AD NUMBER
ADA263680
LIMITATION CHANGES
TO: **Approved for public release; distribution is unlimited.**
FROM: **Distribution authorized to DoD only; Specific Authority; AUG 1991. Other requests shall be referred to Marine Corps Combat Development Command, WF-13, Quantico, VA.**
AUTHORITY
usmc dtd oct 2003

CRM 91-120 / August 1991

Data Quality for the Automotive Maintenance Phase of the Marine Corps Job Performance Measurement (JPM) Project

Neil B. Carey
Catherine M. Hiatt

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

CNA CENTER FOR NAVAL ANALYSES

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268 • (703) 824-2000

26 August 1991

MEMORANDUM FOR DISTRIBUTION LIST

Subj: Center for Naval Analyses Research Memorandum 91-120

Encl: (1) CNA Research Memorandum 91-120, *Data Quality for the Automotive Maintenance Phase of the Marine Corps Job Performance Measurement (JPM) Project*, by Neil B. Carey and Catherine M. Hiatt, Aug 1991

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 automotive maintenance portion of the Marine Corps Job Performance Measurement Project, then describes measures taken to minimize the effect of questionable or missing cases. Particular attention is focused on data inconsistencies and problems associated with operational fielding of the Computer Adaptive Test-Armed Services Vocational Aptitude Battery (CAT-ASVAB) and the Enhanced Computer Administered Test (ECAT).

Lewis R. Cabe
Director
Manpower and Training Program

Distribution List:
Reverse page

Subj: Center for Naval Analyses Research Memorandum 91-120

Distribution List

SNDL

45A2 CG I MEF
45A2 CG II MEF
45A2 CG III MEF
45B CG FIRST MARDIV
45B CG SECOND MARDIV
A1 DASN - MANPOWER (2 copies)
A1H ASSTSECNAV MRA
A2A CNR
A4 PERS-11
A5 PERS-2
A5 PERS-5
A6 CG MCRDAC - WASHINGTON
A6 HQMC AVN
A6 HQMC MPR & RA
Attn: Code M
Attn: Code MR
Attn: Code MP
Attn: Code MM
Attn: Code MA (3 copies)
Attn: Code MPP-54
FF38 USNA
Attn: Nimitz Library
FF42 NAVPGSCOL
FF44 NAVWARCOL
FJA1 COMNAVMILPERSCOM
FJA13 NAVPERSRANDCEN
Attn: Technical Director (Code 01)
Attn: Technical Library
Attn: Dir, Manpower Systems (Code 11)
Attn: Dir, Personnel Systems (Code 12)
Attn: Dir, Testing Systems (Code 13)
Attn: CAT/ASVAB PMO
FJB1 COMNAVCRUITCOM
FT1 CNET
V8 CG MCRD PARRIS ISLAND
V8 CG MCRD SAN DIEGO
V12 CG MAGTEC
V12 CG MCCDC
Attn: Studies and Analyses Branch
Attn: Director, Warfighting Center
Attn: Warfighting Center, MAGTF Proponency and Requirements Branch (2 copies)
V12 CG MCRDAC - QUANTICO
V25 MCAGCC

OPNAV

OP-01

CRM 91-120 / August 1991

Data Quality for the Automotive Maintenance Phase of the Marine Corps Job Performance Measurement (JPM) Project

Neil B. Carey
Catherine M. Hiatt

Operations and Support Division

CNA
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 automotive maintenance portion of the Marine Corps Job Performance Measurement Project, then describes measures taken to minimize the effect of questionable or missing cases. Particular attention is focused on data inconsistencies and problems associated with operational fielding of the Computer Adaptive Test-Armed Services Vocational Aptitude Battery (CAT-ASVAB) and the Enhanced Computer Administered Test (ECAT).

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. For the automotive maintenance phase, more than 1,000 organizational automotive mechanics (MOS 3521) were tested for two days each on a variety of performance measures. Besides taking a eight-hour hands-on test of mechanical performance, examinees took a paper-and-pencil job knowledge test, performed certain portions of the General Aptitude Test Battery (GATB), took the ASVAB by computer (CAT-ASVAB), and were administered several new computerized predictors of job performance (ECAT).

The volume of data was enormous. Many problems could potentially affect the quality or completeness of data. Many precautions were taken to minimize the possibility of poor or missing data--test administrators were extensively trained, data were checked for completeness every day, equipment was checked every day, and the consistency of responses was also monitored daily. Examinees were briefed on the importance of giving their full effort. Despite these precautions, there were still individual cases in which the accuracy of the data was questionable and other instances in which the data were incomplete.

This memorandum quantifies the amount of questionable or incomplete data, then details the procedures used to minimize the effect such data would have on later analyses.

IDENTIFICATION OF UNUSUALLY LOW SCORES

Occasionally, a test may fail to measure properly 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 condition occurred that produced unusually low scores (e.g., lack of motivation, illness, lack of sleep, inattentive marking of the answer sheet, random responses, application of the wrong answer key).

To identify unusually low scores, it is necessary to compare scores to those achieved when the individual was tested under motivated conditions. Enlistment ASVAB scores reflect performance when highly motivated, since enlistment scores determine eligibility for the Marine Corps. Persons whose job knowledge test scores were far below what would be expected on the basis of their enlistment ASVAB scores were assumed to have given less than their full effort on the job knowledge test.

Decision rules were established for the identification of unusually low scores based on the prediction of job knowledge scores from enlistment ASVAB and time in service. Given these criteria, eight scores were declared aberrant for the job knowledge test (JKT). Deleting these aberrant scores increased the means for the test and

decreased the standard deviation. The correlation of the JKT with the enlisted mechanical maintenance (MM) aptitude composite scores increased slightly. These changes in sample statistics indicated that the deleted scores were typically outlier cases.

To check the accuracy of the CAT-ASVAB scores, CAT-ASVAB MM scores were predicted from enlistment MM scores. As with the job knowledge test, persons whose CAT-ASVAB scores were far below what would be expected on the basis of their enlistment scores were assumed to have given less than their full effort on the CAT-ASVAB. In such cases, the CAT-ASVAB score is not reflective of the individual's true aptitude.

The CAT-ASVAB scores of those with extremely low CAT-MM scores relative to their enlisted MM were deleted. Thirteen scores were deleted for this reason. Typically, the deleted scores were the lowest possible value, which indicated that deleted scores reflected random guessing on CAT-ASVAB items.

IMPUTATION OF MISSING DATA

Hands-on performance data were collected at the step level; a person either passed or failed to perform a specific action. Steps were aggregated to form task scores. It was not always possible to collect complete information for each person--there were 391 steps for the hands-on test. Examinees could have incomplete data as a result of weather conditions, equipment failure or unavailability, being called away before completion, or performing a step that was unobservable to the test administrator.

Despite the many ways data could be missing, very few data were missing. Overall, complete data were collected for 96 percent of all tasks administered.

Imputation is the process of estimating the score that would have occurred if circumstances had not prevented actual scoring. Imputation was performed in order to make fullest use of the data. Data were imputed at the step level. Sample statistics for all variables with complete information before the step-level imputation were compared to the sample statistics after imputation. The shifts in mean performance scores were relatively small compared to the standard deviation of performance scores. Correlations of performance with aptitude changed insignificantly as a result of imputation.

PROBLEM LOGS

Problem logs, maintained by field data collectors, recorded instances of difficulty in collecting or maintaining quality of data. Problem logs were kept for the job knowledge test, CAT-ASVAB, ECAT, GATB, and two "administrative duties" tests. Field data collectors' comments noted the reasons that data were lost or questionable due to

lack of effort from examinees or situational disruptions. Considering that data were collected over a period of four and a half months in two different locations, the logs indicated relatively few problems.

There were very few missing cases for any of the data sources covered by the problem logs. The small number of problem log entries was confirmed by inspection of the actual data: The actual amount of fully missing data ranged from a high of 4.9 percent of cases for the ECAT to a low of 0.1 percent for the job knowledge test. Nevertheless, the problem logs showed that each data source presented characteristic challenges to field data collection: Examinees sometimes hurried through the job knowledge test; they ran out of time for the administrative duties test; prior hand injuries occasionally prevented completion of the manual dexterity portion of the GATB; response pedestal problems periodically hampered the ECAT; and disk failures sometimes obstructed CAT-ASVAB data collection.

SUMMARY

Relatively few unusually low scores were observed for the JKT test. The aberrant data cases were outliers so that their deletion generally improved sample correlations and reduced standard deviations. The criteria for identifying unusually low scores were specifically chosen to be conservative. Specific deletions were confirmed against other information whenever possible. Less than 1 percent of job knowledge test scores were deleted as a result of these procedures. Given the verification across different information sources (residual analysis, percent-correct score, problem logs, personal biserial correlation), few if any persons should have been misidentified as having aberrant scores when, in fact, the test score was a reasonable estimate of their ability.

Twenty-one hands-on tasks composed of 391 steps were administered to 1,028 Marines. Overall, complete data were collected for 96 percent of all tasks. For the tasks with at least one missing step, an average of four steps were imputed to achieve complete task-level information. All cases were deemed recoverable by imputation of missing data. 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 automotive maintenance JPM data, further analytic investigations can proceed with confidence in the soundness of the data and the integrity of the results.

CONTENTS

ILLUSTRATIONS

TABLES

INTRODUCTION

The Job Performance Measurement (JPM) Project is a long-term effort to validate the Armed Services Vocational Aptitude Battery (ASVAB) against hands-on measures of job performance. The Automotive Mechanical Maintenance phase of the project tested more than 1,000 mechanics in two test sites. Each mechanic was required to complete 21 hands-on mechanical tasks on five different vehicles. Properly implemented hands-on tests are very resource intensive: Test administrators must individually observe and score performance of job tasks, so administration is considerably costlier than paper-and-pencil testing. Test administrators must have extensive training and be given frequent feedback on their performance judgments. Vehicles must be restored to proper condition before the next mechanic performs a task. Because hands-on tests require a well-organized flow of examinees, hands-on testing requires considerable attention to the assignment and transportation of personnel.

The JPM project also addressed other manpower research issues, such as (1) whether less expensive "surrogate" measures could be used in place of hands-on tests, and (2) whether "new predictors" could enhance the predictive power of the ASVAB. Therefore, examinees were administered a paper-and-pencil test of automotive job knowledge, were administered the ASVAB by computer, and were given a series of computerized "new predictors" designed to enhance the ability of the ASVAB to predict performance. Portions of the General Aptitude Test Battery (GATB) that measure manual dexterity were also administered as part of this project.

The volume of data was enormous. Many problems could affect the quality or completeness of data. Test equipment could break; test sites' equipment setup could differ; a test administrator could fail to see whether a step was completed; or an examinee could fall ill or be called away. Even if these problems did not occur, other situations could affect the quality of data. Examinees could rush through the test, failing to give their full effort; computers, disks, or information networks could fail; the wrong test form could be given; or unavoidable distractions could upset testing.

Because of the many problems that could beset large-scale data collection efforts, significant precautions were taken to minimize the possibility of poor and/or missing data [1]. Preliminary tryout testing and a command review were conducted for all tasks on the performance test. Data quality was affected by the extent to which Marines were motivated to attempt all performance measures seriously. To ensure full effort, each participant was given an illustrated pamphlet that explained the importance of the study to the future of the Marine Corps. Immediately before testing began, a Marine Corps officer gave a short talk that emphasized the importance of giving one's full effort.

Continuous monitoring during the testing identified potential problems so that they could be corrected as soon as possible. Data quality was monitored daily through verification[1] of answer sheets, daily entry of all hands-on responses, and maintenance of problem logs to identify specific problem cases. To encourage effort in taking the CAT-ASVAB, Marines were informed that their scores of record would be changed to their CAT-ASVAB scores if their CAT-ASVAB scores exceeded their scores of record. This could have significant payoff for persons who wanted to transfer to other occupational fields with higher aptitude requirements. In research on the earlier infantry phase of this study, this incentive appeared to be effective [2]. No changes would be made if the CAT-ASVAB scores were lower than the scores of record.

Precautions were also taken to minimize the amount of missing data. A Marine Corps technician was available each day so that mechanical problems and parts failures could be dealt with promptly. Test administrators were instructed in ways to observe all steps being performed. Data were reviewed daily for completeness, and examinees were scheduled to retake any portions of the test that they missed.

Despite these initial tryouts and quality-control procedures, there were still individual cases in which the accuracy of the data was questionable, and other cases in which the data were simply incomplete. Both of these factors affect overall data quality and can affect analyses yet to be conducted on JPM data. For example, Maier [3] has found that data quality-control procedures can make large differences in the computed validity coefficients.

To identify data inaccuracies at the individual level, unusually low scores were compared with scores earned under motivated conditions. Enlisted ASVAB scores are reflective of motivated performance, because such scores determine one's eligibility for the Marine Corps. An unusually low score compared to one's enlisted MM score could reflect such circumstances as fatigue, illness, random guessing, or inadvertently skipping a question so that responses were always meant for the adjacent item. Such causes are unrelated to a person's ability, so they are errors for purposes of this project. Data that have such unusual responses must be declared missing. Identifying unusual response patterns applied to written tests only. For the hands-on performance tests, test administrators served to monitor unusual response patterns.

An examinee might have incomplete data for a number of reasons. The Marine might be called away for an emergency, fall ill, or experience equipment breakage, or the test administrator might be unable to observe the examinee's response. These types of conditions are not under the control of the examinee, and hence are considered random.

---

1. The onsite manager verified answer sheets by reading each sheet and making corrections if necessary.

Because some data are better than none, data that are missing due to such random events can be estimated using data from other parts of the test. Specific procedures were developed for the estimation of missing data at the step level.

Given that JPM analyses, yet to be conducted, are sensitive to outliers and require complete information, this research memorandum presents the procedures used to ensure the quality and completeness of the mechanical maintenance data. Methods for identifying unusual response patterns are described, and deletion of individual cases is justified based on triangulation between different sources of information. The magnitude of missing data at the step level is presented. The impacts of the deletion of aberrant data and the imputation of missing data are documented by noting changes in the sample descriptive statistics.

As part of the data collection procedures, test administrators kept daily problem logs that noted anomalies in testing procedures. Results from the daily problem logs are analyzed to determine common problems in the collection of each type of data (e.g., CAT-ASVAB, new predictors, job knowledge test) and to make recommendations for future data collection efforts.

IDENTIFICATION OF UNUSUALLY LOW SCORES

A test can fail to measure a particular individual's ability even though the test adequately measures abilities in the group. For example, a particular individual's scores might be anomalous because of lack of effort, cheating, random guessing, or unknowingly skipping a question. Such occurrences guarantee that the test is not properly measuring the individual's abilities.

CAT-ASVAB Scores

This project obtained two sets of ASVAB scores. Enlistment scores were obtained under motivated conditions, when the individual's score affected acceptance into the Marine Corps. These scores are thus assumed to reflect accurately the individual's aptitudes. CAT-ASVAB scores, obtained during JPM testing, could be reflective of less motivation, since the examinee was already accepted into the Marines.

To check their accuracy, CAT-ASVAB mechanical maintenance (MM) scores were predicted from enlistment MM scores, using linear regression. Discrepancies between the actual CAT-ASVAB score and the score predicted from the enlistment score are called residuals. Details about linear regression and residual analysis can be found in appendix A. Large negative residuals identified persons whose CAT-ASVAB scores were not accurate indicators of their aptitude. Residuals were computed from the regression and plotted as shown in figure 1. The figure shows 11 scores, below the line, that are highly unlikely to have been the result of full effort. These CAT-ASVAB MM scores are far below what would be predicted from these individuals' enlisted MM scores. Furthermore, the plot shows that these scores would unduly distort

further analyses because they are so different from other scores. CAT-ASVAB scores more than 4 standardized residuals below what would be predicted from enlistment scores were considered indicative of poor motivation, and were dropped from further analyses. These 11 individuals scored the lowest possible score on the CAT-ASVAB. In addition, two other individuals' CAT-ASVAB scores were dropped because (a) extreme lack of motivation was noted in the problem logs and (b) their scores were more than 2.5 standardized residuals below the predicted value. In such cases, it was assumed that the Marine did not give full effort to the CAT-ASVAB test.

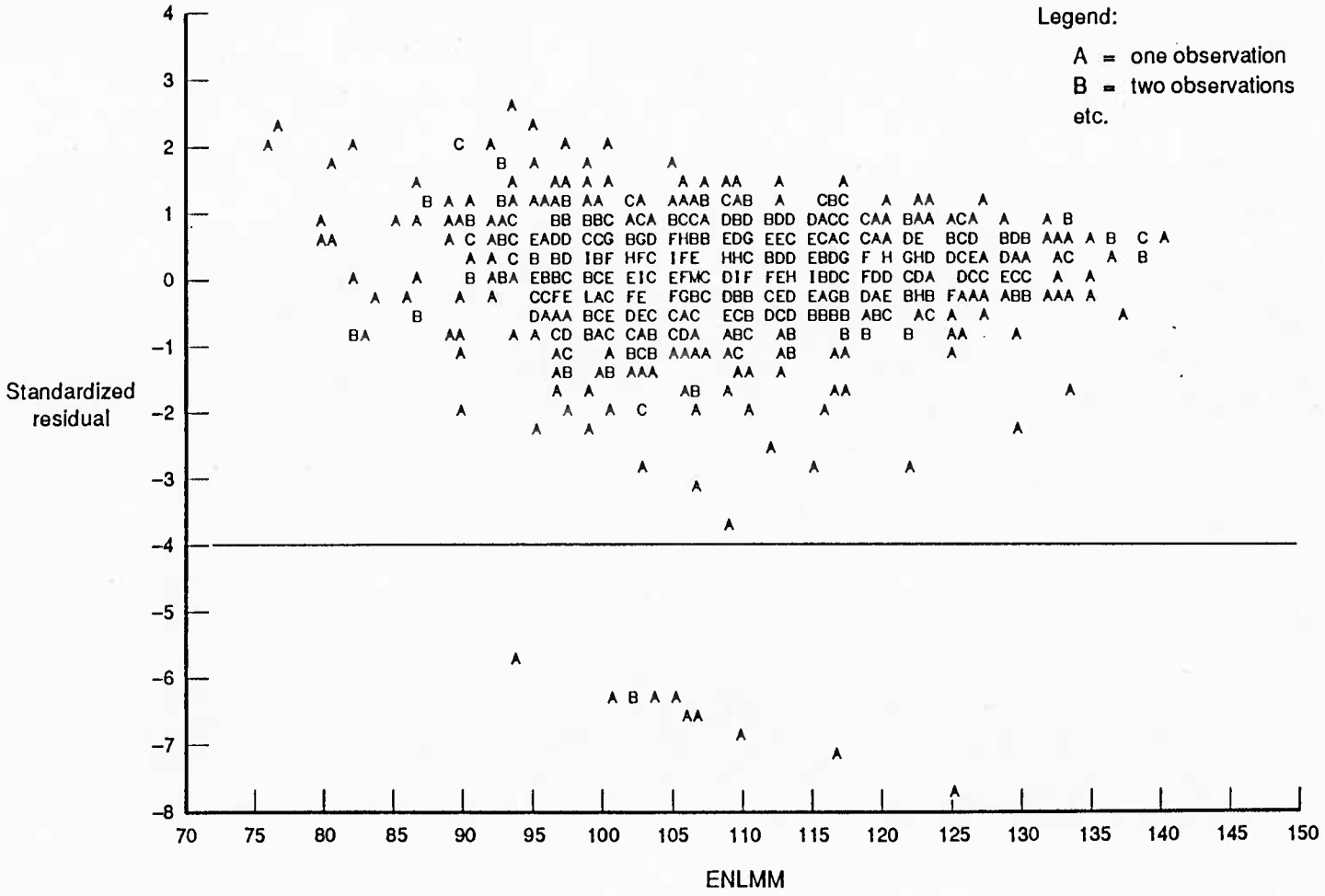
As a check, CAT-ASVAB scores were also predicted from all ten enlistment subtest scores (General Science, Arithmetic Reasoning, Word Knowledge, Paragraph Comprehension, Numerical Operations, Coding Speed, Auto and Shop Information, Mathematical Knowledge, Mechanical Comprehension, and Electronics Information). The adjusted R^2 increased from .35 to .45 using ten subtests. However, the same individuals identified using enlisted MM were also identified as outliers using all ten subtest scores. It was decided to use enlisted MM rather than all ten subtests as the primary method of identifying outliers because 12 percent of individuals did not have complete subtest scores.

Job Knowledge Test

Three quality checks were conducted for the job knowledge test JKT: verification of test form, analysis of item quality, and identification of individuals with questionable job knowledge scores.

Two forms of the JKT were administered. 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 both answer keys. To verify the correct form code, individual total scores resulting from each answer key were compared. A higher total score indicated the correct test form. For borderline cases where the total score was the same for both forms, the reported form was used. Using this method, 12 examinees' scores were changed. Gains resulting from changing the form code ranged from 2 points to 50 points; the average gain was over 26 points.

To assess the measurement quality of items on the JKT, item point-biserial correlations with total test score were computed. Point-biserial correlation is the relation between the scored item response (correct-incorrect) and the total test score. Positive values indicate that the item is probably functioning properly; negative correlations indicate possibly miskeyed items or poorly worded item alternatives. These analyses identified five items with negative correlations that had been miskeyed. Once the key errors were corrected, only one item still appeared questionable. This item had essentially a zero correlation with total score, and the proportion of examinees responding correctly was less than chance. Upon consulting with subject matter experts, it was discovered that this item was ambiguous. Therefore, the item was dropped from further analysis.



NOTE: 20 observations had missing values.

Figure 1. Residuals from regression of CAT-ASVAB MM Composite on enlisted MM

To identify individuals with questionable job knowledge test results, scores were compared with those achieved when taken by the individual under motivated conditions. To check the accuracy of the job knowledge test scores, scores were predicted from enlistment scores and time in service[1] using linear regression. Large negative residuals identified persons whose job knowledge scores were not accurate indicators of their knowledge. Residuals were computed from the regression and plotted as shown in figure 2. Two decision rules were established for the identification of unusually low job knowledge test scores:

- Standardized residual <= -3.0. or
- Standardized residual <= -2.0 and mention in problem logs as an unmotivated performer (e.g., fatigue, illness).

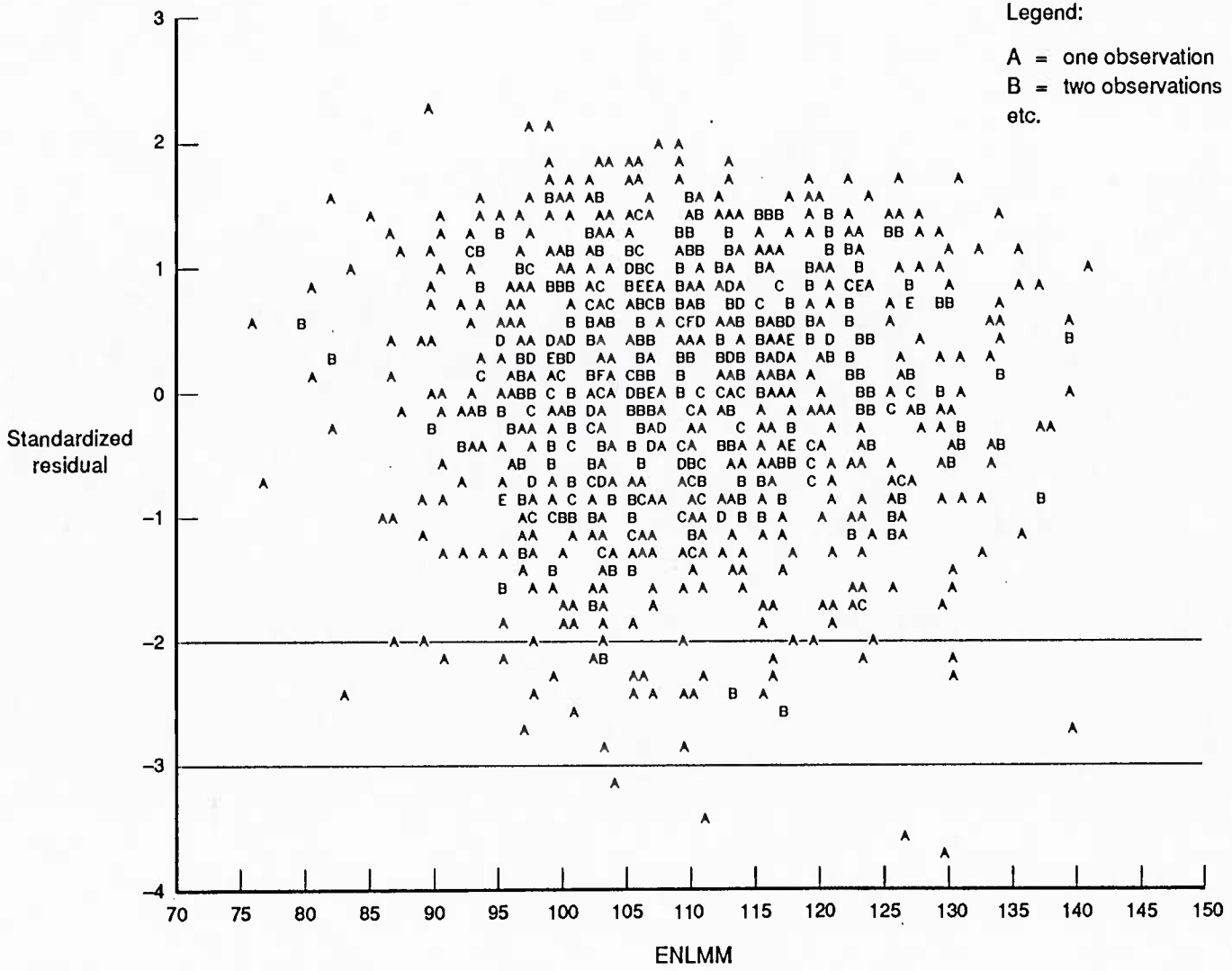
These critical regions for defining aberrant scores are noted on the figure.

Based on these criteria, eight scores were declared aberrant for the JKT.[2] As a result of deleting the eight aberrant scores, the mean for the JKT increased and the standard deviation went down. The correlation of this test with the mechanical maintenance (MM) aptitude composite score increased slightly (table 1). These changes in sample statistics indicate that the scores that were deleted were typically aberrant cases.

A final quantitative verification of residual analyses involved the r_{perbis} statistic [2, 4]. This statistic compares responses to "normal" response patterns. For example, if an individual were guessing, he might answer easy items incorrectly but get difficult items correct. Such a response pattern might also occur if the examinee unknowingly skips a question so that responses are always meant for the adjacent item. The r_{perbis} statistic runs between -1 and 1, with low values indicating an unusual response pattern. Two of the eight individuals identified by residual analysis as unmotivated performers were also identified by r_{perbis}, with r_{perbis} more than 3 standard deviations below the mean value. Further details about r_{perbis} are in appendix B.

---

1. Time in service was used as a predictor for job knowledge and hands-on performance because it was expected that on-the-job experience should predict job knowledge and HOPT. Time in service was not used to predict CAT-ASVAB because experience is not related to mental aptitude.
2. As a check, the ten subtest scores and TIS were also used to predict job knowledge scores. The same individuals were identified as outliers, and no new individuals were identified. Adjusted R^2 increased from .16 to .17 using ten subtest scores.



NOTE: 20 observations had missing values.

Figure 2. Residuals from regression of job knowledge test on enlisted MM and time in service

Table 1. Change in sample statistics due to deleting unusually low scores of JKT

Time	n	Mean	Std	Correlations with Hands-on total	Correlations with Enlisted MM score
JKT					
Before deletion	1,027	65.8	11.5	.44	.38
After deletion	1,019	66.0	11.2	.44	.39

Hands-On Performance Test Scores

Enlisted MM scores reflect motivated test taking, since enlistment scores determine eligibility for the Marine Corps. Although every attempt was made to ensure full effort for JPM testing, some individuals might not have been motivated to perform. Therefore, HOPT scores were regressed on enlisted MM and time in service to determine whether some HOPT performers might have been unmotivated. HOPT scores far below those predicted by enlistment scores were candidates for deletion. Two HOPT cases were deleted after examination of residuals and inspection of the problem logs. In both cases, the examinee's score had a standardized residual of -3.0 or less (that is, the score was extremely unlikely to occur by chance), and the examinees were mentioned in the problem logs as unmotivated performers. This analysis was verified using the ten ASVAB subtests scores as predictors. Adjusted R^2 increased from .23 to .25, but the same individuals were identified as outliers, and no new individuals were identified. The HOPT scores for these two individuals were deleted because their scores did not represent their full effort.

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.

Method

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 [5]. The recommended imputation algorithm is a regression-based procedure that seeks to impute missing values by taking into account the differing levels of task difficulties while maintaining individual differences among examinees [6]. The procedure incorporates a random component equal to the standard error of the estimate to prevent unduly high correlations among variables with imputed values, as compared with

variables with nonimputed values. The procedure also sequentially estimates multiple missing values 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 data imputation is presented in appendix C.

Hands-on performance data were collected at the step level; an examinee was scored as passed or failed. Data were missing at the step level for any of a number of reasons:

- Equipment necessary for completion of the step was missing. This happened at one site for a day because equipment was needed for Operation Desert Shield.
- Broken equipment prevented step completion. This often happened on the "U-Joint" task, early in testing.
- The test administrator's view was blocked, preventing observation of the step performed.
- The examinee was called away from testing before completing a task.
- The examinee refused to perform a task for personal safety reasons.
- The test site was temporarily inoperable due to weather conditions.

The hands-on test was composed of 21 mechanical-maintenance tasks (391 steps), a use-of-manuals test (18 steps), and a forms-completion test (33 steps). Because of the large number of steps, there were many possibilities for missing data. Rather than exclude a data case with some missing data, step scores were imputed as required to obtain a complete record for that individual.

Results

Table 2 details the gains in complete-data cases resulting from imputation of missing data, by task. Overall, complete data were collected for 96 percent of all tasks administered. It can be seen that for every task, more than 85 percent of the cases had a full complement of steps, without need for any imputation whatsoever. Only five tasks had less than 95 percent complete cases: Task 8B (service oil system, 87.0 percent); Task 5B (remove/replace U-joints, 89.6 percent); Task 1A (troubleshoot low-oil-level alarm light, 90.2 percent); Task 8A (remove/replace transmission neutral start switch, 91.2 percent); and Task 4C (bleed brakes, pressure method, 93.3 percent).

Table 2. Complete-step cases, by task (total N = 1,028)

Task no.	Task name	Total steps	Number of complete-step cases	Number of complete-step cases
1A	Troubleshoot low-oil-level alarm light	33	927	90.2
1B	Troubleshoot inoperative folding boom	20	1,006	97.9
2A	General troubleshooting, version 1	3	1,024	99.6
2B	Remove/replace runflat assembly	14	1,019	99.1
2C	Adjust toe in/out	15	1,018	99.0
3A	Adjust/align power steering pump belt	5	1,027	99.9
3B	Repair and replace brake shoes	35	998	97.1
4A	Adjust/align power-steering assist cylinder	10	996	96.9
4B	Troubleshoot inoperative stoplights	39	996	96.9
4C	Bleed brakes, pressure method	23	959	93.3
5A	Troubleshoot engine	13	1,000	97.3
5B	Remove/replace U-joints	18	921	89.6
5C	STE-ICE voltage test	22	1,000	97.3
6A	Troubleshoot winch that will not operate	46	1,008	98.1
6B	Service radiator	16	1,006	97.9
7A	Troubleshoot excessive oil consumption	8	1,024	99.6
7B	General troubleshooting version II	3	1,005	97.8
7C	Remove/replace rear propeller shaft	17	997	97.0
7D	Replace parking brake	10	1,010	98.2
8A	Remove/replace neutral start switch	9	938	91.2
8B	Service oil system	32	894	87.0

The missing data for these tasks were mostly due to difficulties keeping a supply of spare parts. For example, maintaining a full supply of compatible gaskets and oil was a problem during Task 8B; the vice grip and snaprings often broke during Task 5B; there was difficulty during Task 1A keeping a working multimeter and charged battery; there were difficulties during Task 8A keeping an adequate supply of locknuts and neutral start switches; and there were difficulties with breakage when examinees attempted to screw a quick-disconnect into an adapter during Task 4C. In general, table 2 shows a very high percentage of complete data.

Table 3 presents the degree of imputation for the cases that had any missing steps. There were a total of 21 hands-on tasks. Table 3 shows that imputation of a single step completed the data for 38 percent (131/345) of the cases with missing data, and over half (176/345) of the imputed cases had imputed steps for only one task. Table 3 shows that typically a small number of steps were imputed. Over half of the cases with any imputed data had three or fewer imputed steps.

Given this degree of imputation at the step level, what was the impact on the sample statistics of the respective hands-on scores? Table 4 shows the changes in means and standard deviation due to imputation. The shifts in mean performance are relatively small compared to the standard deviation of the performance scores.

Note that the imputation procedure was developed to maintain relations among variables, without unintentionally increasing correlations. Correlations would have increased if the predicted performance were imputed without adding a random component to the computations. Table 5 shows that the corrected correlation[1] of the hands-on test with aptitude composites changed very little as a result of this process. Figure 3 shows that imputed cases fell in the full range of both aptitude and hands-on performance.

---

1. Validities were corrected for multivariate restriction of range, in accordance with the recommendation of the Committee on the Performance of Military Personnel [5], the scientific oversight committee for the joint-service JPM project. The procedure corrects for the fact that observed correlations were computed from a restricted group, i.e., Marines who were specially selected for the job. Since they are a more homogeneous group than the total applicant population, their test scores will have less variance than the total group, and observed correlations are usually lower than would be found for the entire population. Further details concerning multivariate correction procedures can be found in CRC 336 [7] and Gulliksen [8].

Table 3. Number of hands-on tasks and steps inputed for examinees (n = 1,028)

Steps	Tasks 0	1	2	3	4	5	6	7	8	9	10	Total(%)
0	683	0	0	0	0	0	0	0	0	0	0	683(66.4%)
1	0	131	0	0	0	0	0	0	0	0	0	131(12.7)
2	0	17	21	0	0	0	0	0	0	0	0	38 (3.7)
3	0	9	4	4	0	0	0	0	0	0	0	17 (1.6)
4	0	5	8	0	2	0	0	0	0	0	0	15 (1.5)
5	0	1	3	3	1	0	0	0	0	0	0	8 (0.8)
6-10	0	5	4	2	2	3	0	0	0	0	0	16 (1.6)
11-15	0	3	0	10	35	0	0	0	0	0	0	48 (4.7)
16-20	0	2	0	3	13	10	1	0	0	0	0	29 (2.8)
21-35	0	3	3	1	3	5	10	2	0	0	0	27 (2.6)
36-60	0	0	1	0	0	1	1	1	1	0	0	5 (0.5)
>60	0	0	0	1	4	2	0	1	2	0	1	11 (1.1)
	683	176	44	24	60	21	12	4	3	0	1	1,028(100.0)

NOTE: Entries are numbers of examinees. For example, the first entry (683) indicates that, for 683 examinees, no steps and no tasks were imputed.

Table 4. Comparison of task statistics using imputed and complete data

	Original data (complete cases only)					Fully imputed (imputed any missing cases)				
	n	mean	sd	min.	max.	n	mean	sd	min.	max.
1A	926	89.6	11.4	3	100	1,026	89.3	11.6	3	100
1B	1,004	86.5	16.7	0	100	1,026	86.2	16.9	0	100
2A	1,022	80.2	31.9	0	100	1,026	80.1	32.0	0	100
2B	1,017	79.8	14.6	21	100	1,026	79.8	14.6	21	100
2C	1,016	79.3	20.5	0	100	1,026	79.1	20.7	0	100
3A	1,025	82.9	19.4	0	100	1,026	82.8	19.4	0	100
3B	996	75.5	20.9	7	100	1,026	75.3	21.1	7	100
4A	994	86.3	15.8	0	100	1,026	85.9	16.4	0	100
4B	994	84.4	19.1	6	100	1,026	84.0	19.6	6	100
4C	957	73.7	29.9	0	100	1,026	73.2	30.0	0	100
5A	999	80.3	18.7	8	100	1,026	80.2	18.9	8	100
5B	920	88.2	12.5	0	100	1,026	87.8	13.0	0	100
5C	999	77.0	30.1	0	100	1,026	76.7	30.2	0	100
6A	1,007	85.3	12.4	11	100	1,026	85.1	12.8	11	100
6B	1,004	70.8	23.9	0	100	1,026	70.4	24.3	0	100
7A	1,023	59.5	23.0	0	100	1,026	59.4	23.0	0	100
7B	1,004	62.0	37.1	0	100	1,026	61.3	37.5	0	100
7C	996	83.1	15.5	0	100	1,026	82.9	15.8	0	100
7D	1,009	82.2	23.3	0	100	1,026	81.9	23.6	0	100
8A	937	88.9	8.6	22	100	1,026	89.4	8.9	22	100
8B	894	62.7	23.0	6	100	1,026	63.4	23.0	6	100
FRM	1,009	46.2	21.2	0	100	1,022	46.1	21.2	0	100
MAN	1,025	67.9	22.9	0	100	1,026	67.9	22.9	0	100
TOT	670	77.8	8.1	43	95	1,026	76.9	8.3	43	95

NOTE: n = 1,026 because two cases were deleted because of aberrant scores.

Table 5. Validity of MM composite with hands-on total score before and after imputation

	Before imputation		After imputation	
	n	r	n	r
Enlisted MM	674	.70	1,007	.69
CAT-MM	676	.73	1,012	.72

NOTE: The validities are based on the unit-weighted total score. The validities might be slightly different after scores are weighted.

REPORTED DATA COLLECTION PROBLEMS

As described earlier, there were several data collections in addition to hands-on mechanical testing. The ECAT is a computerized test battery of new predictors of job performance, such as one-hand tracking, short-term memory, and reaction time. The CAT-ASVAB is a computerized adaptive version of the regular ASVAB. The "administrative duties" test was a hands-on test of the Marine's proficiency in two tasks--looking up information in manuals and filling out forms. Three aptitudes were tested by the GATB (General Aptitude Test Battery) portions taken: motor coordination, manual dexterity, and finger coordination. The tasks involved writing as many symbols as possible, placing rivets, or assembling rivets within a constrained time. The job knowledge test consisted of multiple-choice paper-and-pencil items written to parallel the hands-on tasks as closely as possible.

Relatively few missing data were reported for the ECAT, CAT-ASVAB "administrative duties," GATB, and JKT data collection, as shown in table 6. Nevertheless, it is useful to pinpoint the reasons for missing data to direct the design of future data collection efforts.

This section analyzes the problems that were reported concerning data collection. This information is important because it describes the difficulties encountered in collecting performance data from multiple sources.

At both sites, test administrators filled out daily logs that described any abnormalities or deviations from expected test procedures for the ECAT, CAT, administrative duties, GATB, and job knowledge test. The frequency of particular categories of reported problems from these logs forms the basis of the following analyses.

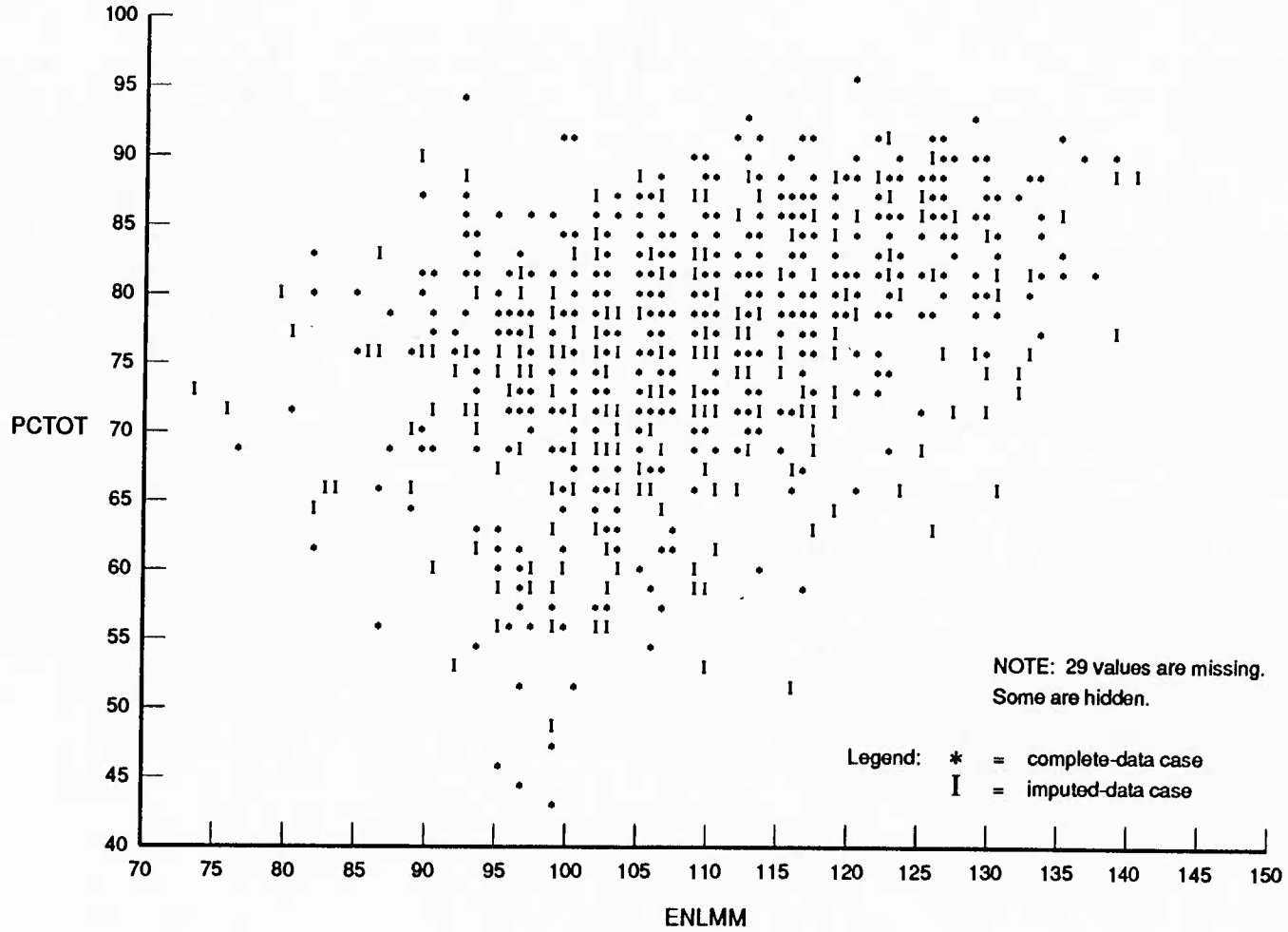


Figure 3. Validity of hands-on total score vs. mechanical maintenance enlisted composite score for both imputed and complete data

Table 6. Amount of missing data for CAT-ASVAB, administrative duties, GATB, and JKT

Test	Amount missing	
	Number of cases	Total percentage of missing cases
ECAT	51	4.9
CAT-ASVAB	17	1.6
Administrative duties	5	0.5
GATB	8	0.8
Job Knowledge Test	1	0.1

Results

The variations in the number and type of problems encountered at the two testing sites were large. Part of this variation across site was due to differences in testing conditions. Camp Pendleton personnel reported the most problems, but that site also was the first to begin collecting data, it tested more examinees, it tested for two weeks longer than Lejeune, and its computer testing site had a considerable problem with dust and heat. There might have been some differences in the ability of test administrators across site (e.g., some test administrators were able to devise more effective strategies for saving computer data, while others had to be retrained in data saving procedures).

Although some cross-site variation is expected, some site differences were due to the degree of attention given to problem logs--some test administrators were more attentive to keeping problem logs than were others. Despite this fact, problem logs could be useful for (1) pointing out relative magnitudes of problems within site, (2) analyzing the degree to which problems are recognized by test administrators, and (3) suggesting the magnitude of problems that might occur if these tests were to be given operationally. For example, depending on policy decisions, some form of the ECAT and CAT-ASVAB could be used operationally as replacements for the currently used paper-and-pencil ASVAB.

Documented testing problems were categorized based on the general duties of the test administrator required for each testing session:

- Site setup
  - -- Determine the completeness and soundness of testing equipment (e.g., computer)
  - -- Prepare testing equipment for administration

- Test administration

  --Deal with problems during the test

- Data collection and verification

  -- Collect test data (e.g., disk information, answer sheets)

  -- Confirm that test data have been collected properly, and if necessary, collect data again

  -- Send data to a central site, with proper identification attached to test responses.

The training sessions for test administrators focused on these test administration tasks in the attempt to prevent missing data and other test administration problems. These are the points where it is most likely for the test data to be lost.

Table 7 shows a list of problems that can occur at each stage of the data collection process. It is clear that the complex and sometimes fragile nature of computers predisposes computer-administered tests to a larger number of possible failures to collect complete data, and requires more highly qualified and trained test administrators. This observation is confirmed by the fact that even though the total amount of missing data was quite small, the computerized CAT-ASVAB resulted in significantly more missing data than did the job knowledge test. Test administrators ideally are able to recognize malfunctioning disk drives, computer boards, response pedestals, and video screens.

Twelve failure types were identified as applicable to the present data collection effort. Table 8 shows the frequencies of computerized testing problems, by testing site. The frequencies indicate that across sites, the largest number of computer testing problems occurred for the ECAT (82). For the ECAT, the most frequent problems involved the response pedestal. Other problems encountered (in decreasing order of frequency) were data collection failure, video computer error, difficulty with the "HELP" key, skipping a test portion, and boot failures.

The CAT-ASVAB had fewer problems (19) than did the ECAT. The most prevalent problems for the CAT-ASVAB were failure to collect and boot failure. Other difficulties occurred relatively infrequently.

Table 9 tabulates the 65 problems common all testing modes. It is striking that insufficient time or scheduling problems for the administrative duties test provided by far the largest single category of difficulties. The number of problems for other testing modes was small.

Table 7. Plausible data collection errors, by source

Source	Before test administration	During test administration	After test administration
ECAT	Bent pedestals	Pedestal breaks	Disk failure during collection of data
	Broken computer boards	Screen goes blank	Running out of disk space during collection
	Broken video screen	Power outage	Mail service failure
	Pedestal fails to calibrate	Disk failure	Difficulty identifying examinee
		Cheating	
		Examinee rushes, doesn't try	
		Examinee physical disability	
CAT	Broken computer boards	Screen goes blank	Disk failure
	Disk failure	Disk failure	Network failure
	Broken video screen	Power outage	Drive failure
		Cheating	Mail service failure
		Examinee rushes, doesn't try	Running out of disk space during collection
		Examinee physical disability	Difficulty identifying examinee
Self-administered	Errors in printing materials	Cheating	Mail service failure
	Disorganized materials	Examinee rushes, doesn't try	Difficulty identifying examinee
	Wrong form administered	Examinee physical disability	
GATB	Bent materials	Cheating	Mail service failure
		Examinee rushes, doesn't try	Difficulty identifying examinee
		Examinee physical disability	
JKT	Errors in printing materials	Cheating	Mail service failure
	Wrong form administered	Examinee rushes, doesn't try	Difficulty identifying examinee
		Examinee physical disability	

Table 8. Problems of computerized data collection, by site

Problem	Test mode				Total
	ECAT		CAT/ASVAB		
	Pendleton	Lejeune	Pendleton	Lejeune	
Response pedestal	19	8	NA	NA	27
Skipped test portion	6	0	1	0	7
Power cord	1	1	0	1	3
Boot failure	5	1	3	1	10
Collection failure	6	9	3	3	21
Help key	4	4	0	0	8
Identification problem	1	0	0	0	1
Video/computer error	8	3	0	0	11
Examinee didn't try	1	0	2	0	3
Examinee unable (e.g., tired, sick)	2	0	2	0	4
Scheduling	2	0	1	0	3
Outside disruption	1	0	2	0	3
Total	56	26	14	5	101

Table 9. Common problems across test mode, overall and by site

	ECAT		CAT		Administrative		GATB		Job Knowledge		
	Pendleton	Lejeune	Pendleton	Lejeune	Pendleton	Lejeune	Pendleton	Lejeune	Pendleton	Lejeune	Total
Insufficient time	0	0	0	0	26	0	0	0	1	0	27
Wrong form	0	0	0	0	0	2	0	0	0	3	5
Miscellaneous	0	0	0	0	0	1	0	0	0	0	1
Scheduling	2	0	1	0	4	0	1	0	1	0	9
Examinee didn't try	1	0	2	0	2	0	1	0	2	1	9
Examinee disabled (e.g., sick, tired)	2	0	2	0	1	0	2	0	0	1	8
Outside distruption	1	0	2	0	1	0	1	0	1	0	6
Total	6	0	7	0	34	3	5	0	5	5	65

When data for the two sites were aggregated, the most frequent difficulties for each testing mode were as follows:

ECAT

(1) Response pedestal/joystick breakage or decalibration (33%)

(2) Failure to collect data onto disk (18%)

(3) Video/computer errors during testing (13%)

(4) Inability to continue testing after "HELP" key depressed (9.8%)

(5) Skipping a test portion (7.3%)

(6) Boot failures (7.3%)

(7) All others (11.6%)

CAT

(1) Failure to collect data (31.6%)

(2) Boot failure (21.1%)

(3) Lack of effort (10.5%)

(4) Disability(10.5%)

(5) Disruption(10.5%)

(6) All others (15.8%)

Administrative duties

(1) Insufficient time (70%)

(2) Scheduling (11%)

(3) All others (19%)

GATB

(1) Physical disability of examinee (40%)

(2) Examinee rushed/didn't try (20%)

(3) All others (40%)

Job Knowledge Test

(1) Examinee rushed/didn't try (30%)

(2) Wrong form (30%)

(3) All others (40%).

The problem logs indicated relatively few difficulties collecting data and maintaining data quality. This confirms the findings from tabulations of missing data. Nevertheless, the problem logs showed that each data source presented characteristic challenges to field data collection: examinees sometimes hurried through the job knowledge test; they ran out of time for the administrative duties test; prior hand injuries occasionally prevented taking the GATB; response pedestal problems periodically hampered the ECAT; and disk failures sometimes obstructed CAT-ASVAB data collection.

Recommendations

Future data collection efforts will benefit if specific procedures are developed to alleviate some of the most common difficulties. Test administrators for the CAT-ASVAB and ECAT tests should always verify that data have been transferred to backup disks immediately. Test administrators should be specifically trained to check data disks for capacity limits so that data overflow problems do not occur. Last, clean, air-conditioned spaces are preferable for the administration of computerized tests.

CONCLUSIONS

HOPT

Relatively few data quality problems were found for the HOPT. Only two examinees both were mentioned by the problem logs and had unexpectedly low scores. These individuals' scores were deleted. Complete data were collected for 96 percent of all tasks administered. The effect of imputing the remaining points was minimal, in terms of mean, standard deviation, and correlation with aptitude composites.

JKT

The JKT had relatively few unusual response patterns. It appears that 12 of more than 1,000 examinees recorded the wrong test form on their answer sheets, and these aberrations were detected by scoring the tests with both answer keys. Only 1 of 145 items had sufficiently poor properties to be deleted from the JKT. Lastly, the patterns of only eight examinees' scores were unusual enough to delete their scores from further consideration. The effect of deleting these few examinees'

scores was to increase the average score slightly, decrease the standard deviation, and increase the correlation of JKT with aptitude. The changes in all cases were extremely small.

CAT-ASVAB

Residual analyses and problem logs pinpointed 13 individuals who apparently did not make a full effort on the CAT-ASVAB. Of these, 11 of the individuals scored the absolute minimum for the CAT-ASVAB, and the other two scored much lower than would be predicted by their enlisted MM scores. The small percentage of unusual CAT-ASVAB scores suggests that the offer to change Marines' scores of record if they improved their enlistment score was a generally successful inducement.

REFERENCES

[1] American Institutes for Research Report AIR-70900-FR 02/91, *Develop and Administer Measures for Mechanical Maintenance Occupational Area,* Volume I: *Test Development*, by Jennifer L. Crafts, et al., 15 Feb 1991.

[2] 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, Unclassified, Mar 1989

[3] M. H. Maier. "On the Need for Quality Control in Validation Research." *Personnel Psychology*, 41 (1988): 497-502

[4] 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

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

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

[7] CNA Research Contribution 336, *A Method To Correct Correlation Coefficients for the Effects of Multiple Curtailment,* by Thomas L. Mifflin and Stephen M. Verna, Aug 1977

[8] Harold Gulliksen. *Theory of Mental Tests*. Hillsdale, NJ: Lawrence Erhlbaum Associates, 1987

APENDIX A

DETAILS OF RESIDUAL ANALYSES

APPENDIX A

DETAILS OF RESIDUAL ANALYSES[1]

Residual analyses are conducted to determine whether the assumptions underlying linear regression are correct, and to check for outlier scores that might not reflect the examinee's true ability.

Figure A-1 shows a relationship between criterion and predictor. Note that the criterion is represented along the vertical (y) axis and the predictor along the horizontal (x) axis. Linear regression calculates the equation for a line that minimizes the distance between predicted scores (points on the regression line) and actual criterion scores. For example, the value of the $\underline{i}$th observation would be

$$y_i = B_0 + B_1 X_i + e_i \ , \qquad (1)$$

where y_i is a value of the dependent variable, x_i is a value of the predictor variable, B_0 and B_1 are unknown parameters to be estimated, and e_i is an error term. As represented on the figure, B_0 would be the y intercept and B_1 is the slope of the line. The line represents a set of predicted y_i, given a value of the predictor. Predicted y_i is often represented as

$\hat{y}_i$, where the "^" indicates that it is a predicted value.

The residual, e_i, is the difference between the observed and predicted criterion values, as shown in figure A-1:

$$e_i = y_i - \hat{y}_i \ .$$

Negative values for residuals are computed when the actual criterion score, y_i, is below the predicted y_i. Positive values are computed when the observed score exceeds the predicted. The variance of the residuals generated under model (1) is:

$$\frac{\Sigma(e_i - \bar{e})^2}{n - 2} = \frac{SSE}{n - 2} = MSE \ .$$

---

1. J. Neter and W. Wasserman's *Applied Linear Statistical Models*. Homewood, Illinois: Richard D. Irwin Co., 1974.

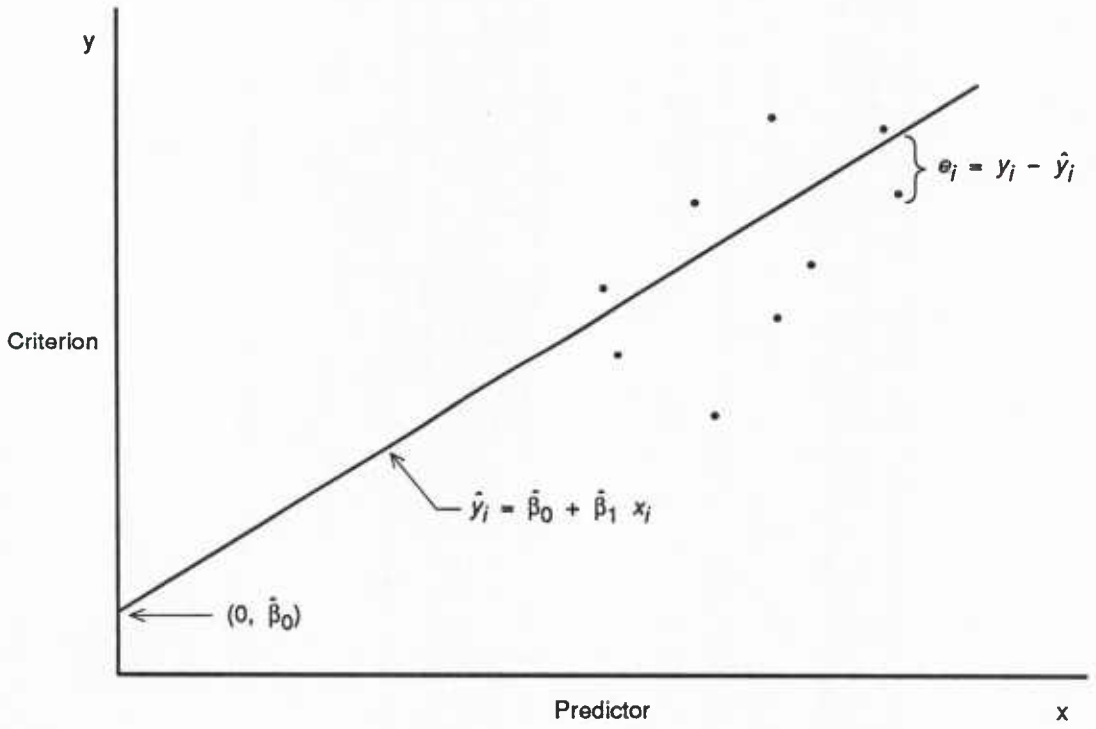


Figure A-1. Illustration of linear regression and residual analysis

Scores with residuals large in absolute value are outliers. It is questionnable whether scores with extremely large negative residuals reflect an individual's true ability to perform.

In order to judge how extreme an outlier is, the residuals must be put on a scale that will assist in its interpretation--i.e., the residual is standardized. Standardized residuals can be calculated by dividing e_i's by the square root of their variance. The standardized residual is:

$$\frac{e_i}{(MSE)^{1/2}}$$

Because this quantity has been standardized, it is expected, under the assumption of a normal distribution, that 99.74 percent of standardized residuals will fall between -3.0 and 3.0. Therefore, scores with standardized residuals below -3.0 are considered outliers.

APPENDIX B

COMPUTATION OF THE R_{PERBIS} STATISTIC

APPENDIX B

COMPUTATION OF THE R_{PERBIS} STATISTIC

Mayberry has described the computation of the r_{perbis} statistic in earlier JPM work.[1] What follows is a summary of that description.[2]

The personal biserial correlation (r_{perbis}) was proposed by Donlon and Fischer as a means of evaluating the appropriateness of a person's total score in measuring his or her ability. Determinations of appropriateness are made relative to the responses of a reference sample. The r_{perbis} statistic quantifies the similarity of item difficulties 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 is that 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.

The personal biserial correlation (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.

---

1. CRM 88-259, *Analysis of Data Quality for the Infantry Phase of the Marine Corps Job Performance Measurement Project*, by Paul W. Mayberry, Mar 1989

2. Further details can be found in Mayberry or in T.F. Donlon's and F.E. Fisher's "An Index of an Individual's Agreement Group-Determined Item Difficulties," in *Educational and Psychological Measurement* 28 (1968): 105-113.

APPENDIX C

DATA IMPUTATION PROCEDURES

APPENDIX C

DATA IMPUTATION PROCEDURES

The imputation procedure used for this study attempts to maintain the correlational structure of the original data, unlike many other imputation methods.[1]

COMPUTATIONS REQUIRED FOR IMPUTATION

The initial step in the imputation 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 the magnitude of their 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 B-1 presents a hypothetical expectancy table for a discrete variable (e.g., a rating scale with values ranging from 1 to 5).

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.

---

1. Imputation procedures for the estimation of incomplete data are fully described in CRM 88-259, *Analysis of Data Quality for the Infantry Phase of the Marine Corps Job Performance Measurement Project*, by Paul W. Mayberry, March 1989, and L.L. Wise's and D. McLaughlin's *Guidebook for the Imputation of Missing Data* (Palo Alto, CA: American Institutes for Research, 1980).

APPENDIX D

UNCORRECTED VALIDITIES WITH HANDS-ON TOTAL SCORE BEFORE AND AFTER IMPUTATION

The validities below are based on the unit-weighted total score. The validities might be slightly different after scores are weighted.

	Before imputation		After imputation	
	n	r	n	r
Enlisted MM	674	.47	1,007	.43
CAT MM	676	.58	1,012	.55

REPORT DOCUMENTATION PAGE

Form Approved
OPM No. 0704-0188

1. AGENCY USE ONLY *(Leave Blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	August 1991	Final

4. TITLE AND SUBTITLE

Data Quality for the Automotive Maintenance Phase of the Marine Corps Job Job Performance Measurements (JPM) Project

5. FUNDING NUMBERS

C - N00014-91-C-0002

PE - 65153M

PR - C0031

6. AUTHOR(S)

Neil B. Carey, Catherine M. Hiatt

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

Center for Naval Analyses
4401 Ford Avenue
Alexandria, Virginia 22302-0268

8. PERFORMING ORGANIZATION REPORT NUMBER

CRM 91-120

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

Commanding General
Marine Corps Combat Development Command (WF 13F)
Studies and Analyses Branch
Quantico, Virginia 22134

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

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 automotive maintenance portion of the Marine Corps Job Performance Measurement Project, then describes measures taken to minimize the effect of questionable or missing cases. Particular attention is focused on data inconsistencies and problems associated with operational fielding of Computer Aptitud Test-Armed Services Vocational Aptitude Battery (CAT-ASVAB) and the Enhanced Computer Administered Test (ECAT).

14. SUBJECT TERMS	15. NUMBER OF PAGES
ASVAB (armed services vocational aptitude battery), CAT (computer adaptive test), Data acquisition, JPM (job performance measurement), Maintenance personnel, Marine Corps personnel, Performance (human), Performance tests, Problems, Quality, Test methods, Validation	46
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
CPR	CPR	CPR	SAR

Work conducted under contract N00014-91-C-0002.