion of the Marine Corps Job Performance Measurement Project. Particular attention is focused on data inconsistencies and imputation of missing data.

DTIC
COPY
INSPECTED
1

Accession For
NTIS GRA&I ☑
DTIC TAB □
Unannounced □
Justification

By
Distribution/
Availability Codes

Dist	Avail and/or Special
A-1	

EXECUTIVE SUMMARY

The Marine Corps ~~Job Performance Measurement (JPM)~~ Project is a large-scale effort to validate the Armed Services Vocational Aptitude Battery (ASVAB) against measures of job performance. Over 2,500 infantrymen in five military occupational specialties (MOSs) were tested for two days each on a variety of performance measures. They were also readministered the ASVAB and given a battery of other new predictor tests. Although significant precautions were taken to minimize the possibility of poor or missing data, there were still individual cases in which the accuracy of the data was questionable and other instances in which the data simply were incomplete. Keywords; aptitude tests, job analysis, statistical data, statistical analysis, test scores -(KR)

IDENTIFICATION OF UNUSUAL RESPONSE PATTERNS

Occasionally, a test may fail to properly measure the ability of a particular person even though the test may provide excellent measurement for a group. For such persons, it is possible that some anomaly occurred while taking the test that produced unusual patterns of responses (e.g., inattentive marking of the answer sheet, random responses, application of wrong answer key). To identify aberrant response patterns, it is necessary to characterize the properties of normal response patterns and then contrast the individual responses to this norm while accounting for the probability of variation in response patterns. The personal biserial correlation (r_{perbis}) is a statistic that specifically quantifies the consistency of a person's item responses relative to the difficulty of each item.

Decision rules were established for the identification of aberrant response patterns based on r_{perbis} and percent correct score. Given these criteria, 36 scores were declared aberrant for the job knowledge test (JKT), 12 for the ASVAB, and 59 for the new predictor tests. Deleting these aberrant scores increased means for each test and decreased standard deviations, as shown in table I. The correlation of these three tests with hands-on total score (HOTS) and the General Technical (GT) aptitude composite scores improved slightly or remained the same. These changes in sample statistics indicated that these scores were typically outlier cases.

Table I. Change in sample statistics due to deleting aberrant scores

	Change in				
				Correlation with	
Test	N	Mean	SD	HOTS	GT
JKT	-36	.31	-.29	.02	.01
ASVAB	-12	.10	-.11	.00	.01
New predictors	-59	.67	-.58	.00	.00

Other information was examined to confirm the r_{perbis} statistic. Comparisons of enlistment ASVAB scores were made to current ASVAB scores to identify infantrymen with large negative discrepancies. Records that were maintained during the hands-on testing that identified persons having difficulty or lacking motivation in taking the tests were very consistent with the r_{perbis} statistic. Other self-report questionnaires asking the extent to which an examinee tried on the test were supportive but incomplete.

IMPUTATION OF MISSING DATA

Hands-on performance data were collected at a step level; a person either passed or failed performing a specific action. Steps were aggregated to form task scores, task scores were combined to produce duty area scores, and duty area scores were weighted to create a total score. It was not always possible to collect complete information for each person – there were over 600 steps for each hands-on test. Examinees could have incomplete data as a result of weather conditions, equipment failure or unavailability, being called away before completion, unobservable response by test administrator, etc. Decisions were made concerning the conditions under which data were sufficient so that imputation of missing data points was warranted.

Table II details the gain in complete-data cases resulting from imputation at each of the three imputation stages. Approximately 10 percent of the cases tested had complete data (i.e., no missing steps), except for the mortarman specialty (0341). While this may appear to be a low percentage, the majority of incomplete cases were missing only a few steps. As a result of step imputation, data cases were complete for approximately 75 percent of all persons tested. On average, about 5 step scores were imputed to complete these cases for four MOSs; over 31 step scores were imputed on average for the mortarmen. Relatively few cases were gained by imputation conducted at the task stage. The final stage of imputation at the duty area level resulted in over 95 percent of all cases tested having complete data.

Given this degree of imputation at the step, task, and duty area levels, what was the impact on the sample statistics of the respective HOTS? Sample statistics for all variables with complete information after the step-level imputation were compared to the sample statistics after imputation at the duty area level. The shifts in mean performance scores were relatively small compared to the standard deviation of the performance scores. The largest standardized change in means was observed for the mortarman specialty. Standard deviations increased slightly in all cases, as would be expected because the imputation was not based on a "substitution of the mean" process. Intercorrelation among the core infantry content and primary and secondary MOS scores were also relatively unchanged. Change statistics computed for all duty areas of each MOS were also insignificant. Across the five MOSs, the validity of the performance scores versus the GT aptitude composite was differentially affected, but again changes were insignificant.

Table II. Gains in complete-data cases resulting from imputation of missing data

Imputation stage	MOS 0311	0331	0341	0351	0369
Complete data	102	45	0	15	49
Additional cases					
Step level	883	205	221	223	266
Task level	27	6	3	2	15
Duty area level	262	53	83	73	65
Irretrievable cases	58	6	12	8	20
Total cases	1,332	315	319	321	415
Average number of steps imputed					
Step level	5.0	4.3	31.1	5.6	5.0
Task level	3.5	1.0	6.0	3.5	1.7
Duty area level	1.2	1.0	1.1	1.2	1.1
Percent of irretrievable cases	4%	2%	4%	2%	5%

CONCLUSIONS

Relatively few unusual response patterns were observed for the written tests. The aberrant data cases tended to be outliers so that their deletion generally improved sample correlations and reduced standard deviations. The criteria established to identify aberrant response patterns were specifically chosen to be conservative. Certainly, arguments could be made for different criteria. However, given the verification across different information sources (personal biserial correlation, percent correct score, residual analysis, problem logs, and self-report of effort), it was believed that few persons were misidentified as having aberrant patterns when, in fact, the test score was a reasonable estimate of their ability.

Imputation of missing data was required, in varying degrees, for over 90 percent of the examinees. Decisions were required that defined the circumstances in which sufficient data were available to warrant the imputation of missing data. Again, conservative ranges were established to mark the level of acceptable data for imputation. Sample statistics were insignificantly affected by imputation. Indeed, this was the intended outcome sought by employing an imputation procedure that incorporated steps to minimize the impact of imputed values.

As a result of these data quality analyses that identified unusual response patterns and imputed missing data for the infantry JPM data, further analytic investigations can proceed with confidence in the soundness of the data and the integrity of the results.

CONTENTS

ILLUSTRATIONS

ILLUSTRATIONS (Cont.)

TABLES

INTRODUCTION

The Marine Corps Job Performance Measurement (JPM) Project is a large-scale effort to validate the Armed Services Vocational Aptitude Battery (ASVAB) against measures of job performance. Extensive resources, time, and personnel have been devoted to the development and administration of performance tests for the infantry occupational field. In total, over 2,500 infantrymen in five military occupational specialties (MOSs) were tested for two days each on a variety of performance measures. They were also readministered the ASVAB and given a battery of other new predictor tests. The volume of data collected was enormous.

Because of the many potential problems that may beset large-scale data collection efforts, significant precautions were taken to minimize the possibility of poor and/or missing data. Particular attention was devoted to the design of all testing material to preclude the possibility of not being able to collect data as a result of the testing process [1]. Specifically, pilot testing and tryouts were conducted for all tasks of the hands-on performance tests. Clarity of administrator instructions and scoring procedures was established before full-scale testing commenced. Standardized training of test administrators was conducted to ensure that administrators accurately, objectively, and reliably scored the performance that they observed. Administrators were also instructed in the management and setup of the testing station so that testing would be orderly and systematic. Estimates of completion times were obtained for each testing station so that tasks could be reallocated to ensure that examinees had ample time to complete each testing station. For examinees unable to complete a testing station during the allotted time, efforts were made to finish the test during lunch or at the end of the day. Continuous monitoring during the testing identified potential problems so that corrective actions could be taken. This monitoring included the verification of all answer documents, daily computer entry of all hands-on responses, and maintenance of problems logs to identify specific problem cases.

Despite these initial preparations and quality-control procedures, there were still individual cases in which the accuracy of the data were questionable and other instances in which the data simply were incomplete. Both of these factors are critical to the overall data quality and may affect analyses yet to be conducted on the JPM data.

To identify data inaccuracies at the individual level, item responses were examined for unusual patterns (such as getting very easy items wrong while responding correctly to very difficult items.). Such data inaccuracies could be caused by random responses, guessing, cheating, misunderstanding test instructions, accidentally responding to wrong item numbers on the answer sheet, and so on. Therefore, unusual response patterns are not a true measure of a person's ability. There is no recovery of data identified as having unusual response patterns; the data must be declared missing. Identifying unusual response patterns applied to written tests only. Because hands-on performance testing was one-on-one, the test administrator served as a monitor to correct any misconceptions or random responses as they occurred.

An examinee may also have incomplete data as a result of weather conditions, equipment failure or unavailability, being called away before completion, unobservable response by test administrator, etc. These conditions imply that missing data were the result of a random event that was not under the control of the examinee. This was in contrast to persons who did not know the answer (and thereby did not record a response) or did not complete the test because of time constraints. In these instances, the responses were marked as wrong, not missing. For those persons with missing data, some data are better than no data (within limits) and the available data can be used to estimate missing data. Specific rules were established at the step, task, and duty area levels defining conditions in which too much missing data made a case irretrievable.

Given that the analyses to be conducted on the JPM data are sensitive to outliers and generally require complete information, this research memorandum presents specific procedures to ensure the quality and completeness of the infantry data of the Marine Corps JPM Project. Methods for identifying unusual response patterns in the written tests of the project are described, and the deletion of such aberrant data is justified based on a verification across different information sources. The magnitude of missing data for the hands-on performance tests at the step, task, and duty area levels is presented. The impact on sample descriptive statistics and intercorrelations due to deleting aberrant data from the written tests and estimating missing data points for the hands-on tests is noted.

IDENTIFICATION OF UNUSUAL RESPONSE PATTERNS

Occasionally a test may fail to properly measure the ability of a particular *person* even though the test may provide excellent measurement for a group. For such persons, it is possible that some anomaly occurred while taking the test that produced unusual patterns of responses. As discussed earlier, examples of such anomalies may include inattentive marking of the answer sheet, random responses, or application of the wrong answer key.

Methodology

Four forms of the job knowledge test (JKT) were administered to the rifleman MOS and only two forms to the other specialties. Two forms of the ASVAB were also administered. Examinees marked the test-form identifier on their scannable answer sheets. To verify the form code for each written test (or to determine a form code if one was not marked), all answer sheets were scored against all answer keys. To verify the correct form code, comparisons of individual total scores resulting from each answer key were made. Typically, higher total scores indicated the correct test form. For borderline cases in which there was no difference in total scores, the reported test form was used. For the ASVAB, the speeded tests (numerical operations and coding speed) readily identified the correct test form because these tests are composed of very easy items that should be correctly solved. These procedures corrected 32 cases of misidentified or missing form codes for the JKTs and 43 cases for the ASVAB.

To identify other aberrant response patterns, it is necessary to characterize the properties of normal response patterns and then contrast the responses of individual examinees to this norm while accounting for the probability of variation in response patterns. Patterns of "normal" responses can be defined as a function of persons tested; however, such patterns are sample dependent. Therefore, caution must be exercised to minimize the possibility of incorrectly identifying serious attempts by examinees (particularly persons of low ability) as inappropriate measures of ability.

Donlon and Fischer [2] proposed a statistic that specifically quantifies the internal consistency of a person's item responses relative to the difficulty of each item. Called the personal biserial correlation (r_{perbis}), the statistic quantifies the similarity between item difficulties as experienced by a particular person relative to the item difficulties computed for a reference sample. The statistic ranges from 1 to -1 and is interpreted as any correlation coefficient. High positive values indicate that the pattern of responses for one examinee is quite similar to the pattern of item difficulties experienced by the reference sample. Low or negative values indicate that the pattern of responses for a single examinee is poorly or inversely related to the item difficulties of the reference sample, and thus the response pattern is atypical relative to the expectation. Computation of the r_{perbis} statistic is discussed in appendix A.

Given that no absolute standard exists against which to validate or invalidate a person's score based on the magnitude of r_{perbis}, the use of this correlation as the sole criterion would be questionable. Correlations can be insensitive to departures from linearity and to individual differences in the measurement consistency of one's test score. Therefore, other information was used to supplement r_{perbis}. This information included verification against the daily problem logs that identified specific examinees noted as having difficulty or lacking motivation. "About taking these tests" questionnaires were administered, asking the extent to which the person tried on the test. This information was useful in examining individual cases. For the ASVAB testing, residual analyses were conducted based on the regression of enlistment scores on the concurrent ASVAB scores. Large negative discrepancies between enlistment and concurrent aptitude scores identified persons whose concurrent scores were not accurate indicators of their ability.

Results

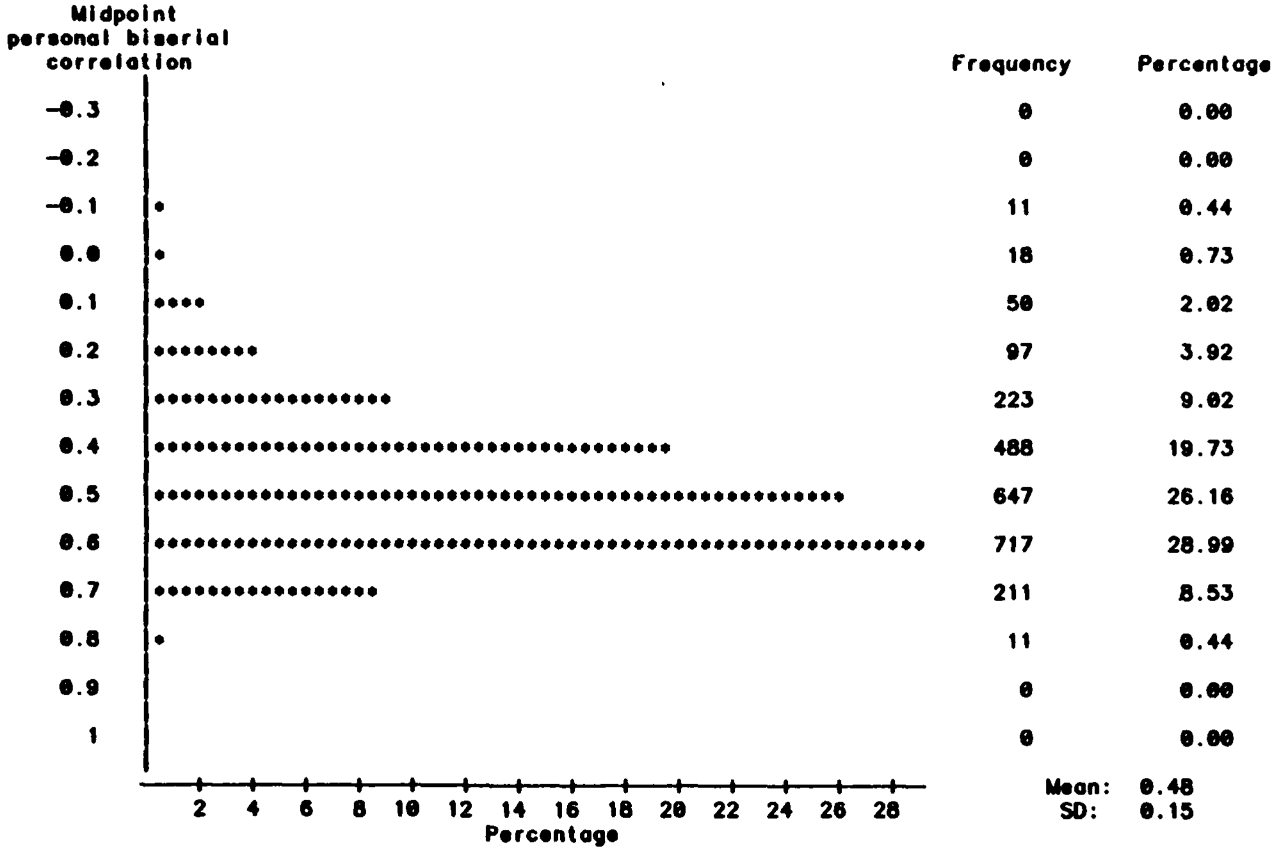
Figures 1 through 3 report the distributions of r_{perbis} for the JKT, the ASVAB, and the new predictor tests. Although the mean values for r_{perbis} were in the range of 0.50, the lower tail of each distribution was the primary area of interest. These lower correlations possibly identified examinees for whom test performance (at the item level) was not consistent with normal or expected performance relative to the entire sample. It was interesting to note that the highest mean r_{perbis} was computed for the ASVAB. As a motivational incentive to seriously take this aptitude test in a research setting, examinees were instructed that their scores of record would be permanently changed if their performance exceeded their previous aptitude performance (however, lower aptitude scores would not become part of the permanent record). This could have significant payoff for persons who wanted to transfer to other occupational fields with higher aptitude requirements. Given the limited number of persons with a low ASVAB r_{perbis}, it appeared that this incentive was effective.

The magnitude of r_{perbis} is not sufficient to invalidate a person's test score. The relationship between r_{perbis} and total score is not necessarily linear, but more typically quadratic. That is, high-ability examinees may also have low r_{perbis} by missing extremely easy items while performing correctly on all the difficult items. Figures 4 through 6 illustrate the relationship between the r_{perbis} and test performance (percent correct score) for each of the three written tests. Two decision rules were established for the identification of aberrant response patterns:

- $r_{perbis} <= 0.15$ and percent correct score $<= 25\%$, or
- $r_{perbis} <= 0.00$.

These critical regions that define aberrant scores are noted on the figures.

Based on these criteria, 36 scores were declared aberrant for the JKT, 12 for the ASVAB, and 59 for the new predictor tests. As a result of deleting these aberrant scores, means for each test increased and standard deviations went down (see table 1). The correlation of these three tests

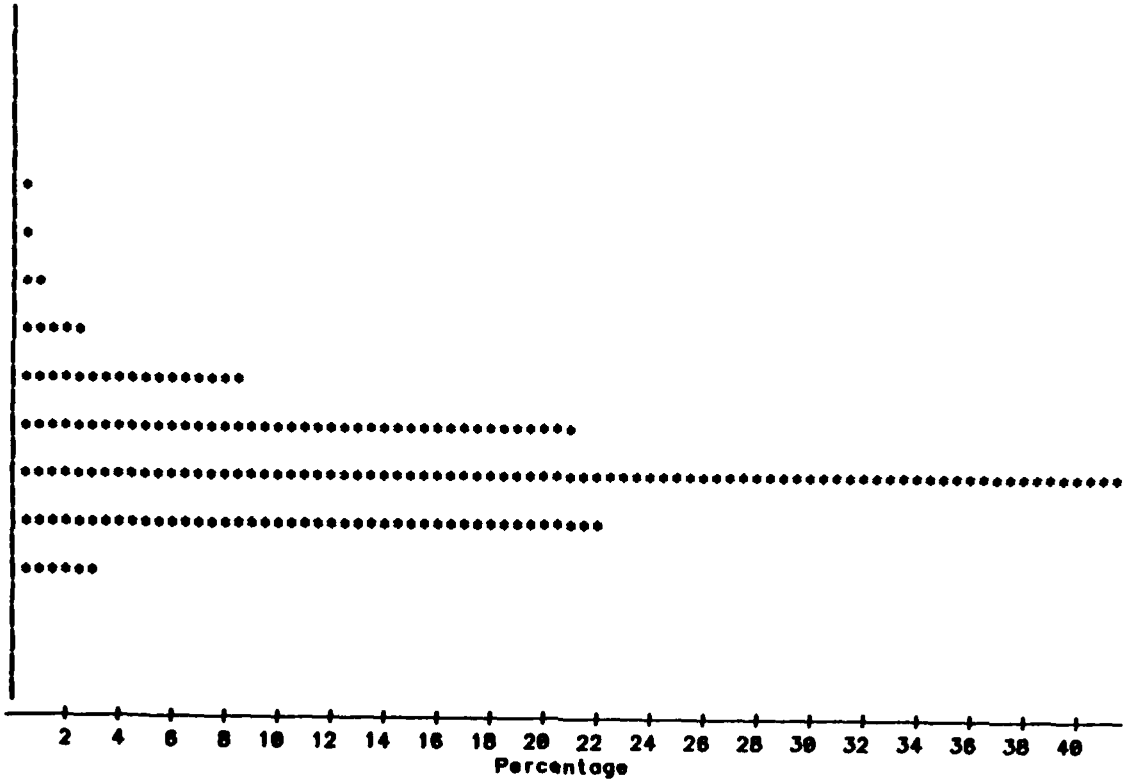


Midpoint personal biserial correlation	Frequency	Percentage
−0.3	0	0.00
−0.2	0	0.00
−0.1	11	0.44
0.0	18	0.73
0.1	50	2.02
0.2	97	3.92
0.3	223	9.02
0.4	488	19.73
0.5	647	26.16
0.6	717	28.99
0.7	211	8.53
0.8	11	0.44
0.9	0	0.00
1	0	0.00

Mean: 0.48
SD: 0.15

Figure 1. Frequency histogram of personal biserial correlation for job knowledge test

Midpoint personal biserial correlation	Frequency	Percentage
−0.3	0	0.00
−0.2	0	0.00
−0.1	3	0.12
0.0	8	0.32
0.1	10	0.41
0.2	24	0.97
0.3	56	2.27
0.4	212	8.59
0.5	523	21.19
0.6	1019	41.29
0.7	539	21.84
0.8	73	2.96
0.9	1	0.04
1	0	0.00

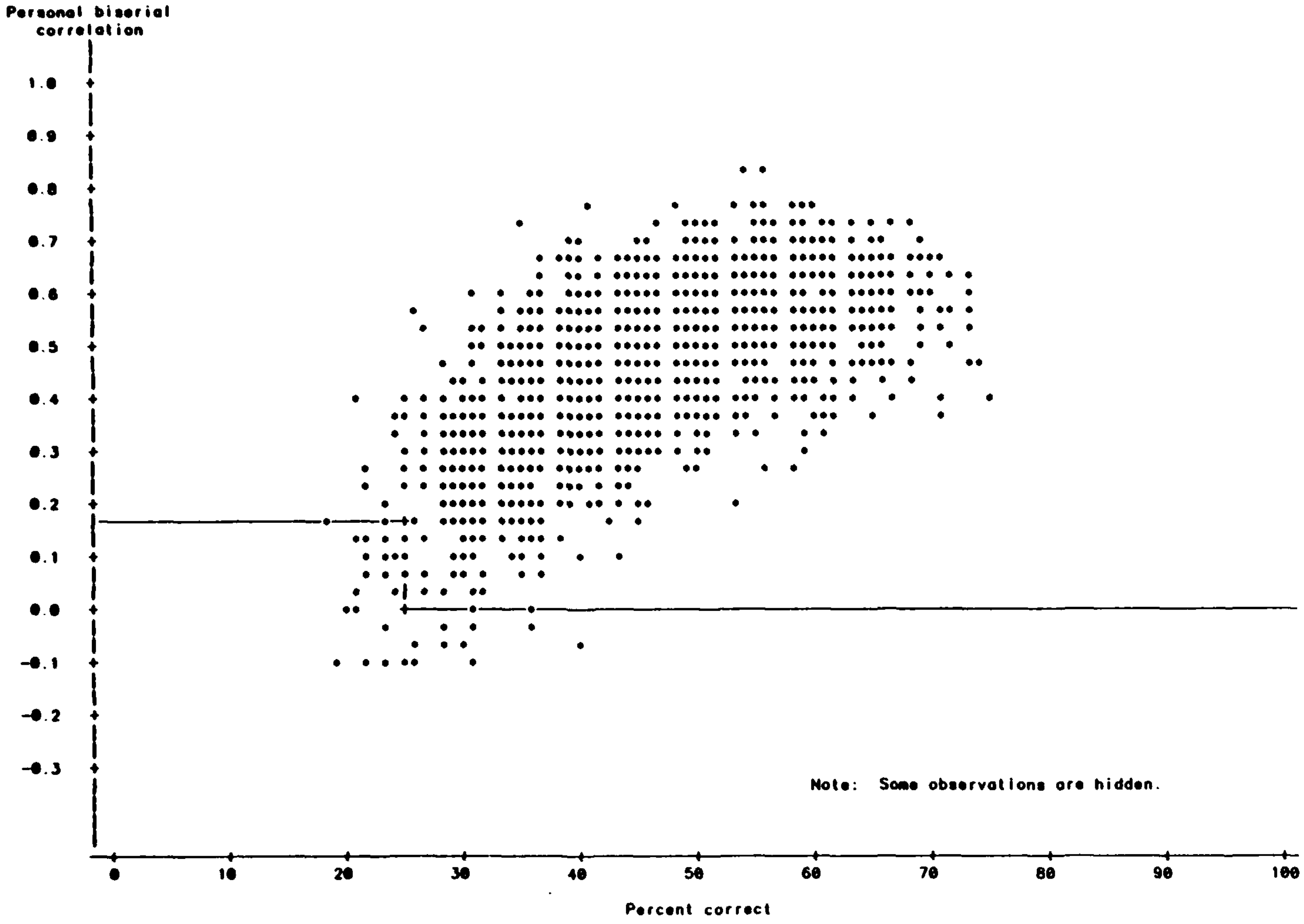


Mean: 0.57
SD: 0.12

Figure 2. Frequency histogram of personal biserial correlation for ASVAB

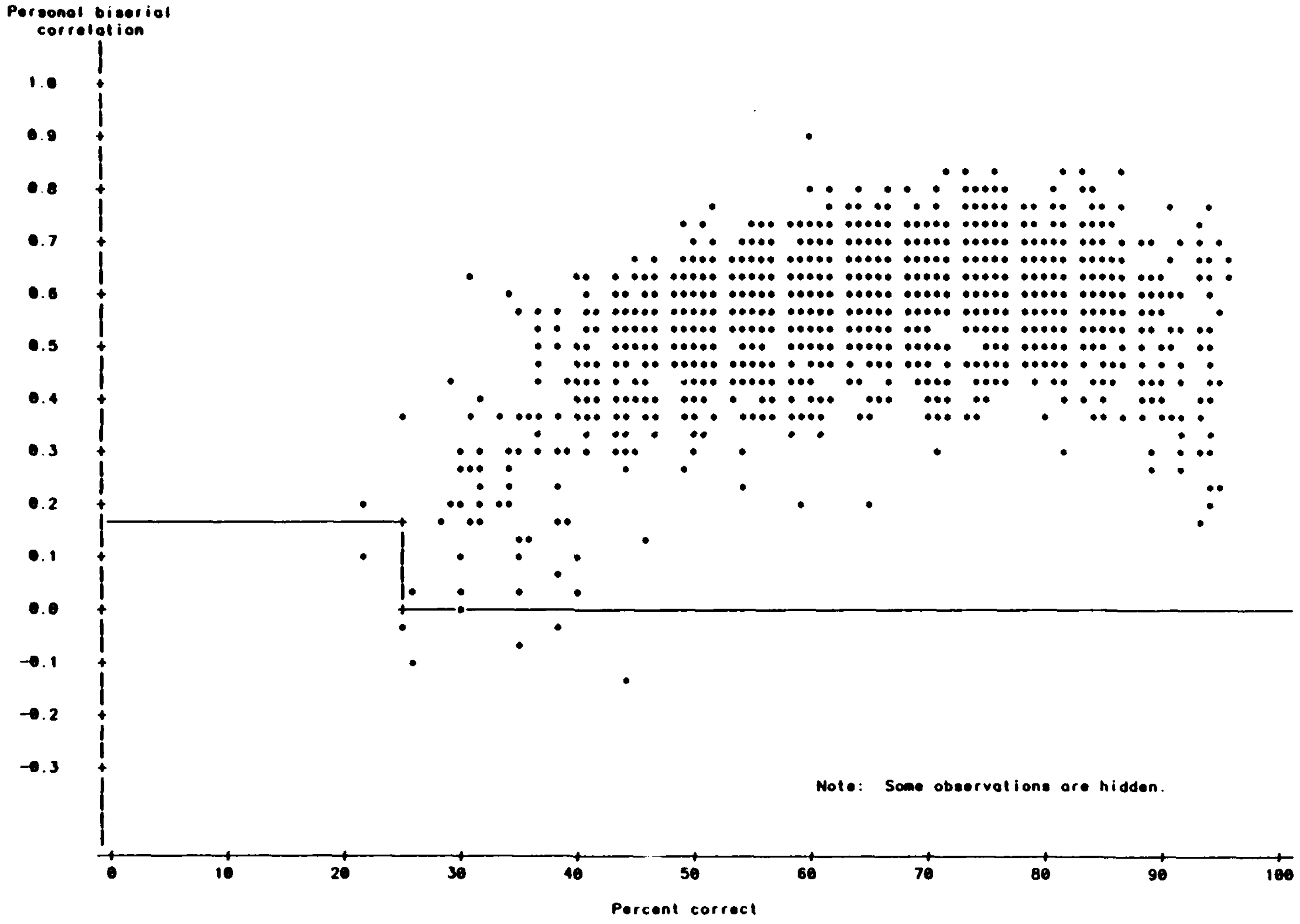
```
    Midpoint
personal biserial
  correlation                                                     Frequency   Percentage
                 |
     -0.3        |                                                       2        0.10
                 |
     -0.2        |                                                       3        0.14
                 |
     -0.1        |*                                                     10        0.48
                 |
      0.0        |**                                                    20        0.96
                 |
      0.1        |*****                                                 57        2.73
                 |
      0.2        |*********                                             93        4.45
                 |
      0.3        |****************                                     166        7.94
                 |
      0.4        |*****************************                        302       14.44
                 |
      0.5        |****************************************             417       19.94
                 |
      0.6        |*********************************************        475       22.72
                 |
      0.7        |*******************************                      325       15.54
                 |
      0.8        |****************                                     165        7.89
                 |
      0.9        |*****                                                 56        2.68
                 |
        1        |                                                       0        0.00
                 |
                 ----+---+---+---+---+---+---+---+---+---+---+-       Mean:   0.52
                     2   4   6   8  10  12  i4  16  18  20  22         SD:   0.19
                                Percentage
```

Figure 3. Frequency histogram of personal biserial correlation for new predictor tests



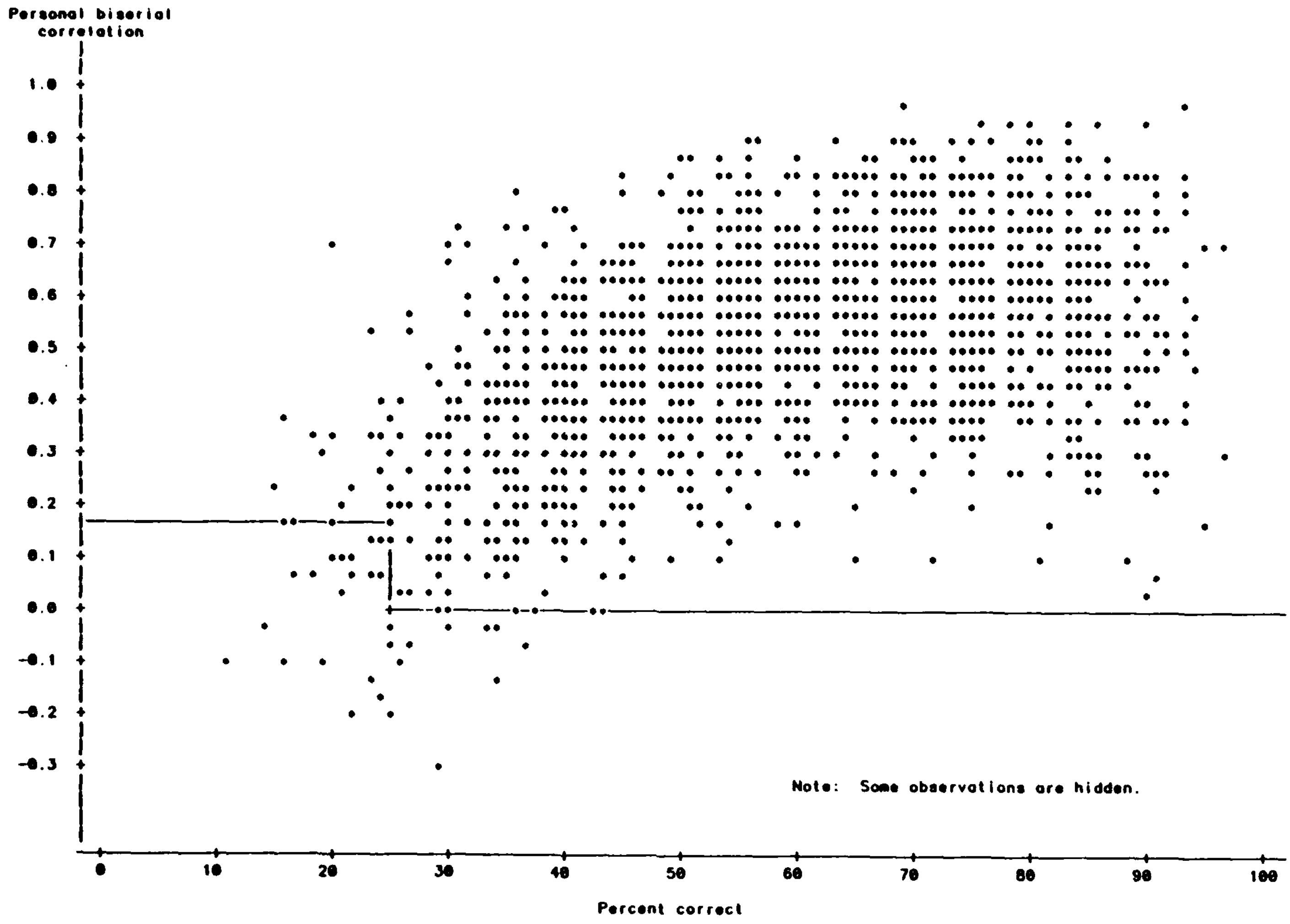
Note: Some observations are hidden.

Figure 4. Relationship between personal biserial correlation and percent correct score for job knowledge test



Note: Some observations are hidden.

Figure 5. Relationship between personal biserial correlation and percent correct score for ASVAB



Note: Some observations are hidden.

Figure 6. Relationship between personal biserial correlation and percent correct score for new predictor tests

with both hands-on total score (HOTS) and the General Technical (GT) aptitude composite score improved slightly or remained the same. These changes in sample statistics indicate that these scores were typically outlier cases.

Table 1. Change in sample statistics due to deleting aberrant scores

	Change in				
				Correlation with	
Test	N	Mean	SD	HOTS	GT
JKT	-36	.31	-.29	.02	.01
ASVAB	-12	.10	-.11	.00	.01
New predictors	-59	.67	-.58	.00	.00

An additional check was applied to identify aberrant scores for the ASVAB. The GT aptitude composite score from the ASVAB administered during the JPM testing was regressed on the GT composite score obtained at the time of enlistment in the Marine Corps. Residuals were computed from this regression and plotted against r_{perbis}, as shown in figure 7. In this manner, those who had ASVAB scores extremely below their enlistment scores (greater than -3 standard deviations from the mean) and also a low aberrant index (r_{perbis} less than 0.25) were determined to have invalid scores. Those persons who had significantly improved their score were not of concern. Although five examinees satisfied these criteria, only three were unique and had not been excluded based on earlier tests.

A final qualitative verification of r_{perbis} involved the problem logs maintained for each of the written tests. These logs identified examinees having difficulty or lacking motivation in taking the tests. Other circumstances that potentially affected test performance were also noted in the logs: ex-

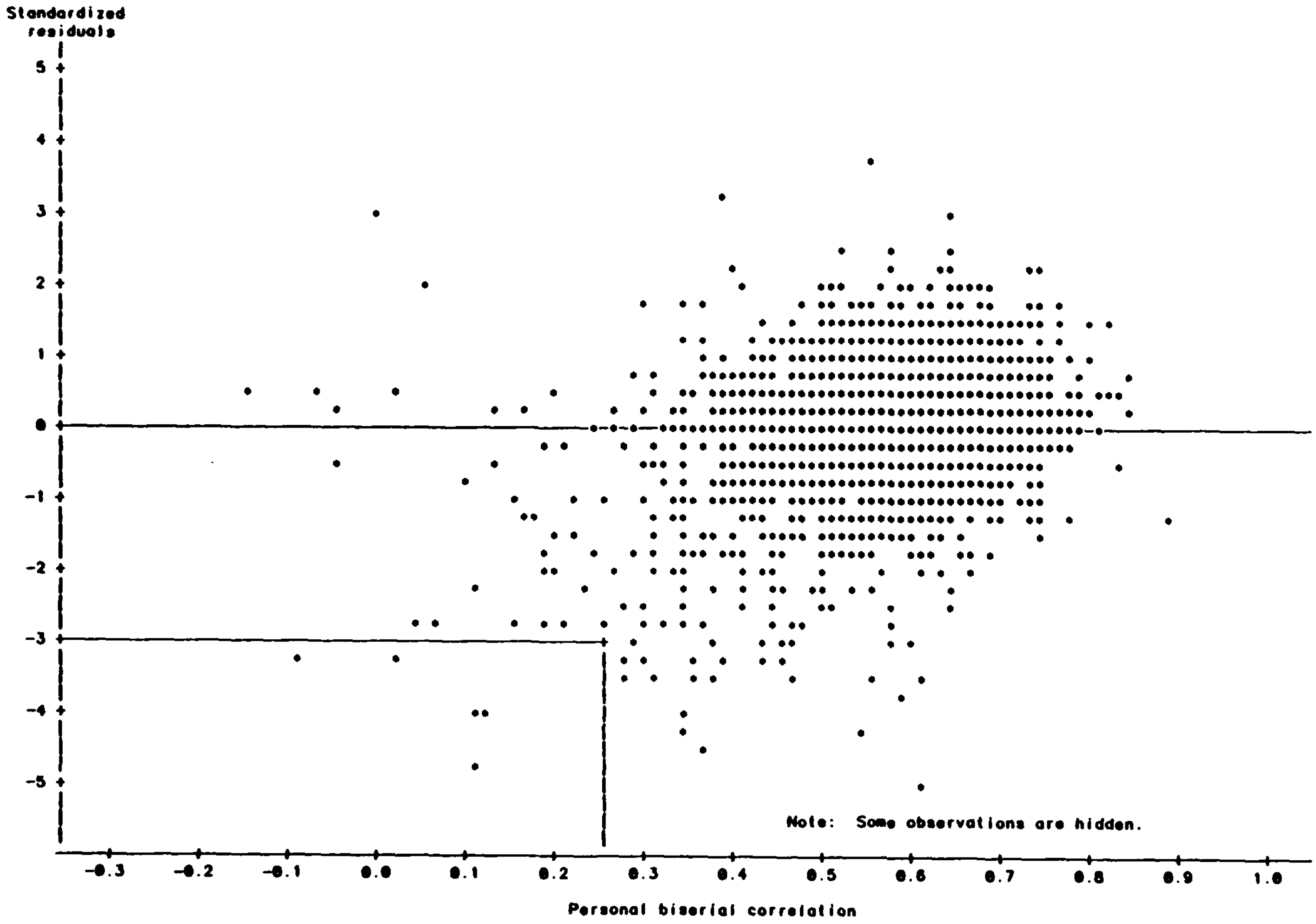


Figure 7. Relationship between standardized residual of general technical aptitude composite score and personal biserial correlation

aminees taking medications or who had been on firewatch (patrol duty) for the 24-hour period before testing. Because the problem logs were based on observable abnormalities in the testing behaviors of examinees (or the self-reporting of a problem by an examinee), the problem logs were not as extensive in identifying problem cases as were the quantitative r_{perbis} statistics. However, the problem logs were very consistent with the r_{perbis} statistics – examinees noted as having difficulties by the test administrators also had relatively low r_{perbis} values. The "about taking these tests" questionnaires, which asked the examinee to rate the extent to which he tried on the test, were also a source of information. However, these ratings were not consistently related to either r_{perbis} or the problem logs.

IMPUTATION OF MISSING DATA

Data collected for the Marine Corps JPM Project were extremely difficult and expensive to obtain. Despite the best of intentions, it was not always possible to collect complete information for each person. Given the extensive resources devoted to the project, every effort should be made to use whatever data were collected for each case.

Methodology

The National Academy of Sciences Committee on the Performance of Military Personnel, an oversight committee for the Joint Service JPM Project, recommended employing an imputation procedure that estimates missing data so that complete-case analysis can be conducted [3]. The recommended imputation algorithm, developed by Wise and McLaughlin [4], was a regression-based procedure. The procedure seeks to impute missing values by taking into account the differing levels of item difficulties while also maintaining individual differences among examinees. The technique incorporates a random component equal to the error of estimate to prevent unduly high correlations among variables with imputed values compared to variables with nonimputed values. The procedure also sequentially estimates multiple missing variables for the same person using a multistage process that relies on previously imputed values for the imputation of successive missing values. Further discussion of the computational procedures for the imputation of missing data is presented in appendix A. Before such an imputation procedure can be implemented, decision rules must be established to specify the conditions under which there are sufficient data to warrant the use of imputation.

Hands-on performance data were collected at a step level; an examinee either passed or failed performing a specific action. Steps were then aggregated to form task scores, task scores were combined to produce duty area scores, and duty area scores were weighted to create a total score. In total, over 600 steps were accumulated into more than 65 task scores, which were reduced to at least 13 duty area scores, which were combined into a single total score. Based on this hierarchy of scores, decisions had to be made at each level before computation of scores could proceed at higher levels. Figure 8 diagrams the sequence of events and decisions rules required for imputation of missing data and computation of scores at each of the three score levels.

The imputation process began by computing the percentage of missing steps within a duty area. If this was less than 15 percent, it was determined that imputation of missing step data was appropriate. The imputation of missing steps was based on all available step information within the duty area.

Each task within the duty area was then examined for complete step information. A task score was computed for those tasks with no missing steps (defined as either nonmissing or imputed steps). Tasks that had missing steps were assigned missing task scores because imputation was determined to be inappropriate due to the large number of missing steps.

Next, the percentage of missing task scores within the duty area was computed. If this did not exceed 20 percent, task scores were imputed based on the remainder of the task scores of the duty area. A duty area score was then computed for those persons with complete task information. In those cases for which over 20 percent of the tasks within a duty area were missing scores, the duty area score was assigned a missing value.

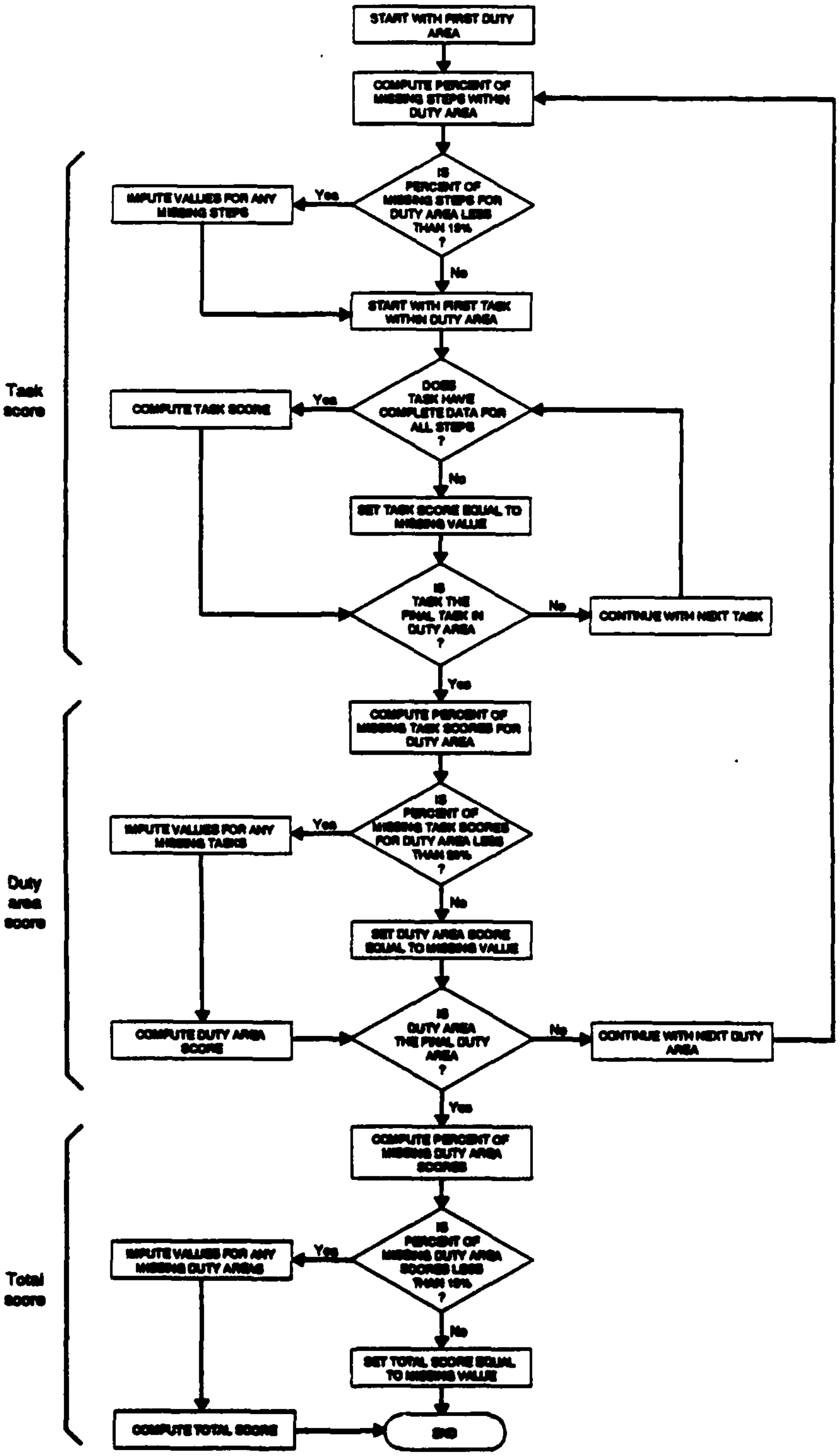

START WITH FIRST DUTY AREA
COMPUTE PERCENT OF MISSING STEPS WITHIN DUTY AREA
IS PERCENT OF MISSING STEPS FOR DUTY AREA LESS THAN 15% ?
Yes
IMPUTE VALUES FOR ANY MISSING STEPS
No
START WITH FIRST TASK WITHIN DUTY AREA
Task score
DOES TASK HAVE COMPLETE DATA FOR ALL STEPS ?
Yes
COMPUTE TASK SCORE
No
SET TASK SCORE EQUAL TO MISSING VALUE
IS TASK THE FINAL TASK IN DUTY AREA ?
No
CONTINUE WITH NEXT TASK
Yes
COMPUTE PERCENT OF MISSING TASK SCORES FOR DUTY AREA
Duty area score
IS PERCENT OF MISSING TASK SCORES FOR DUTY AREA LESS THAN 20% ?
Yes
IMPUTE VALUES FOR ANY MISSING TASKS
No
SET DUTY AREA SCORE EQUAL TO MISSING VALUE
COMPUTE DUTY AREA SCORE
IS DUTY AREA THE FINAL DUTY AREA ?
No
CONTINUE WITH NEXT DUTY AREA
Yes
COMPUTE PERCENT OF MISSING DUTY AREA SCORES
Total score
IS PERCENT OF MISSING DUTY AREA SCORES LESS THAN 15% ?
Yes
IMPUTE VALUES FOR ANY MISSING DUTY AREAS
No
SET TOTAL SCORE EQUAL TO MISSING VALUE
COMPUTE TOTAL SCORE
END

Figure 8. Process for imputing missing data and computing hands-on performance scores

This process continued for each duty area until all had been processed. At that point, the percentage of duty areas with missing scores was determined. If a person had all duty area scores, a hands-on total score was computed. If the percentage of missing duty area scores was less than 15 percent, the missing duty area score was imputed based on the other nonmissing duty area scores. A total score was then computed from the nonmissing and imputed duty area scores. For those cases in which the percentage of missing duty area scores exceeded 15 percent, no imputation was conducted and the total score was declared missing.

Results

The imputation strategy was applied separately for each MOS. Table 2 details the gain in complete-data cases resulting from imputation at each of the three stages. Approximately 10 percent of the cases tested had complete data (i.e., no missing steps), except for the mortarman specialty *(0341). While this may appear to be a low percentage of complete* data, the majority of incomplete cases were missing only a few steps (over 600 steps composed each hands-on test). Reasons for examinees missing steps included:

- Equipment failure – low batteries in night vision device
- Equipment unavailable for a limited time – atropine injectors, claymore mine
- Refusal of subject to perform – mouth-to-mouth resuscitation on artificial dummy
- Weather conditions – lightning during outdoor testing
- Inconsistent scoring by administrators – visual inspection of grenade launcher.

Table 2. Gains in complete-data cases resulting from imputation of missing data

Imputation stage	MOS				
	0311	0331	0341	0351	0369
Complete data	102	45	0	15	49
Additional cases					
Step level	883	205	221	223	266
Task level	27	6	3	2	15
Duty area level	262	53	83	73	65
Irretrievable cases	58	6	12	8	20
Total cases	1,332	315	319	321	415
Average number of steps imputed					
Step level	5.0	4.3	31.1	5.6	5.0
Task level	3.5	1.0	6.0	3.5	1.7
Duty area level	1.2	1.0	1.1	1.2	1.1
Percent of irretrievable cases	4%	2%	4%	2%	5%

For examinees with less than 15 percent missing steps within a duty area, step scores (1/0) were imputed. As a result of step imputation, data cases became complete for approximately 75 percent of all examinees tested. On average, about five step scores were imputed to complete these cases for four MOSs. However, over 31 step scores were imputed on average for the mortarmen. These missing steps dealt primarily with tasks of the 81-mm mortar duty area that were not completed because of equipment failure (broken lensatic compass).

Relatively few cases were gained by imputation conducted at the task stage because most examinees had complete task scores after the step imputation. The degree of imputation at this level was also minimal with one to six task scores being imputed on average across the five MOSs.

Duty area scores were imputed for those remaining incomplete cases that had only one or two missing duty area scores. This final stage of imputation resulted in 95 percent or better of all cases tested having complete duty area and total score data. Table 3 notes the frequency of imputation for each duty area by MOS. The rifleman specialty (0311) required imputation primarily on duty areas tested outdoors: M16A2 rifle, mines, and hand grenades. Difficulties in maintaining adequate nuclear, biological, and chemical defense supplies resulted in imputation for the mortarman and unit leader specialties. Imputation for the dragon duty area was necessary for the assaultman specialty (0351) due to some equipment unreliability.

Given this degree of imputation at the step, task, and duty area levels, what was the impact on the sample statistics of the respective hands-on performance scores? Tables 4 through 8 present the changes in means, standard deviations, and correlations for each MOS as a result of gains in complete-data cases due to imputation from the step level to the duty area level. The shifts in mean performance scores are relatively small compared to the standard deviation of the performance scores. These standard deviations of the performance scores are reported in the table footnotes. The largest standardized change in means was observed for the mortarman specialty. The 0.6 change in HOTS represents a 0.07 change in standard score units. Standard deviations increased slightly in all cases, as would be

Table 3. Frequency of imputation for each duty area by MOS

Duty area	MOS 0311	 0331	 0341	 0351	 0369
Communications	9	24	0	0	4
First aid	4	3	0	0	1
Grenade launcher	19	0	8	0	0
Hand grenade	56	7	5	7	4
Light antitank weapon	6	0	0	1	2
Land navigation	8	0	0	0	3
Mines	54	2	1	4	9
Nuclear, biological, chemical defense	8	3	23	6	15
Night vision device	20	6	4	0	10
Security and intelligence	10	0	6	1	0
Tactical measures	5	0	1	1	2
Squad automatic weapon	13	1	0	13	6
M16A2 rifle firing	94	5	22	a	a
Machine gun	a	4	a	a	a
Mortar	a	a	18	a	3
Dragon	a	a	a	44	a
Shoulder-mounted assault weapon	a	a	a	7	a
Operations order	a	a	a	a	10

a. Duty area is not a job requirement for this specialty.

Table 4. Change in sample statistics due to adding imputed values: MOS 0311 (rifleman)

Performance score	N	Mean[a]	SD	Correlation with				
				HOTS	CORE	MOS1	MOS2	GT
HOTS	+289	-.2	.1	-	.00	-.01	.00	-.01
CORE	+289	-.3	.1	.00	-	-.02	.01	-.01
MOS1	+289	.0	.1	-.01	-.02	-	.00	-.02
MOS2	+289	-.1	.1	.00	.01	.00	-	-.02
Average[b] over 13 duty areas	+289	.4	.1	.01	.01	.01	.01	.01

NOTE: Change reflects differences in sample statistics after imputation at the step level versus at the duty area level. The results at the step level serve as the base.

a. The original standard deviations of these performance scores against which to compare changes in means are as follows: HOTS, 9; CORE, 9; MOS1, 16; and MOS2, 16.

b. Absolute change averaged over all duty areas.

Table 5. Change in sample statistics due to adding imputed values: MOS 0331 (machinegunner)

Performance score	N	Mean[a]	SD	Correlation with HOTS	CORE	MOS1	MOS2	GT
HOTS	+59	-.1	.1	-	-.01	-.01	-.03	.01
CORE	+59	.1	.2	-.01	-	-.03	-.04	.01
MOS1	+59	-.1	.6	-.01	-.03	-	-.01	.01
MOS2	+59	-.5	.0	-.03	-.04	-.01	-	-.01
Average[b] over 14 duty areas	+59	.5	.3	.02	.02	.03	.02	.02

NOTE: Change reflects differences in sample statistics after imputation at the step level versus at the duty area level. The results at the step level serve as the base.

a. The original standard deviations of these performance scores against which to compare changes in means are as follows: HOTS, 8; CORE, 9; MOS1, 10; and MOS2, 15.

b. Absolute change averaged over all duty areas.

Table 6. Change in sample statistics due to adding imputed values: MOS 0341 (mortarman)

Performance score	N	Mean[a]	SD	Correlation with HOTS	CORE	MOS1	MOS2	GT
HOTS	+86	.6	.2	-	.00	.01	.03	.00
CORE	+86	.5	.1	.00	-	.00	.02	.00
MOS1	+86	.6	.3	.01	.00	-	.06	.03
MOS2	+86	.9	.1	.03	.02	.06	-	-.03
Average[b] over 14 duty areas	+86	.5	.4	.02	.02	.02	.03	.03

NOTE: Change reflects differences in sample statistics after imputation at the step level versus at the duty area level. The results at the step level serve as the base.

a. The original standard deviations of these performance scores against which to compare changes in means are as follows: HOTS, 9; CORE, 9; MOS1, 14; and MOS2, 16.

b. Absolute change averaged over all duty areas.

Table 7. Change in sample statistics due to adding imputed values: MOS 0351 (assaultman)

Performance				Correlation with				
score	N	Mean[a]	SD	HOTS	CORE	MOS1	MOS2	GT
HOTS	+75	-.4	.4	-	.02	.02	.04	-.02
CORE	+75	-.4	.2	.02	-	.03	.06	.02
MOS1	+75	-.1	.1	.02	.03	-	.02	.01
MOS2	+75	-1.3	1.0	.04	.06	.02	-	-.03
Average[b] over 14 duty areas	+75	.8	.5	.04	.04	.03	.02	.02

NOTE: Change reflects differences in sample statistics after imputation at the step level versus at the duty area level. The results at the step level serve as the base.

a. The original standard deviations of these performance scores against which to compare changes in means are as follows: HOTS, 7; CORE, 9; MOS1, 6; and MOS2, 22.

b. Absolute change averaged over all duty areas.

Table 8. Change in sample statistics due to adding imputed values: MOS 0369 (unit leader)

Performance				Correlation with			
score	N	Mean[a]	SD	HOTS	CORE	MOS1	GT
HOTS	+80	.0	.2	-	.01	.00	.04
CORE	+80	-.1	.2	.01	-	.02	.04
MOS1	+80	.1	.0	.00	.02	-	.06
Average[b] over 12 duty areas	+80	.4	.3	.01	.01	.2	.03

NOTE: Change reflects differences in sample statistics after imputation at the step level versus at the duty area level. The results at the step level serve as the base.

a. The original standard deviations of these performance scores against which to compare changes in means are as follows: HOTS, 10; CORE, 10; and MOS1, 12.

b. Absolute change averaged over all duty areas.

expected because the imputation was not based on a "substitution of the mean" process. Intercorrelation among the core infantry content (CORE), primary (MOS1), and secondary (MOS2) scores were also relatively unchanged. The larger changes in intercorrelations tended to involve MOS2, which was a shorter block of test content and therefore less reliable. These same change statistics were computed for all duty areas of the MOSs but were based on absolute change. Again, imputation did not severely affect the sample statistics. Means and standard deviations are reported for each duty area in appendix B at both the step and duty area level of imputation.

Across the five MOSs, the validity of the performance scores versus the GT aptitude composite was differentially affected, but again changes were insignificant. Validities dropped slightly for the rifleman MOS, whereas they improved for the unit leader MOS. Figures 9 through 13 illustrate the change in validities by showing the scatterplots for data points noted as complete data versus imputed data. Note that imputation occurred across all points of the aptitude scales; in fact, imputation even resulted in some outlying cases. Thus, imputation was independent of aptitude (i.e., the frequency of missing data was similar for high- and low-aptitude persons).

CONCLUSIONS

Relatively few unusual response patterns were found in the written tests. Given the number of test forms, it was not surprising that some mistakes were made in coding answer sheets. The other aberrant data cases tended to be outliers and their deletion generally improved sample correlations and reduced standard deviations. The criteria established to identify aberrant response patterns were specifically chosen to be conservative. Although arguments could be made for different criteria, given the verification across different information sources (personal biserial correlation, percent correct score, residual analysis, problem logs, and self-report of effort), it was believed that few persons were misidentified as having aberrant patterns when, in fact, the test score was a reasonable estimate of their ability.

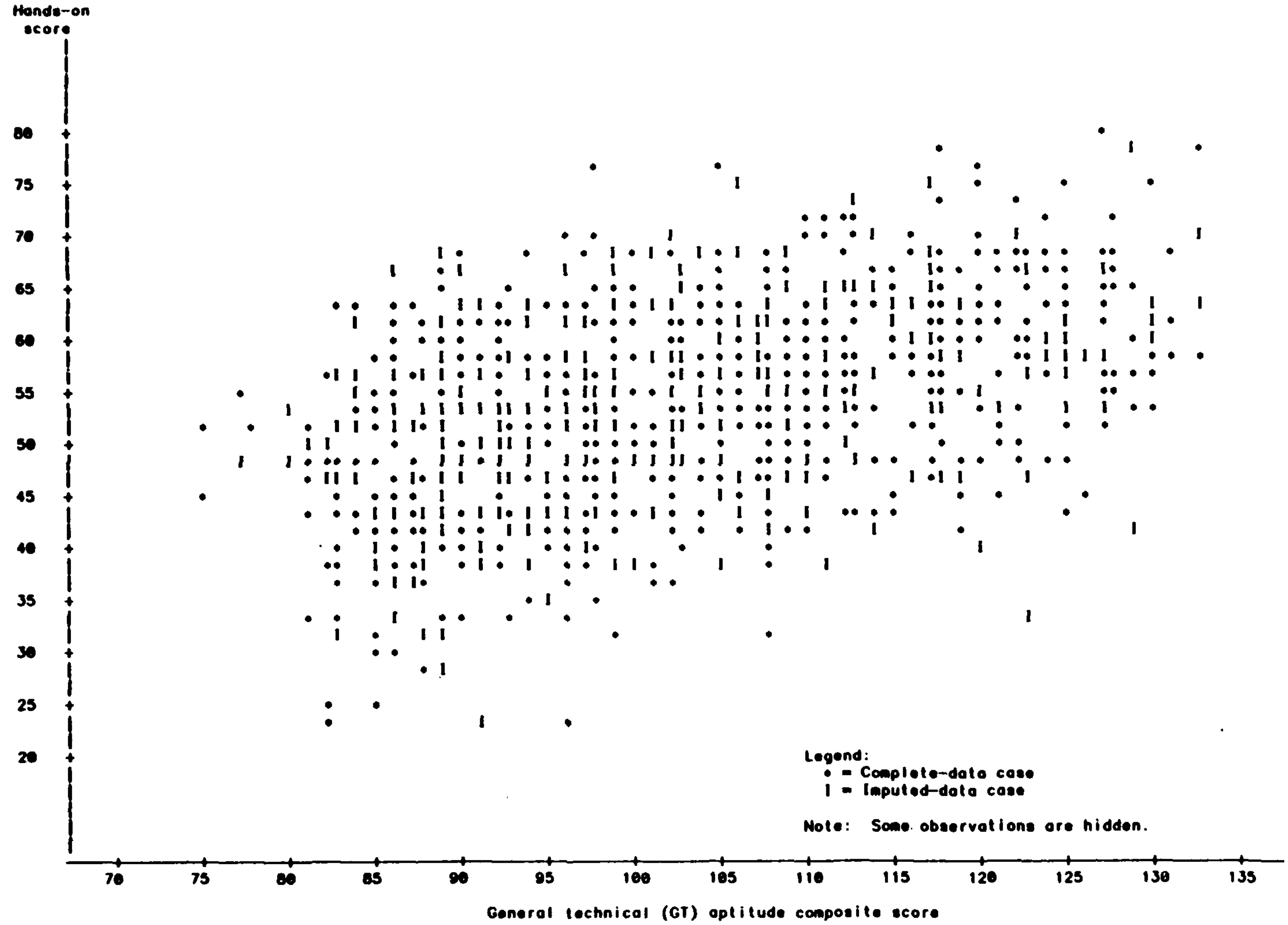


Figure 9. Validity of hands-on total score versus general technical aptitude composite score for both imputed and complete data: MOS 0311 (rifleman)

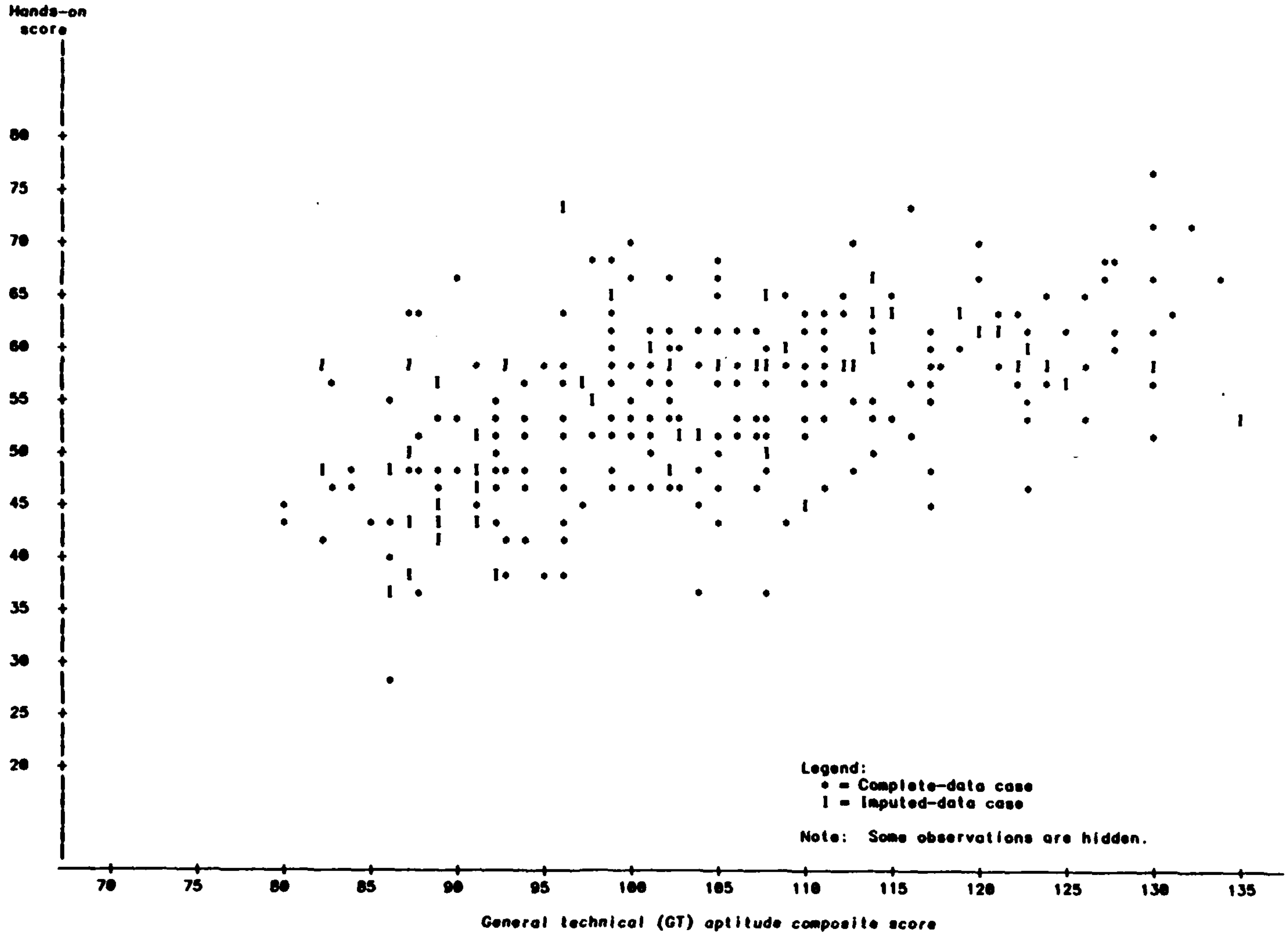


Figure 10. Validity of hands-on total score versus general technical aptitude composite score for both imputed and complete data: *MOS 0331 (machinegunner)*

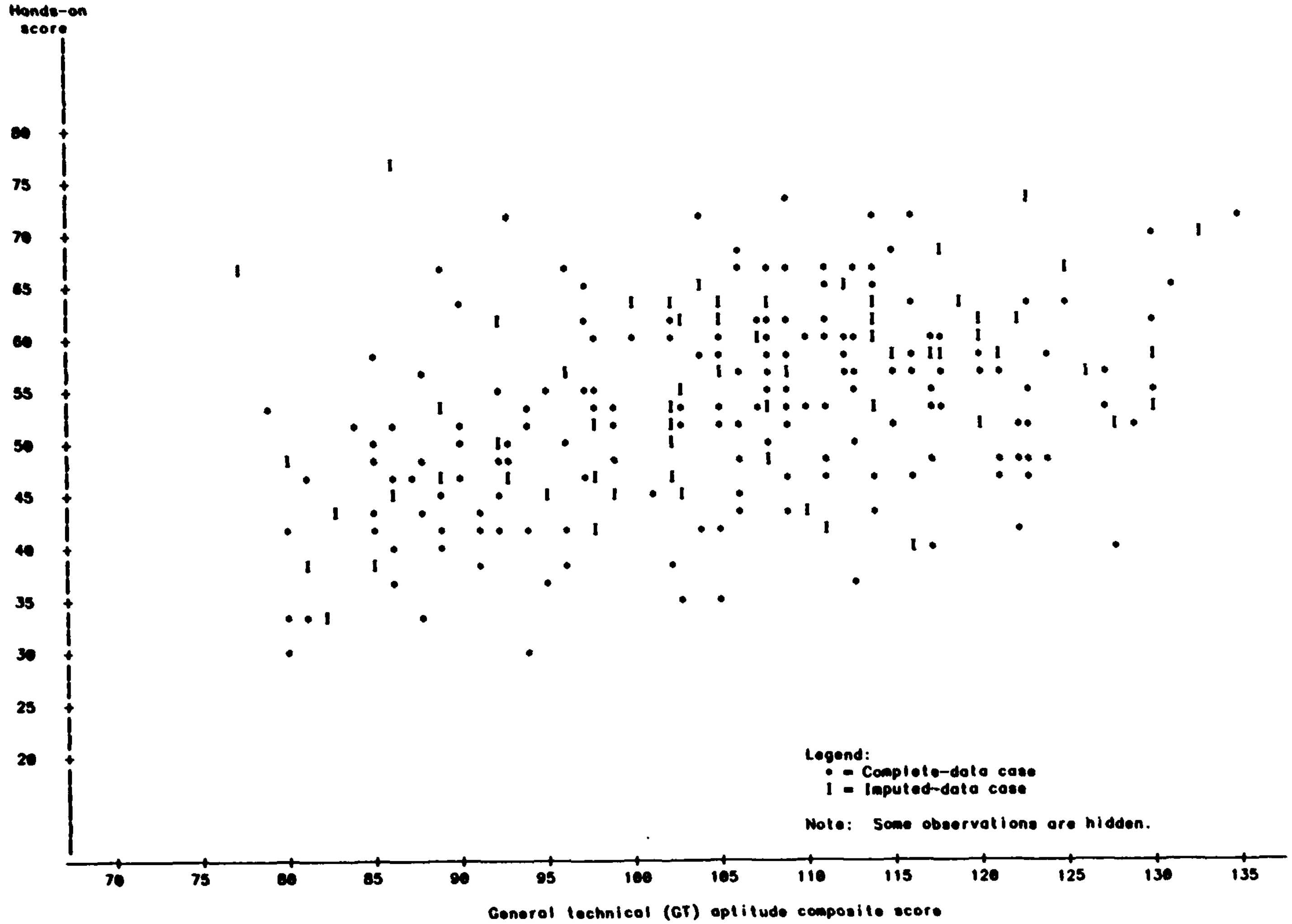


Figure 11. Validity of hands-on total score versus general technical aptitude composite score for both imputed and complete data: MOS 0341 (mortarman)

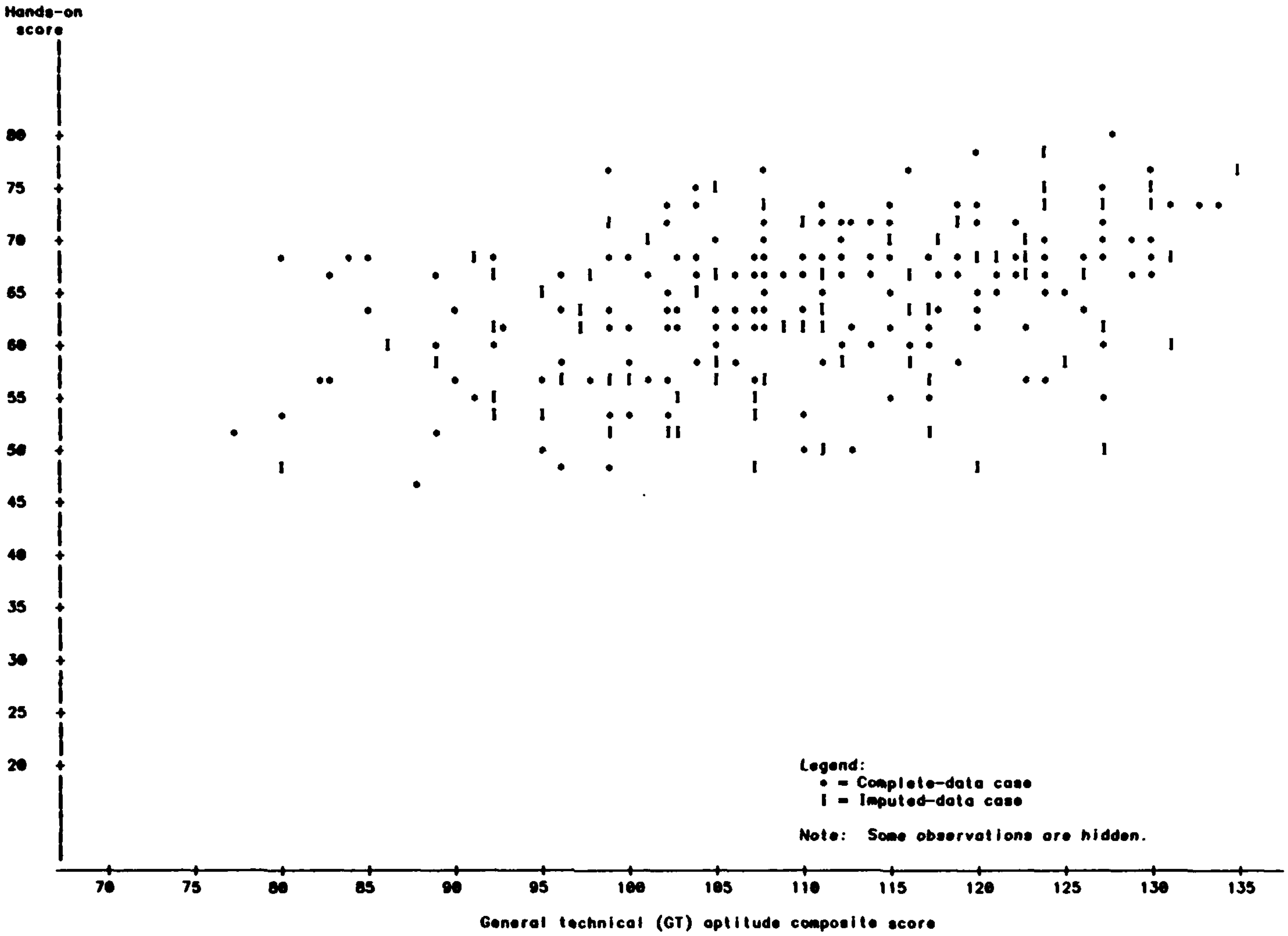


Figure 12. Validity of hands-on total score versus general technical aptitude composite score for both imputed and complete data: MOS 0351 (assaultman)

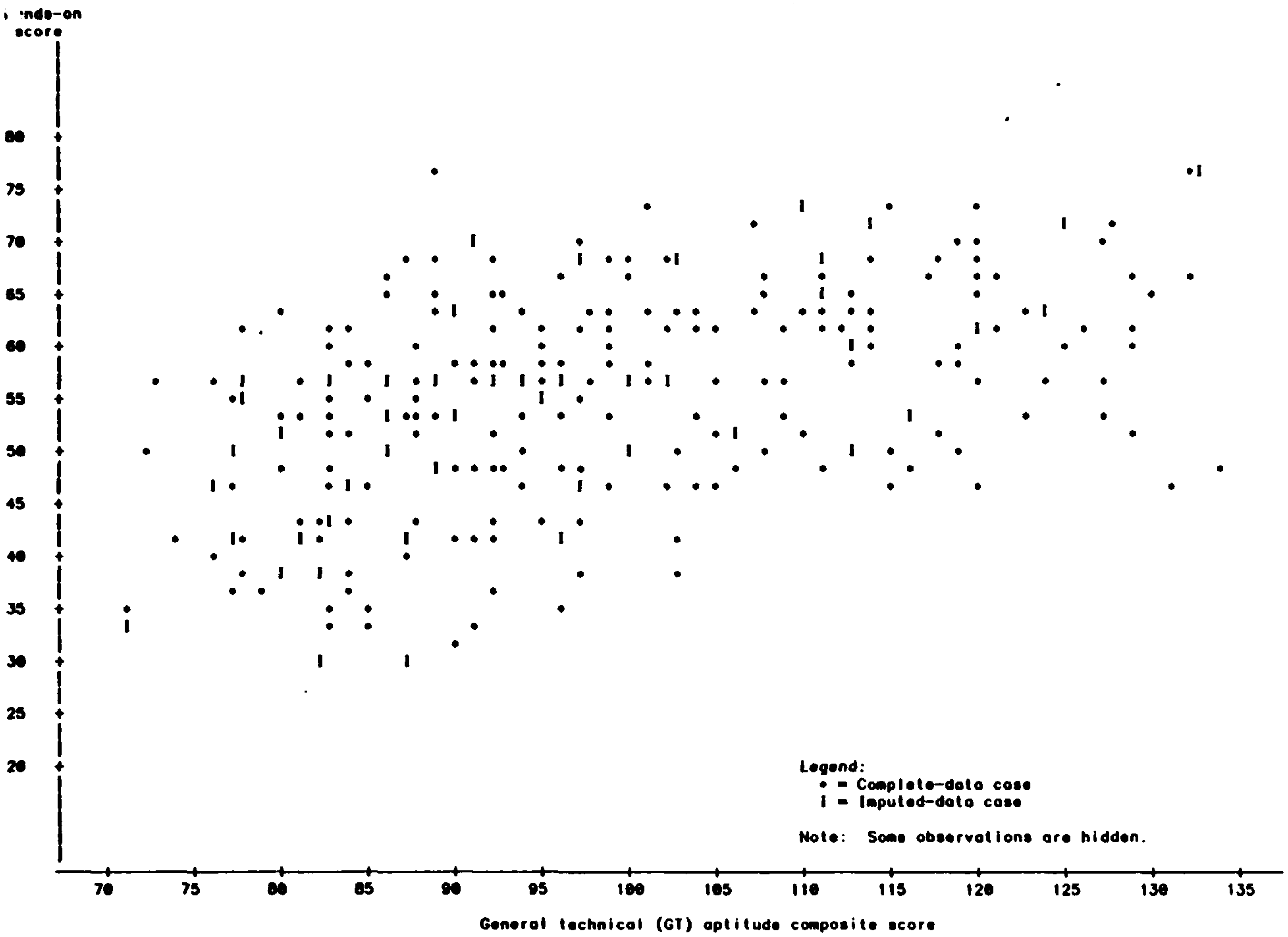


Figure 13. Validity of hands-on total score versus general technical aptitude composite score for both imputed and complete data: MOS 0369 (unit leader)

Imputation of missing data was required, in varying degrees, for over 90 percent of the examinees. The technique is intuitively straightforward, although statistically complex. Decisions were required that defined the circumstances in which sufficient data were available to warrant the imputation of missing data. Again, conservative ranges were established to mark the level of acceptable data for imputation: less than 15 percent missing steps, less than 20 percent missing tasks, and less than 20 percent missing duty areas. Sample statistics were insignificantly affected by imputation. Indeed, this was the outcome that was sought by employing an imputation procedure that incorporated procedures to minimize the impact of imputed values.

As a result of these data quality analyses that identified unusual response patterns and imputed missing data for the infantry JPM data, further analytic investigations can proceed with confidence in the soundness of the data and the integrity of the results.

REFERENCES

[1] American Institutes for Research, AIR-47500-FR, *Developing Job Performance Tests for the United States Marine Corps Infantry Occupational Field*, 29 Sep 1988

[2] T. F. Donlon and F. E. Fischer. "An Index of an Individual's Agreement Group-Determined Item Difficulties." *Educational and Psychological Measurement* 28 (1968): 105-113

[3] B. F. Green and H. Wing, eds. *Analysis of Job Performance Measurement Data: Report of a Workshop.* Washington, DC: National Academy Press, 1988

[4] L. L. Wise and D. McLaughlin. *Guidebook for the Imputation of Missing Data.* Palo Alto, CA: American Institutes for Research, 1980

APPENDIX A

COMPUTATIONAL PROCEDURES FOR IDENTIFICATION OF INCONSISTENT RESPONSE PATTERNS AND IMPUTATION OF MISSING DATA

APPENDIX A

COMPUTATIONAL PROCEDURES FOR IDENTIFICATION OF INCONSISTENT RESPONSE PATTERNS AND IMPUTATION OF MISSING DATA

COMPUTATION OF PERSONAL BISERIAL CORRELATION

The personal biserial correlation (r_{perbis}) was proposed by Donlon and Fischer [A-1] as a heuristic means of evaluating the appropriateness of a person's total score in measuring his or her ability. The approach is heuristic in that no assumptions or theories are made concerning a person's underlying ability; rather, determinations of appropriateness are made relative to the responses of a reference sample. The r_{perbis} statistic quantifies the similarity between the item difficulties as experienced by a particular examinee relative to the item difficulties computed for a reference sample.

The r_{perbis} statistic requires two basic assumptions. First, there is a latent variable that underlies a person's observed item responses and this variable is normally distributed across items. If the magnitude of this latent variable is greater than some threshold, the examinee responds correctly to the item; otherwise, the item is incorrectly answered. Excessive guessing by examinees for any item invalidates this assumption. The second assumption requires a linear regression of item difficulties experienced by the reference sample onto the item difficulties experienced by a particular examinee. In other words, the relative ordering of items with respect to difficulty is similar for both the individual examinee and the reference sample.

Given these assumptions, r_{perbis} can be computed as the biserial correlation between the examinee's pattern of item responses (1s and 0s) and the item difficulties in the reference sample. (This is the transpose of the computations required for an item-total correlation.) However, Donlon and Fischer first transformed the item difficulty statistics because they tend not to be normally distributed:

$$\Delta_i = 4\Phi^{-1}(1 - \hat{p}_i) + 13, \qquad (A-1)$$

where

Δ_i = the transformed item difficulty for item i

Φ^{-1} = a probit transformation

$\hat{p}_i$ = item difficulty statistic – proportion correct – for item i.

This Δ_i is more normally distributed than the original item difficulties and has a mean of 13 and standard deviation of 4. The r_{perbis} is then computed for each examinee as:

$$r_{perbis} = \frac{\bar{\Delta}_r - \bar{\Delta}_c}{s_\Delta} \frac{k}{h}, \qquad (A-2)$$

where

$\bar{\Delta}_r$ = the mean Δ for items reached by an examinee

$\bar{\Delta}_c$ = the mean Δ for all items correctly answered

s_Δ = the standard deviation of the Δs across all items reached

k = the number of items correctly answered divided by the number of items reached

h = the height of the standard normal curve at the point dividing the area under the curve into sections with areas k and (1-k).

As stated in the text, r_{perbis} ranges from -1 to 1, with negative and low values representing negative or inconsistent relationships between an examinee's set of responses and the item difficulties experienced by the reference sample. Caution should be used in interpreting r_{perbis} because it is a heuristic statistic. Without a specific theory of measurement, it is difficult to characterize the properties of normal response patterns and, therefore, difficult to definitively determine inconsistent response patterns.

IMPUTE PROCEDURE FOR ESTIMATING MISSING DATA

Imputation procedures for the estimation of incomplete data can be divided into four basic types. Each type is briefly reviewed and the inherent problems associated with each are discussed. The IMPUTE procedure that was used in this research memorandum [A-2] is described within the context of other imputation procedures. Particular attention is given to the assumptions of the IMPUTE procedure and further details concerning its computations are provided.

The first type of imputation procedure makes use of only summary-level data in the estimation of missing values. These procedures typically compute means based on complete data and then substitute these values for all missing cases. For example, an examinee's mean performance on all available tasks can be substituted for any task that has a missing value. However, tasks differ in their difficulty of performance. Therefore, substitution of an examinee's average task score for any missing task introduces systematic bias to the extent that the missing task differs in difficulty from the average task difficulty. Conversely, the mean could be computed over all examinees but separately for each task to account for task difficulty. This technique also introduces systematic error by not recognizing differences in individual performance. If missing data points are few and the intended use of the data set is simply to estimate population means or totals, such summary-level substitution procedures may suffice.

Weighting methods are another means of "imputing" missing data. Missing values are implicitly accounted for by increasing the weights assigned to similar cases that have complete data. This technique is primarily employed in the survey research community and assumes that nonresponse cases (incomplete data) are consistent with the persons who did respond. (This assumption is required of all imputation procedures.)

The third type of imputation procedure can be called single-iteration imputation because explicit values are determined at the individual level fcr all missing values based on a single manipulation of the data. Three specific techniques fall within this category. First, regression-based imputation procedures can simultaneously account for multiple predictors in the estimation of a missing value (e.g., task difficulty and individual differences). However, regression procedures may distort the distributions of variables and thereby bias variance and covariance statistics so that imputed values become overly correlated with the predictor variables from which they were imputed. Regression procedures may also result in values that are outside the range of actual observed values. A second procedure, called a "hot deck" estimate, limits imputed values to the observed range. Although hot-deck procedures maintain the distributional characteristics of variables, the whole case is replaced, not just the missing values for specific variables. The final technique typifies the IMPUTE procedure in that the missing values are distributional estimates – responses are randomly assigned from an appropriately generated distribution of estimates. In this manner, the IMPUTE procedure is an *extension* of the regression procedures but preserves the multivariate distributions of variables and thereby accurately reproduces means, variances, and covariances.

The final category of imputation procedures is extremely computation laden – multiple iteration imputation. Missing values are imputed multiple times based on different random numbers to generate multiple data sets. Complete-data analyses are conducted for each data set, and the variance in the results provides an estimate of the error due to imputation. Such analyses appear to be excessive in the context of hands-on job performance measurement and the validation of the Armed Service Vocational Aptitude Battery (ASVAB).

Computations Required for IMPUTE

The initial step in the IMPUTE procedure computes basic descriptive statistics – mean, standard deviation, minimum, maximum, and number of missing values for each variable. Intercorrelations among the variables are also computed based on all pairwise combinations of the variables; again, missing variables within each pair are noted. The variables are then ordered on the basis of their magnitude of missing data and relative intercorrelations with other variables. A stepwise regression is computed for the first variable in this ordered list that has missing data. The regression uses all prior variables in the list as predictors and stops when no further variables contribute to the prediction of the variable being imputed. Based on this regression, expectancy tables are constructed relating actual values to the predicted regression values. If the imputed variable is discrete, the predicted regression values are categorized into the discrete intervals of the criterion. If the imputed variable is continuous, the regressed values are categorized so that each interval contains a sufficient number of subjects. (The continuous scale of the criterion is regenerated once an imputed value is determined by interpolation between the means of the regressed predicted values for adjacent categories.) Table A-1 presents a hypothetical expectancy table for a discrete variable (e.g., a rating scale with values ranging from 1 to 5).

Table A-1. Expectancy table relating actual values to predicted regression values

Predicted regression	Percentage of actual rating values at each predicted regression value				
value	1	2	3	4	5
1	50	45	5		
2	15	65	20		
3		30	40	30	
4		5	20	60	15
5			20	25	55

For each missing value, a predicted value is generated using the regression function, and then an "actual" value is selected randomly with probability proportional to the percentages of the expectancy table. Such a procedure yields only values that actually occurred and ensures an appropriate variation of the imputed values.

Each variable from the ordered list is processed in turn. Those variables that have imputed values are considered as potential predictors in the later stepwise regressions. Once all missing values have been imputed, a second stage of imputation is conducted to determine if any variables in the later part of the ordered list would have been significant predictors of previous variables requiring imputation. If so, the procedure is repeated for that particular variable and new imputed values are computed. In this way, any significant relationships between variables with missing values are preserved because each is used in the prediction of the other. Although it may appear that using imputed values to impute other values only builds error on error, such redundancy is necessary to reproduce the multivariate structure of a data set. A much more complete description of the imputation procedures is provided in [A-2].

Assumptions for IMPUTE Procedure

The primary assumption of the IMPUTE procedure is that persons with incomplete data are thought to be similar to those with complete data (once any particular differences have been controlled for). This assumption requires a constant relationship between variables regardless of group membership (those with complete information versus those with missing data). This assumption cannot be validated directly because data are not available for the incomplete data group. An indirect validation of this assumption can be obtained by examining the consistency of relationships across various groups of the data set. Such comparisons have been conducted in the context of other JPM analyses with respect to groups defined by pay grade, time in service, and other demographic variables. Consistency of the relationships between duty areas was found for these designated groups (such analyses have not been conducted at the task or step level).

Additional assumptions are also required because regression procedures are used in the IMPUTE procedure. These are the typical regression requirements for linearity and errors - expected value of zero, uncorrelated errors, homoscedasticity, and errors uncorrelated with the predictors.

REFERENCES

[A-1] T. F. Donlon and F. E. Fischer. "An Index of an Individual's Agreement Group-Determined Item Difficulties." *Educational and Psychological Measurement* 28 (1968): 105-113

[A-2] L. L. Wise and D. McLaughlin. *Guidebook for the Imputation of Missing Data.* Palo Alto, CA: American Institutes for Research, 1980

APPENDIX B

SAMPLE STATISTICS AFTER IMPUTING AT STEP AND DUTY AREA LEVELS

APPENDIX B

SAMPLE STATISTICS AFTER IMPUTING AT STEP AND DUTY AREA LEVELS

The tables of this appendix detail the means and standard deviations resulting from imputation at the step and duty area levels. The change statistics reported in tables 4 through 8 were based on these values. Also, the statistics are given for each duty area within each military occupational specialty (MOS). These numbers were only summarized in the tables in the text. The standard deviations of the duty areas will help in interpreting the magnitude of the change in duty area means.

Abbreviations used in tables B-1 through B-5 are defined below:

HOTS	hands-on total score
HOCORE	hands-on core content score
MOS1	primary MOS score
MOS2	secondary MOS score
CR	communications
FA	first aid
GL	grenade launcher
HG	hand grenade
LA	light antitank weapon
LN	land navigation
MI	mines
NB	nuclear, biological, chemical defense
NV	night vision device
SI	security and intelligence

TM	tactical measures
SH	squad automatic weapon
RF	M16A2 rifle firing
MG	machine gun
MO	mortar
DR	dragon
SM	shoulder-mounted assault weapon
OP	operations order

Table B-1. Sample statistics after imputing at step and duty area levels: MOS 0311 (rifleman)

Content area	Level of imputation			
	Step		Duty area	
	Mean	SD	Mean	SD
HOTS	54.2	9.0	54.0	9.1
CORE	56.7	9.4	56.4	9.5
MOS1	54.1	16.6	54.1	16.7
MOS2	44.5	16.5	44.4	16.6
CR	57.0	13.8	56.8	13.8
FA	48.3	17.2	48.1	17.1
GL	54.5	9.1	54.2	9.2
HG	52.4	19.9	52.3	19.6
LA	57.5	23.1	56.6	23.1
LN	50.8	23.6	50.1	23.9
MI	35.7	28.0	36.2	28.0
NB	57.3	14.1	57.3	14.1
NV	64.4	25.1	63.4	24.9
SI	61.0	18.8	61.1	19.0
TM	53.7	11.1	53.4	11.1
SH	48.2	16.4	48.9	16.7
RF	54.9	25.9	55.7	25.5

Table B-2. Sample statistics after imputing at step and duty area levels: MOS 0331 (mortarman)

Content	Level of imputation			
	Step		Duty area	
area	Mean	SD	Mean	SD
HOTS	55.2	7.8	55.1	7.9
HOCORE	54.1	9.1	54.2	9.3
MOS1	60.1	9.7	60.0	10.3
MOS2	50.4	15.8	49.9	15.8
CR	52.6	14.2	52.8	14.4
FA	50.0	16.2	50.1	16.0
GL	52.5	7.7	51.9	7.7
HG	51.6	20.2	51.5	19.8
LA	49.9	23.2	49.4	22.8
LN	46.6	22.7	46.5	23.6
MI	38.1	26.6	38.7	26.6
NB	55.9	14.4	56.8	14.6
NV	74.9	19.3	74.6	19.3
SL	60.0	18.9	59.9	19.2
SI	54.5	20.5	55.8	20.2
TM	51.6	10.8	52.0	10.7
MG	61.0	9.7	60.0	10.3
RF	37.2	26.6	37.6	27.2

Table B-3. Sample statistics after imputing at step and duty area levels: MOS 0341 (machinegunner)

Content area	Level of imputation			
	Step		Duty area	
	Mean	SD	Mean	SD
HOTS	53.1	8.9	53.7	9.1
HOCORE	53.8	9.4	54.3	9.5
MOS1	55.1	13.9	55.7	14.2
MOS2	46.1	16.3	47.0	16.4
CR	59.0	12.7	59.6	12.6
FA	47.9	15.5	47.3	15.3
GL	54.7	9.3	55.1	10.4
HG	49.5	18.8	50.0	18.7
LA	55.4	24.7	55.3	25.8
LN	49.5	23.2	50.4	23.1
MI	40.4	24.1	39.9	23.4
NB	56.3	15.1	55.9	14.9
NV	57.0	27.5	57.0	27.9
SL	49.6	27.0	51.3	26.6
SI	60.2	18.6	59.6	18.2
TM	52.0	10.3	52.1	10.5
MO	55.1	13.9	55.7	14.2
RF	43.2	26.6	42.7	26.6

Table B-4. Sample statistics after imputing at step and duty area levels: MOS 0351 (assaultman)

Content	Level of imputation			
	Step		Duty area	
area	Mean	SD	Mean	SD
HOTS	64.6	6.4	64.2	6.8
HOCORE	59.8	8.4	59.4	8.6
MOS1	79.2	5.8	79.1	5.9
MOS2	59.7	21.7	58.4	22.7
CR	63.4	12.1	62.5	12.5
FA	52.1	13.9	52.0	14.3
GL	55.8	10.4	56.0	9.9
HG	48.9	18.9	48.4	18.9
LA	71.4	18.4	72.0	17.6
LN	56.5	22.7	55.9	23.0
MI	48.3	27.0	50.5	26.5
NB	61.7	13.7	61.7	13.7
NV	66.4	24.7	67.2	24.8
SL	61.9	20.1	59.1	22.9
SI	62.9	18.4	62.4	18.6
TM	54.4	10.7	54.2	10.8
SM	59.7	21.7	58.4	22.7
DR	50.4	14.4	50.0	14.6

Table B-5. Sample statistics after imputing at step and duty area levels: MOS 0369 (unit leader)

Content	Level of imputation			
	Step		Duty area	
area	Mean	SD	Mean	SD
HOTS	55.1	9.4	55.1	9.6
HOCORE	61.8	9.5	61.7	9.7
MOS1	45.2	12.0	45.3	12.0
CR	61.8	11.5	61.6	12.0
FA	55.2	16.8	54.6	17.2
GL	59.4	9.1	59.6	9.1
HG	49.8	18.2	51.0	18.5
LA	62.4	19.7	62.5	20.0
LN	69.9	22.3	70.0	22.4
MI	38.7	22.3	38.2	22.0
NB	63.7	14.4	63.1	14.6
NV	69.4	25.3	70.6	25.2
SI	70.9	17.4	70.5	18.2
TM	57.1	10.4	57.0	10.7
SH	43.2	16.7	43.9	17.2