

AL-TP-1992-0025

AD-A252 330



**ASSESSING THE SUBSTITUTABILITY OF
SURROGATE MEASURES OF JOB PERFORMANCE
FOR HANDS-ON WORK SAMPLE TESTS**



ARMSTRONG

Frances J. Laue

UES, Incorporated
Human Factors Division
4401 Dayton-Xenia Road
Dayton, OH 45432-1894

**DTIC
ELECTE
JUL 01 1992
S B D**

**Mark S. Teachout
Donald L. Harville**

**HUMAN RESOURCES DIRECTORATE
TECHNICAL TRAINING RESEARCH DIVISION
Brooks Air Force Base, TX 78235-5000**

LABORATORY

June 1992

Final Technical Paper for Period January 1989 - January 1990

Approved for public release; distribution is unlimited.

92 6

079

92-17244



**AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5000**

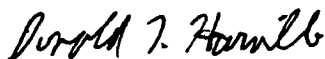
NOTICES

This technical paper is published as received and has not been edited by the technical editing staff of the Armstrong Laboratory.

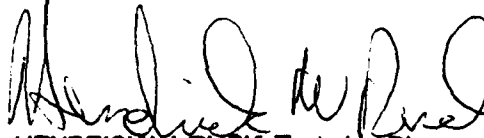
When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way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

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

This paper has been reviewed and is approved for publication.



DONALD L. HARVILLE
Project Scientist



HENDRICK W. RUCK, Technical Director
Technical Training Research Division



RODGER D. BALLENTINE, Colonel, USAF
Chief, Technical Training Research Division

REPORT DOCUMENTATION PAGEForm Approved
OMB No. 0704-0188

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

1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE June 1992	3. REPORT TYPE AND DATES COVERED Final - January 1989 - January 1990	
4. TITLE AND SUBTITLE Assessing the Substitutability of Surrogate Measures of Job Performance for Hands-on Work Sample Tests			5. FUNDING NUMBERS C - F41869-86-D-0052 PE - 63227F PR - 2922 TA - 01 WU - 01	
6. AUTHOR(S) Frances J. Laue Mark S. Teachout Donald L. Harville				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) UES, Incorporated Human Factors Division 4401 Dayton-Xenia Road Dayton, OH 45432-1894			8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES) Armstrong Laboratory Human Resources Directorate Technical Training Research Division Brooks Air Force Base, TX 78235-5000			10. SPONSORING/MONITORING AGENCY REPORT NUMBER AL-TP-1992-0025	
11. SUPPLEMENTARY NOTES Armstrong Laboratory Technical Monitor: Dr. Donald L. Harville, (512) 536-2932				
12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.			12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 words) The Air Force developed Hands-On work sample tests to serve as (a) criteria for validation of the Armed Services Vocational Aptitude Battery (ASVAB) and (b) benchmarks for assessment of less expensive and easier to administer surrogate measures (i.e., interview-format work sample tests, rating forms, job knowledge tests, training school grades). Measures were administered to 1,491 enlisted airmen in eight different jobs. Results indicated that the Hands-On measures were most strongly related to interview tests and job knowledge tests. Factor analyses of performance criteria revealed that technical proficiency was represented by three distinct factors (i.e., self ratings, supervisor ratings, and performance scores), an indication that these criteria measure multiple constructs. None of the surrogates could be considered interchangeable or substitutable for the Hands-On measures. Future work should investigate whether selection and classification decisions and minimum aptitude standards set for selection and classification would differ using Hands-On versus surrogate measures.				
14. SUBJECT TERMS Job knowledge Performance measures			15. NUMBER OF PAGES 40	
Test validation Work samples			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UL	

TABLE OF CONTENTS

	Page
INTRODUCTION	1
METHOD	4
Subjects.	4
Measures.	5
Hands-on Work Sample Tests	5
Interview Work Sample Test	5
Rating Scales.	6
Aptitude	6
Experience	8
Procedure	8
RESULTS.	10
DISCUSSION	21
REFERENCES	26

LIST OF TABLES

Table	Page
1 Measures Included in Analyses	7
2 Corrected Correlations Between Hands-On Score and JPM Variables	11
3 Percent Hands-On Variance Accounted for by the JPM Variables	13
4 Percent Hands-On Variance Accounted for by the JPM Variables (Without Interview Test)	15
5 Regression with Sets of JPM Variables: Variance Accounted for in Hands-On Dependent Measure	18
6 Patterns of Relationships Between Performance and Measures of Aptitude and Experience	20



Accession For	
NTIS GRA&I	<input checked="" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By _____	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

PREFACE

The research described in this paper was conducted under Contract No. F41689-86-D-0052, awarded to Universal Energy Systems, Inc. The purpose of the research described in this paper was to compare the usefulness of surrogate measures of job performance against hands-on job performance test scores.

SUMMARY

As part of a joint-Service project on the assessment of enlisted personnel performance, the Air Force developed hands-on work sample tests to serve as (a) criteria for validation of the Armed Services Vocational Aptitude Battery (ASVAB) and (b) benchmarks for assessment of less expensive and easier to administer surrogate measures (i.e., interview-format work sample tests, rating forms, job knowledge tests, training school grades). Measures were administered to 1491 enlisted airmen in eight different jobs. Results indicated that the hands-on measures were most strongly related to interview tests and job knowledge tests (mean $r = .73$ and $.58$, respectively). Factor analyses of performance criteria revealed that technical proficiency was represented by three distinct factors (i.e., self ratings, supervisor ratings, and performance scores), an indication that these criteria measure multiple constructs. None of the surrogates could be considered interchangeable or substitutable for the hands-on measures. Future work should investigate whether selection and classification decisions and minimum aptitude standards set for selection and classification would differ using hands-on versus surrogate measures.

ASSESSING THE SUBSTITUTABILITY OF SURROGATE MEASURES OF JOB PERFORMANCE FOR HANDS-ON WORK SAMPLE TESTS

The Joint-Service Job Performance Measurement/ Enlistment Standards Project was initiated in the early 1980s to provide data that could address questions arising about the validity of military classification and selection systems. This initial objective was extended by the Air Force Human Resources Laboratory (AFHRL) to provide data for other research issues (Hedge & Teachout, 1986). A measurement technology was needed to address a wide variety of purposes including selection, classification, training evaluation, training needs analysis, and other personnel research. As a result of these diverse research needs, the Job Performance Measurement System (JPMS) consists of several different approaches to performance assessment.

Hands-on testing was designated as the benchmark measure for the Joint-Service Project since it most closely represents on-the-job behaviors, work settings, and performance, and, therefore, elicits the truest sample of job performance (Wigdor & Green, 1986). The Air Force's hands-on work sample test contained detailed step-by-step checklists specifying the conditions, standards, and behaviors for successful performance on a set of tasks representative of the job of the first term airman. Other measures were designed to serve as surrogates or supplements to hands-on testing. An interview work sample test required the job incumbent to describe the steps necessary for

task completion in a "show-and-tell" manner without the aid of technical information. The use of the interview test format allowed researchers to expand the range of tasks to be tested by including those where time constraints or potential safety hazards were concerns. Together, the hands-on and interview tests comprised the Walk-Through Performance Test (WTPT). Four rating forms, varying in level of specificity, were created for administration to three rating sources (i.e., self, supervisor, and peer). Questionnaires were designed to gather background, task experience, and attitudinal data. Paper-and-pencil multiple-choice Job Knowledge Tests (JKTs) were developed, paralleling the task content of the work sample tests. Finally, archival data from Air Force documents and data files were accessed as potential supplemental measures of job performance.

The Air Force focused on the technical proficiency of airmen, an important aspect of overall job performance, for the development of the hands-on testing methodology. Alternative measures were also designed to measure technical proficiency of enlisted personnel, yet each was unique in design and format. As a result, there are differences among measures in their fidelity as gauges of actual job performance. As Green (1984) stated, "the difference between the surrogate and the hands-on measure is more than a nuisance, it is a research topic" (p. 2). A key research question for the Joint-Service Job Performance Measurement (JPM) Project was the issue of comparative adequacy of performance criteria and the substitutability of measures for

one another. More specifically, the psychometric adequacy and statistical properties of the surrogate measures must be established and evaluated against a standard (i.e., the benchmark hands-on approach). Ideally, a surrogate would have high content validity, be highly reliable, and be construct valid, much like the benchmark. The methodology of WTPT task sampling (Lipscomb, 1987) and subject-matter validation of test content resulted in high content validity for the WTPT and the surrogates which were based on WTPT content (e.g., JKT, Task Rating Form). Other studies of the JPMS data have revealed acceptably high reliability of instruments, ranging from .7 to .9 across measures (Hedge, Teachout, & Laue, 1990; Kraiger, 1989, 1990). This current research, among others being conducted, addresses the issue of construct validity.

This research examines whether any of the JPMS measures, or combinations of measures, could serve as substitutes for hands-on performance assessment. Although no single analytic approach will provide enough evidence for the establishment of substitutability, Wigdor and Green (1986) suggest that a correlation between the benchmark and a surrogate of .90 be a criterion for initially identifying substitutability. Gottfredson (1986) provides additional guidance for making comparison among alternative measures (e.g, factor analyses, regression, correlational studies). The current study uses these approaches to address the substitutability issue.

Administrative and economic issues drive the search for

suitable surrogate measures. Although the Hands-on Testing procedure may achieve the highest possible level of fidelity, developmental and administrative costs make it an unaffordable choice for future large-scale research efforts or operational implementation. More cost-effective performance measures (e.g., ratings, paper-and-pencil tests) may be technically suitable surrogates for Hands-on Testing and might better meet the needs of the users such as Department of Defense manpower, personnel, and training communities.

Method

Subjects

Over a five-year period (1982-1987), eight enlisted specialties (i.e., jobs) were studied and instruments were developed for each. Two jobs from each of the four Aptitude Index (AI) areas used by the Air Force for classification were selected for inclusion in the JPM Project as follows: (a) Mechanical Aptitude, Jet Engine Mechanic (AFS 426X2, $N = 255$) and Aerospace Ground Equipment Specialist (AFS 423X5, $N = 261$); (b) Administrative Aptitude, Information Systems Radio Operator (AFS 492X1, $N = 156$) and Personnel Specialist (AFS 732X0, $N = 197$); (c) General Aptitude, Air Traffic Control Operator (AFS 272X0, $N = 191$) and Aircrew Life Support Specialist (AFS 122X0, $N = 195$); and Electronic Aptitude, Avionic Communications Specialist (AFS 328X0, $N = 98$) and Precision Measurement Equipment Laboratory Specialist (AFS 324X0, $N = 138$). A total of 1491 job incumbents in their first enlistment were participants.

Measures

Hands-on work sample tests. The hands-on work sample test was constructed to assess incumbent job proficiency on tasks representative of the job. A domain task sampling plan was developed (Lipscomb, 1984), and tasks were sampled with stratified random sampling procedures (Lipscomb, 1984; Lipscomb & Dickinson, 1987). Test developers used technical orders and manuals (i.e., descriptions of work procedures), as well as input from subject matter experts (SMEs) to define and describe the procedural steps required for successful task completion. A "Yes/No" format was used to score each step to be performed and the proportion of steps performed correctly was calculated for each task.

The work sample tests were administered to job incumbents by active duty noncommissioned officers who had extensive work experience in the jobs tested. These administrators received one to two weeks of observation and scorer training (Hedge, Lipscomb, & Teachout, 1988) which resulted in high level of inter-scorer agreement (median $r = .97$) (Hedge et al., 1990).

Interview work sample test. Performance of interview tasks required the job incumbent to describe the steps required to perform the identified task in a "show-and-tell" manner. No technical data or information were allowed to be used although examinees could point to specific tools and equipment in explaining their approach to task performance. Items included in the interview portion of the WTPT were constructed in a manner

identical to that described for the hands-on items. Inter-scorer agreement with the interview test approach was also found to be high (median $r = .93$) (Hedge et al., 1990).

Rating scales. Included in this study were three different graphic rating scales which measured performance on a 5-point adjectivally anchored scale ranging from 1, with an anchor of "Never meets acceptable level of proficiency," to 5, with an anchor of "Always exceeds acceptable level of proficiency." Performance on tasks included in the hands-on work sample tests were measured by the Task Rating Form; each task was described by its statement from the Air Force Occupational Survey. Dimensional Rating Forms included four to six dimensions reflective of job-specific performance. Each scale contained anchors with specific behavioral examples of performance. The Global Rating Form consisted of two items for assessment of "Technical" and "Interpersonal" performance and was identical in form across jobs.

Aptitude. The Armed Services Vocational Aptitude Battery is a multiple aptitude test battery composed of ten subtests as shown in Table 1. The ASVAB is used by all the Armed Services for enlistment qualification and initial job assignment. It is normed on a weighted nationally representative sample of 18-to 23-year-old youths (Maier & Sims, 1986). The battery has been used in this subtest configuration since 1980, and is highly reliable (Palmer, Hartke, Ree, Welsh, & Valentine, 1988) and valid (Wilbourn, Valentine, & Ree, 1984).

Table 1

Measures Included in Analyses

<u>Performance Measures</u>	<u>ASVAB Subtests</u>
WTPT Scores:	General Science (GS)
Hands-on Test	Arithmetic Reasoning (AR)
Interview Test	Word Knowledge (WK)
Job Knowledge Test (JKT)	Paragraph Comprehension (PC)
Ratings (Self and Supervisor):	Numerical Operations (NO)
Task Rating Form	Coding Speed (CS)
Dimensional Rating Form	Auto Shop Information (AS)
Global-Technical Rating	Mathematics Knowledge (MK)
Technical Training Final Grade	Mechanical Comprehension (MC)
<u>Experience Measures</u>	Electronics Information (EI)
Time in Service (months)	
Task Experience Rating	
Number of Times Performed	
Last Time Performed (weeks)	

Note. ASVAB = Armed Services Vocational Aptitude Battery; WTPT = Walk-Through Performance Test. Standard scores on the ASVAB subtests were used.

Experience. Four measures of job experience were included in this study. Time in Service data were accessed from personnel records. The other three measures related specifically to work experience relevant to those tasks included in the WTPT and are self-report measures gathered as a part of the JPMS (see Table 1). These measures related the frequency of task performance (Number of Times Performed), recency of task performance (Last Time Performed), and self-ratings of the amount of on-the-job experience for each task in the WTPT (Task Experience Rating).

Procedure

In a group session, participants were introduced to the research project, participation conditions were explained, and participants were familiarized with each measure used in the project. This orientation was followed by one hour of frame-of-reference and rater error training (McIntyre, Smith, & Hassett, 1984). Immediately following rater training, raters completed a series of rating forms and questionnaires. Next, for four of the jobs, incumbents were administered job knowledge tests in a four-hour testing period.

The final testing stage, WTPT administration, occurred over several days at each site. Each incumbent was tested individually by a trained test administrator; administration usually required four to eight hours per incumbent, with test length dependent on the job. Task performance was measured with the hands-on method, the interview method, or both.

As recommended (Dickinson, 1989; Dunbar & Linn, 1986; Green,

1984; Wigdor & Green, 1986), these data were corrected for restriction in range (i.e., corrected for curtailment). A multivariate correction on the ten ASVAB subtests was used with 1980 youth population normative data (Mifflin & Verna, 1977). Except where noted, all analyses were performed on the corrected data.

Since the purpose of these analyses was identification of potential surrogate measures for use across Air Force enlisted occupations, all eight data sets were treated in the same manner. The exception was the presence of JKT data which were available for only the final four jobs included in the JPM Project.

A primary goal of this project was determination of the relationships among the JPM measures, especially how each of the surrogates compared to the hands-on measure. The main approach used was the examination of the hands-on score relative to each surrogate measure and sets of measures. The questions to be answered through these analyses of the JPMS data are:

1. Do the measures order people differently as to their technical proficiency?
2. What is being measured by each instrument?
3. Are the patterns of relationship to other related measures (i.e., aptitude, job experience) the same for the benchmark and surrogates?

The JPMS data were analyzed with correlation, regression, and factor analysis approaches. JPM measures were then studied in relation to other relevant information such as incumbent

aptitude and experience. Again, the focus was on identification of the relationships of these measures to Hands-on Test performance, then in comparison to the surrogates. Consistent findings across statistical approaches and occupations provide strong evidence that the measures are suitable surrogates for the Hands-on Testing.

Previous research findings indicated that the interview score should account for the largest amount of variance in the hands-on score (Hedge et al., 1990; Laue & Hedge, 1989). Although it is less costly to administer, Interview Testing may not be a practical substitute for Hands-on Testing since it is as equally expensive and time-consuming to develop. Thus, while the relationship between Interview Testing and Hands-on Testing was investigated in this study, emphasis was on the other more affordable alternate measures.

Results

Summary statistics on the corrected correlations between the hands-on score and each of the surrogate measures are in Table 2. Note that the highest correlation is with Interview Testing, followed by job knowledge testing. As expected, a strong relationship between the two work sample measures (i.e., Hands-on and Interview) was demonstrated across the eight jobs. Zero-order correlations between hands-on and interview scores ranged from .56 (Jet Engine Mechanic) to .88 (Personnel Specialist). Correlations with technical training grade and ratings are much lower, although statistically significant.

Table 2**Corrected Correlations Between Hands-On Score and JPM Variables**

Variable	Range	Mean r	Median r
Interview Test	.56 to .88	.73	.76
Job Knowledge Test	.50 to .80	.59	.53
Training Grade	.09 to .68	.40	.42
Task Ratings			
Self	.13 to .44	.29	.30
Supervisor	.14 to .51	.29	.29
Dimensional Ratings			
Self	.13 to .50	.31	.29
Supervisor	.13 to .59	.33	.33
Global Ratings			
Self	.05 to .58	.31	.32
Supervisor	.12 to .36	.25	.27

There appears to be a good deal of dispersion of correlations across the jobs, particularly for the ratings and training grade.

The first set of regression analyses used the hands-on score as the dependent variable and the surrogate measures as the independent variables. This method allowed the identification of the unique variance attributed to each surrogate. Then, the analyses were replicated with the exclusion of the Interview Testing score from the independent variable list.

Unique variance in the hands-on score accounted for by each individual predictor is indicated in Table 3. F -tests of the significance of the change in the R^2 of the model were performed and statistically significant results were found as noted. As expected, the interview measure consistently accounted for a significant amount of variance in the hands-on score, far more than any other single measure. Few other variables accounted for much variance in the dependent variable and there were no clear trends across occupations. No measure met the .90 criterion suggested by Wigdor and Green (1986).

Table 4 presents the results of the multiple regression with the exclusion of the interview data. The R^2 associated with the full model for seven jobs was considerably reduced from those presented previously. However, the WTPT for the Precision Measurement Equipment Laboratory career field contained only one interview task that was not also tested by the hands-on approach, thus limiting the amount of interview-unique variance. Across jobs, ratings rarely accounted for a significant amount of

Table 3

Percent Hands-On Variance Accounted for by the JPM Variables

Variable	Percent Variance			
	Aptitude Index			
	Mech	Admin	Gen	Elect
	JETENG	RADIO	ATC	AVCOM
Full Model	37.7**	64.1**	67.0**	72.7**
Interview	15.9**	43.6**	61.4**	18.8**
Training Grade	1.1	0.0	1.0**	2.4*
Task-Self	.6	.7	.2	.1
Task-Supervisor	0.0	.1	.1	2.1*
Dimensional-Self	1.1	.2	.5	.1
Dimensional-Supervisor	.1	.3	0.0	0.0
Global-Self	.1	1.7*	0.0	.1
Global-Supervisor	.8	.2	.3	.2
	AGE	PERSON	LIFESUP	PMEL
Full Model	61.2**	79.1**	38.4**	71.9**
Interview	21.9**	34.2**	11.8**	1.4*
JKT	.9*	0.0	2.6**	6.0**
Training Grade	.2	.5*	.2	.8
Task-Self	.1	.2	.1	.2

(table continues)

Table 3 (continued)

	Percent Variance			
	Aptitude Index			
	Mech	Admin	Gen	Elect
	AGE	PERSON	LIFESUP	PMEL
Task-Supervisor	.4	.1	0.0	.8
Dimensional-Self	.2	.1	0.0	0.0
Dimensional-Supervisor	1.0*	.2	0.0	0.0
Global-Self	.9*	.4	.2	.9
Global-Supervisor	.1	0.0	.5	1.5*

Note. Mech = Mechanical; Admin = Administrative, Gen = General; Elect = Electronic; JETENG = Jet Engine; RADIO = Radio Operator; ATC = Air Traffic Control; AVCOM = Avionic Communications; AGE = Aerospace Ground Equipment; PERSON = Personnel; LIFESUP = Life Support; PMEL = Precision Measurement Equipment Laboratory; JKT = Job Knowledge Test.

*p ≤ .05

**p ≤ .01

Table 4

Percent Hands-On Variance Accounted for by the JPM Variables
(Without Interview Test)

Variable	Percent Variance			
	Aptitude Index			
	Mech	Admin	Gen	Elect
	JETENG	RADIO	ATC	AVCOM
Full Model	21.8**	20.5**	5.6	55.0**
Training Grade	7.4**	2.6	.6	11.5**
Task-Self	1.6	1.3	0.0	.5
Task-Supervisor	.6	.2	0.0	1.6
Dimensional-Self	.6	.2	1.5	.7
Dimensional-Supervisor	0.0	.2	0.0	.6
Global-Self	0.0	1.5	0.0	0.0
Global-Supervisor	1.3	.2	0.0	.3
	AGE	PERSON	LIFESUP	PMEL
Full Model	39.3**	44.9**	26.6**	70.5**
JKT	4.5**	5.1**	19.2**	11.5**
Training Grade	.3	.7	.1	1.8**
Task-Self	.4	.9	0.0	.3
Task-Supervisor	.2	0.0	0.0	.6
Dimensional-Self	0.0	0.0	.2	.1

(table continues)

Table 4 (continued)

	Percent Variance			
	Aptitude Index			
	Mech	Admin	Gen	Elect
	AGE	PERSON	LIFESUP	PMEL
Dimensional-Supervisor	1.5*	1.4*	.3	.1
Global-Self	2.6**	2.9**	.7	.9
Global-Supervisor	.4	.2	.3	1.6**

Note. Mech = Mechanical; Admin = Administrative, Gen = General; Elect = Electronic; JETENG = Jet Engine; RADIO = Radio Operator; ATC = Air Traffic Control; AVCOM = Avionic Communications; AGE = Aerospace Ground Equipment; PERSON = Personnel; LIFESUP = Life Support; PMEL = Precision Measurement Equipment Laboratory; JKT = Job Knowledge Test.

*p ≤ .05

**p ≤ .01

variance and the training grade was only significant in three jobs. JKT, however, was significant for all four jobs.

Combinations or batteries of JPM variables were created and tested for potential as substitutes for Hands-on Testing in multiple regression procedures. The amount of unique variance attributable to each set of variables was assessed; analyses yielding a significant change in R^2 in 75% of the jobs are listed in Table 5. Again, analyses were conducted with inclusion and exclusion of the interview score. The amounts of R^2 explained by these sets of variables were small, with the sets including interview data predicting the largest amount of the variance in the hands-on score.

Principal factors analysis was selected for exploratory analysis of the internal structure of these data and construct validity. Factor analyses of the JPM variables were performed on the raw, uncorrected data in accordance with related JPM research efforts (e.g, Dickinson, 1989). The resulting factor solutions indicated that technical proficiency is represented by (a) Supervisory Ratings, (b) Self Ratings, and (c) Performance Measures (i.e., hands-on score, interview score, training grade, JKT). These findings demonstrate that these three groups of measures were tapping different underlying constructs, which appear to be very stable across jobs. The sole exception to this factor structure was in Air Traffic Control where the training grade loaded on the supervisory rating factor.

Finally, data were examined for relationships to other

Table 5

Regression with Sets of JPM Variables: Variance Accounted for in
Hands-On Dependent Measure

Sets of Variables	Percent Variance		
	Range	Mean	Median
With Interview Data			
Interview & Training Grade	3.12 - 62.01	29.22	26.78
Interview & JKT	12.86 - 39.31	27.38	28.68
Without Interview Data			
Training Grade & Global Ratings	.66 - 14.88	5.71	4.60
JKT & Supervisor Ratings	7.91 - 21.09	11.88	13.19
JKT & All Ratings	17.47 - 24.54	22.14	23.28
JKT & Training Grade	10.23 - 28.47	17.56	15.78
JKT & Task Ratings	5.58 - 19.63	11.62	10.63

Note. The sets of variables shown here were found to be significantly related to the hands-on dependent variable in at least 75 percent of the jobs studied (i.e., 3 of 4, 6 of 8); all other combinations tested were not statistically related to the dependent variable; JKT = Job Knowledge Test.

relevant variables (i.e., aptitude and experience variables) such that the pattern of relationships of hands-on to the other measures could be examined in relation to the surrogates' relationships with the other measures. Mean zero-order correlations among measures are displayed in Table 6.

In regard to the relationships between performance and aptitude measures, the technical training school grade showed the highest correlation with the ASVAB. This was expected since training performance is the criterion currently used for ASVAB validation. The relatively strong relationship between the ASVAB and JKT indicated on Table 6 had been anticipated due to similarity of the paper-and-pencil testing formats that capitalize on verbal and/or writing abilities. Both the hands-on and interview work sample tests related similarly to the ASVAB, exhibiting moderate correlations with the aptitude measures. Performance ratings did not appear to be consistently, nor significantly, related to the aptitude measures.

The correlations of experience and performance revealed strong similarities across hands-on and three surrogate measures (i.e., interview, JKT, and ratings). Time in Service was found to relate statistically to these four performance measures, indicating that time on the job relates positively to performance. Number of Times Performed also correlated significantly with these measures, suggesting that repeated practice increases performance. The self ratings of Task Experience were also consistently positively correlated with

Table 6

Patterns of Relationships Between Performance and Measures of
Aptitude and Experience

Measure	Mean <u>r</u>				
	Hands-on	Interview	JKT	Ratings	Grade
<u>Aptitude</u>					
ASVAB	.30**	.31**	.52**	.14	.61**
<u>Experience</u>					
Time in Service	.31**	.30**	.27**	.27**	-.03
Number of Times Performed	.27**	.29**	.26**	.26**	.01
Task Experience	.23*	.25**	.24*	.35**	-.03
Rating					
Last Time Performed	.09	.04	-.01	.09	-.02

Note. These data were not corrected for curtailment. ASVAB = Armed Services Vocational Aptitude Battery; JKT = Job Knowledge Test.

*p ≤ .05

**p ≤ .01

hands-on and the three above-mentioned surrogates. None of the experience measures were statistically related to training grades and Last Time Performed was not related to any of the performance measures.

The relationships of performance and measures of aptitude and experience can be briefly summarized. The findings were as follows: (a) the correlations between performance and aptitude measures were quite variable; (b) technical training grade did not relate to any experience measure; (c) a measure of recency of task performance, Last Time Performed, was not related to any of the performance measures; and (d) Experience-Surrogate relationships were quite consistent across the remaining three experience measures, and three surrogates generally paralleled the correlations between hands-on and experience.

Discussion

Results of these analyses provide evidence that hands-on work sample testing provides data that are somewhat independent of other job performance measures, job experience variables, and aptitude data. Although statistically significant results were found in many of the comparisons between the hands-on and surrogate measures, the impact of many of these results is minimal. That is, only small portions of variance in the hands-on measure are explained by a relationship with the surrogates. This was especially true for the ratings by self and supervisor.

The lack of support for the substitutability of the

alternate measures does not negate the value of research in this area. None of the JPM surrogates measures met the criterion of a .90 correlation with Hands-on Testing, making the analysis of other aspects more important (e.g., factor analysis, relationships with aptitude and experience measures). Results of these analyses rather consistently indicate that:

1. None of the surrogates are interchangeable with the Hands-on Testing of job performance.
2. Technical proficiency is represented by three distinct factors (i.e., self ratings, supervisor ratings, performance on work sample and job knowledge tests)
3. There may be systematic differences between job groupings (i.e., aptitude composites) with regard to feasibility of JPMS development and suitability for implementation.

Note that the two jobs with the highest percent of variance accounted for were Avionics Communication Specialist and Precision Measurement Equipment Laboratory Specialist, the two Electronic AI jobs (see Table 4). These data indicate that these types of jobs are possibly better suited for this approach to criterion development. The Electronic AI jobs exhibited the highest R^2 values for the regression of the hands-on measure on the surrogates when interview testing scores were omitted from

the regression equation. These two jobs also showed the strongest ASVAB and hands-on correlational relationships. Results such as these suggest that the JPMS approach may be better suited for some jobs and less appropriate, or even inappropriate, for others.

It is important to note a high level of confidence in the first two conclusions (i.e., 1 and 2 above) since they are based on numerous and varied analyses; however, the final conclusion is highly speculative and worthy of additional research. Recent work by Hunter, Schmidt, and Judiesch (1990) focused on the differences in complexity among job groups. Their findings indicated that worker output is related to complexity of jobs, resulting in higher validity of selection procedures and greater payoff for the organization in the selection of highly complex jobs. Thus, these present data may be reflective of the general conclusions of the Hunter et al. (1990) research.

Two approaches were found to have potential as surrogates for Hands-on Testing: interview testing and job knowledge testing. Both of these measures were related to Hands-on Testing in regression and correlation studies and yielded similar profiles of correlations with measures of experience and aptitude. Also, factor analyses indicated that these measures load together with the hands-on measure on the same factor, indicating that they measure similar underlying constructs which are independent of the other performance measures.

While JKTs or Interview Tests may prove feasible for future

research and development, it is important to note that a good deal of variance in the hands-on score remains unexplained. That is, there is job performance data elicited uniquely by the hands-on approach that is not measured by any single surrogate measures or combinations of measures. Surrogates tend to test procedural knowledge and general understanding and not whether the person can actually perform the job (Green, 1984). An ideal testing battery might rely heavily on the use of selected surrogates and include an abbreviated hands-on performance measure for certain tasks that cannot be evaluated by alternative means. In this manner, the job domain can be more completely covered to include a representative sample of job tasks.

Research on the systematic differences between job groupings should center on identifying key indicators that serve to differentiate them. Aptitude of job incumbents may be one factor; both Electronic jobs had a classification cutoff on the Electronic composite which was quite high relative to others included in the JPM Project. Research into the relationships of these indices to general intelligence and reading level, for example, may yield a greater understanding of the general conditions that facilitate development and applicability of performance measures.

Another area that may impact the development of performance measures is related to the technical requirements of task performance. Both were supported by well written, detailed technical orders, manuals, and resources. The highly technical

and precise nature of the Electronics field may lend itself well to the step-level performance evaluation of the WTPT and JKT. Thus, these may be best suited to jobs of this type. As noted by Gottfredson (1986), different measures may be required for different classes of jobs due to systematic differences in job demands, and these data tentatively support this hypothesis.

One final area of future research on the assessment of substitutability of JPM instruments focuses on the types of selection and classification decisions made when using Hands-on Testing versus surrogates for setting aptitude standards. The decision-making usefulness of these measures is a key question. If both the benchmark and surrogate lead to the same personnel decisions, then they may be considered equivalent in that context. Future research should examine the outcomes of using the various JPMS components in setting aptitude requirements and other decisions.

References

- Cohen, J., & Cohen, P. (1975). Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum.
- Dickinson, T. L. (1989). Structure of the Air Force's Job Performance Measurement System and predictability of the Armed Services Vocational Aptitude Battery (UES Report 788-039-004). San Antonio, TX: Universal Energy Systems, Inc. Prepared under Contract #F41689-86-D-0052, Air Force Human Resources Laboratory, Brooks AFB, TX.
- Dunbar, S. B., & Linn, R. L. (1986). Range restriction adjustments in the prediction of military job performance. Prepared for the Committee on the Performance of Military Personnel, Commission on Behavioral and Social Sciences and Education (National Research Council/National Academy of Sciences).
- Finn, J. D. (1974). A general model for multivariate analysis. New York: Holt, Rinehart, and Winston.
- Gottfredson, L. S. (1986). The evaluation of alternative measures of job performance. Paper prepared for the Committee on the Performance of Military Personnel, Commission on Behavioral and Social Sciences and Education, National Research Council/National Academy of Sciences.
- Green, B. F., Jr. (1984). Measure surrogates and the problem of substitutability. Unpublished document.
- Hedge, J. W., & Teachout, M. S. (1986). Job Performance

- Measurement: A systematic program of research and development (AFHRL-TP-86-37). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.
- Hedge, J. W., Teachout, M. S., & Laue, F. J. (1990). Interview testing as a work sample measure of job proficiency (AFHRL-TP-90-61). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.
- Hunter, J. E., Schmidt, F. L., & Judiesch, M. K. (1990). Individual differences in output variability as a function of job complexity. Journal of Applied Psychology, 75, 28-42.
- Kraiger, K. (1989). Generalizability theory: An assessment of its relevance to the Air Force job performance measurement project (AFHRL-TP-87-70, AD-A207-107). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.
- Kraiger, K. (1990). Generalizability of Walk-Through Performance Tests, job proficiency ratings, and job knowledge tests across eight Air Force specialties (AFHRL-TP-90-14). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.
- Laue, F. J. (1990). Interim report on R & D for substitutability analyses of job performance measures (UES Report 788-040-1). San Antonio, TX: Universal Energy Systems, Inc. Prepared under Contract #F41689-86-D-0052, Air Force Human Resources Laboratory, Brooks AFB, TX.

- Laue, F. J., & Hedge, J. W. (1989). The evaluation of alternate measures of job performance (UES Report 788-029-1). San Antonio, TX: Universal Energy Systems, Inc. Prepared under Contract #F41689-86-D-0052, Air Force Human Resources Laboratory, Brooks AFB, TX.
- Lipscomb, M. L. (1984). A task-level domain sampling strategy: A content valid approach. Paper presented at the annual meeting of the American Psychological Association, August, Toronto.
- Lipscomb, M. S. (1987). A task-level domain sampling strategy: A content valid approach. In J. W. Hedge and M. S. Lipscomb (Eds.), Walk-Through Performance Testing: An innovative approach to work sample testing (AFHRL-TP-87-8). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.
- Maier, M. H., & Sims, W. H. (1986). The ASVAB score scales: 1980 and World War II (CNR 116). Alexandria, VA: Center for Naval Analyses.
- McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. Journal of Applied Psychology, 69, 147-156.
- Mifflin, T. L., & Verna, S. M. (1977). A method to correct correlation coefficients for the effects of multiple curtailment (CRC 336). Arlington, VA: Marine Corps Operations Analysis Group, Center for Naval Analyses.

Prepared for Office of Naval Research.

Palmer, P, Hartke, D. D., Ree, M. J., Welsh, J. R., & Valentine, L.D. Jr. (1988). Armed Services Vocational Aptitude Battery (ASVAB): Alternate form reliability (Forms 8, 9, 10, and 11) (AFHRL-TP-87-48, AD-A191 658). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Wigdor, A. K., & Green, B. F., Jr. (1986). Assessing the performance of enlisted personnel: Evaluation of a Joint-Service Research Project. Washington, DC: National Academy Press.

Wilbourn, J. M., Valentine, L. D., Jr., & Ree, M. J. (1984). Relationships of the Armed Services Vocational Aptitude Battery (ASVAB): Forms 8, 9, and 10 to the Air Force technical school final grades (AFHRL-TP-84-8, AD-A144 213). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.