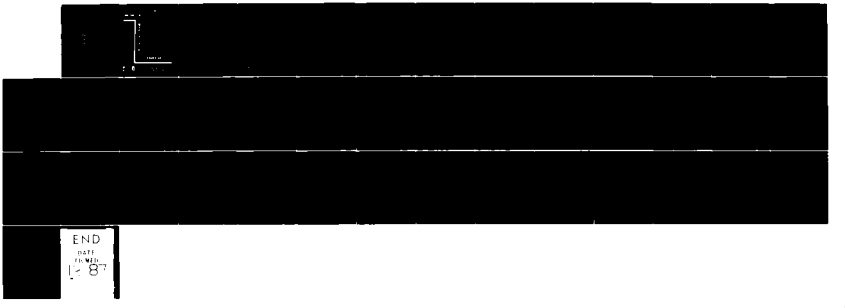
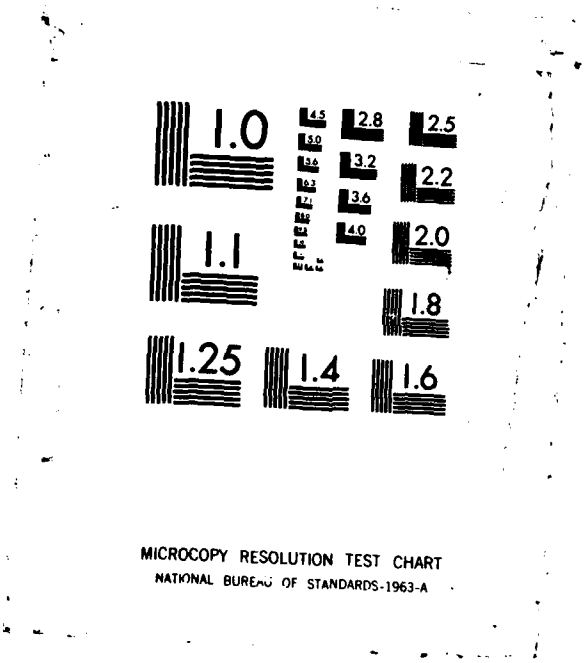
AD-A185 479 WALK-THROUGH PERFORMANCE TESTING: AN INNOVATIVE APPROACH TO WORK SAMPLE TESTING(U) AIR FORCE HUMAN RESOURCES LAB BROOKS AFB TX J W HEDGE ET AL. SEP 87 1/1

UNCLASSIFIED AFHRL-TP-87-8 F G 5/9 NL



END
DATE FILMED
1-88



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AFHRL-TP-87-8

(12)

AIR FORCE HUMAN RESOURCES LABORATORY

WALK-THROUGH PERFORMANCE TESTING: AN INNOVATIVE APPROACH TO WORK SAMPLE TESTING

Jerry W. Hedge
M. Suzanne Lipscomb

TRAINING SYSTEMS DIVISION
Brooks Air Force Base, Texas 78235-5601

September 1987

Interim Technical Paper for Period May 1983 - August 1984

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

NOTICE

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 Public Affairs Office 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.

HENDRICK W. RUCK, Technical Advisor
Training Systems Division

GENE A. BERRY, Colonel, USAF
Chief, Training Systems Division

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE

AD-A185 479

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) AFHRL-TP-87-8	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Training Systems Division 6b. OFFICE SYMBOL (If applicable) AFHRL/IDE	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code) Air Force Human Resources Laboratory Brooks Air Force Base, Texas 78235-5601	7b. ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Air Force Human Resources Laboratory 8b. OFFICE SYMBOL (If applicable) HQ AFHRL	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State, and ZIP Code) Brooks Air Force Base, Texas 78235-5601	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 62703F; PROJECT NO. 7734; TASK NO 08; WORK UNIT ACCESSION NO 22

11. TITLE (Include Security Classification)
Walk-Through Performance Testing: An Innovative Approach to Work Sample Testing

12. PERSONAL AUTHOR(S)
Hedge, J.W.; Lipscomb, M.S.

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Interim	FROM May 83 TO Aug 84	September 1987	44

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	criterion development; work sample; job performance; performance measurement
05	09		
05	10		

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

Work sample tests have been advocated as reliable, valid measures of job proficiency. Work sample procedures normally identify critical tasks, discard those tasks not practically observable and measure the remainder with hands-on tests. For the Air Force, hands-on testing is a particular problem because of the complexity and expense involved in performing many tasks. Walk-Through Performance Testing (WTPT) is being developed to expand the range of job tasks measured, to include tasks that do not lend themselves to hands-on testing. WTPT is a task-level job performance measurement system which combines hands-on task performance and interview testing procedures to provide a high-fidelity measure of an individual's technical job competency. The addition of an interview testing component allows the test administrator to evaluate by means of a "show-and-tell" procedure individual competencies on tasks not measurable by the hands-on process. This technical paper details the background, theoretical basis, and development procedures involved in WTPT. (Keywords: Air Force research; work measurement).

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION
22a. NAME OF RESPONSIBLE INDIVIDUAL Nancy J. Allin, Chief, STINFO Office	22b. TELEPHONE (Include Area Code) (512) 536-3877 22c. OFFICE SYMBOL AFHRL/TSR

DD Form 1473, JUN 86 Previous editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

AFHRL Technical Paper 87-8

September 1987

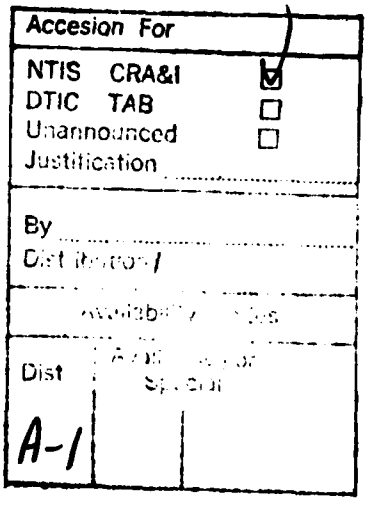
WALK-THROUGH PERFORMANCE TESTING: AN INNOVATIVE APPROACH TO WORK SAMPLE TESTING

Jerry W. Hedge
M. Suzanne Lipscomb

TRAINING SYSTEMS DIVISION
Brooks Air Force Base, Texas 78235-5601

Reviewed and submitted for publication by

Rodger D. Ballentine, Lt Col, USAF
Chief, Skills Development Branch
Training Systems Division



SUMMARY

As part of an extensive job performance measurement research and development program, the Air Force Human Resources Laboratory has developed a new methodology called Walk-Through Performance Testing (WTPT). WTPT is a task-level job performance measurement system which combines hands-on testing and interview testing to provide a high-fidelity measure of an individual's technical job competency. This document contains a series of papers, originally presented at the 92nd Convention of the American Psychological Association, which outline: (a) the conceptual frame-of-reference within which the original planning for WTPT took place, (b) the WTPT methodology and the rationale and overall approach to hands-on and interview test development, (c) the sampling strategy used to select tasks for work sample development, and (d) the approach used to analyze selected tasks for WTPT development. A final section discusses the implications of the measurement strategy.

PREFACE

The Air Force Human Resources Laboratory (AFHRL) is engaged in a long-term research and development effort to develop criteria for validation of Air Force selection and classification procedures. Both work sample and rating forms at varying levels of specificity are currently being developed. The work sample measurement approach being developed and evaluated by AFHRL is the topic of this technical paper.

The basic content of these four papers was presented originally at the 92nd Annual Convention of the American Psychological Association in Toronto, Canada. The symposium was chaired by Dr. Sheldon Zedeck (University of California at Berkeley). Dr. Terry Dickinson (Old Dominion University) served as symposium discussant. A paper based on comments provided at the symposium by Dr. Dickinson is included in the report.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

WALK-THROUGH PERFORMANCE TESTING:
AN INNOVATIVE APPROACH TO WORK SAMPLE TESTING

INTRODUCTION

This series of papers presents a newly developed performance measurement methodology, and provides a detailed explanation of the rationale, developmental process and potential payoffs of such a technology. This new approach, Walk-Through Performance Testing (WTPT), is being developed to expand the range of job tasks measured, to include tasks that do not lend themselves to hands-on testing. The first paper, by Gould and Hedge, outlines the general conceptual performance measurement frame-of-reference within which the original planning for WTPT took place. In the next paper, Hedge discusses in detail the WTPT methodology, and the rationale and overall approach to hands-on and interview test development. Lipscomb's paper follows with an explanation of the sampling strategy used to select representative tasks for work sample development, an initial step in the development process. Next, Ballentine and Lipscomb present the approach used in task analysis for WTPT development. Finally, Dickinson provides comments on these papers and discusses implications of this new measurement strategy. Taken together, these papers present the most comprehensive discussion of the WTPT process to date.

HISTORY, BACKGROUND, AND THEORETICAL BASES OF WALK-THROUGH PERFORMANCE TESTING

R. Bruce Gould
and
Jerry W. Hedge

Air Force Human Resources Laboratory

This paper describes the Air Force Human Resources Laboratory's (AFHRL) research and development (R&D) program for development of individual job performance measures. The job performance measurement literature indicates that most previous efforts have used broad-based generic indices, performance ratings, or operational measures, with their inherent problems of inflation and halo effects. These broad measures were unable to take into account task-level-specific influences such as training differences or differences in opportunities to perform; hence, such efforts have been largely unsuccessful. However, it appears that current interest, added resources, and technology developments have now significantly increased the probability of developing successful measures of job performance to be used as criteria in evaluating manpower, personnel, and training programs.

Several influences have highlighted the Air Force's need for job performance measurement and brought ongoing and planned programs to their current state. Planning for the R&D program began 3 years ago, on the recommendation of an AFHRL Research Advisory Panel (composed of knowledgeable scientists from academia and industry, as well as peers from the Army and Navy).

The panel reviewed the entire AFHRL manpower, personnel, and training R&D program and recommended consolidation of separate job performance measurement efforts into a single unified R&D program. At the same time, the Uniform Guidelines for Employee Selection (1978) and a review of case law mandated that Air Force civilian selection systems be validated against job performance measures. Military tests are exempted from this legal mandate by the Office of Management and Budget (OMB), but OMB has been reviewing that exemption. Finally, Congress mandated that military selection tests be validated against hands-on job performance measures. These operational, legal, and Congressional imperatives have thus provided the impetus to planning and obtaining support for a lengthy, high resource R&D effort.

Twenty years of extensive occupational R&D and a commitment of significant resources provided the backdrop, data base, and means to solve a portion of the "criterion problem" for Air Force researchers and program evaluators. In addition, the Air Force has now completed the second year of a 7-year R&D effort to systematically obtain job performance measures that will serve as criteria in validating selection systems and in evaluating training programs and the effects of personnel policies and procedures. Previous R&D concerning Air Force occupations has identified the major job tasks in enlisted specialties, the types of individuals who perform them, the relative difficulties in learning to perform them, and the relative aptitude requirements of the tasks. This occupational data base provides the initial reference point for identifying job tasks to be measured, as well as objective indices of moderator variables such as task-level experience which otherwise would contribute error variance to the measurement of job performance.

A conceptually based theoretical framework of performance measurement is presented to summarize and organize research progress in terms of previous empirical work and to identify future R&D needs. The present program is unique in its "research purposes only" orientation, its concentration on individual-job-specific tasks rather than tasks common to all jobs in a specialty, and its consideration of different types of measures (job sample testing; objective indices of productivity; and supervisory, peer, and self ratings) as tapping both overlapping and unique components of the job performance criterion space. A novel job measurement approach called "walk-through performance testing" is used as the high-fidelity benchmark against which

less time-consuming and expensive procedures can be compared through a successive approximation research strategy.

The Air Force job performance measurement R&D plan has not been developed in isolation from the other Services. We served with representatives from the Navy Personnel Research and Development Center (NPRDC) on the Army Research Institute's (ARI) Armed Services Vocational Aptitude Battery (ASVAB) validation contract evaluation panel. We also sponsored an informal tri-Service workshop on job performance evaluation. We are now working with the other Services to coordinate a Joint-Service Job Performance Measurement Program. In effect, the Services are pooling their resources by dovetailing research plans and sharing results.

The short-term objective of the Air Force's job performance measurement program is the development of on-the-job performance measures to validate Air Force selection and classification procedures. Guidelines for developing and obtaining the performance measures will be established for a wide range of enlisted, officer, and civilian jobs. Once obtained, the measures will be placed in a data base for test validation use.

The long-term goal is to establish an operational performance measurement program for the evaluation of selection and training procedures and personnel policies and practices; that is, to operationalize procedures such that the performance measurement, validation, and evaluation can be carried on by technicians. In this way, R&D resources will be freed for other projects.

Conceptual Performance Measurement Model

The first step in the present effort was to develop a conceptually based descriptive model of performance measurement that could be used to summarize and organize R&D progress in terms of previous empirical work, and to identify and prioritize future research needs for the program. Time will not permit a detailed examination here of model development, results of the literature review, and research issues to be studied. These details were described in a separate report (Kavanagh, Borman, Hedge, & Gould, 1987). The development process, resulting model, and some general conclusions will, however, be outlined.

Guidelines established for model development were that the model should: (a) focus on performance measurement used in the military; (b) describe performance measurement used for "research purposes only"; (c) consider all variables, based on the theoretical or empirical literature, that could affect job performance or performance measurement; (d) use a general classical test score theory perspective for identifying sources of true and error variance in observed scores; (e) specify classes of variables rather than detailed individual variables; (f) be descriptive rather than prescriptive because tested causal linkages to job performance are as yet too incomplete; and (g) use an iterative process that begins with a general model of job performance and ends with a model of measurement quality where the measures are to be used as criteria in validation or evaluation projects only.

First, a general conceptually and empirically based model was developed which identified individual characteristics that may interact with supervisory and work group factors to influence job performance. Organizational factors and situational constrains were also included. Next, a restriction was imposed on the general model that only factors affecting the quality of performance measurement would be considered, and a second more detailed model resulted. The emphasis had thus shifted to include not only a person's job performance but also the measurement method used to record that performance, as well as the characteristics of the person recording the scores. The experience of job incumbents, including opportunities to perform, remained a prominent factor. A reasonably exhaustive list of the variables that have been empirically demonstrated to affect, or potentially affect, performance measurement quality was compiled at this point and is shown in Table 1. One of the critical variables identified is the measurement purpose.

Table 1. Variables That Can Impact Measurement Quality

1. Individual characteristics
   a. Cognitive variables: rater or ratee
   b. Rater/ratee intelligence
   c. Rater/ratee knowledge of the job being evaluated
   d. Rater/ratee personal characteristics
   e. Rater/ratee interpersonal trust

2. Relationship between ratee and rater/observer
   a. Sex congruence
   b. Race congruence
   c. Job tenure together
   d. Age congruence
   e. Off-the-job relationship
   f. History of conflict or cooperation

3. Method/Source of measurement
   a. Supervisor ratings
   b. Peer ratings
   c. Self ratings
   d. Subordinate ratings
   e. Assessment center (team) ratings
   f. Work samples/simulations
   g. Productivity records

4. Scale development
   a. Critical incidents used
   b. Job description/job requirements-based
   c. Employee participation
   d. Top management support during development

5. Rating scale characteristics
   a. Content of the scale
   b. Anchors versus no anchors
   c. Behaviors versus traits
   d. Format type
   e. Number of anchors/scale points
   f. Single versus multiple dimensions
   g. Scaling metric/approach

6. Performance standards/goals
   a. Present or not
   b. Standards versus goals
   c. Participatively set and communicated
   d. Specificity of behavior or accomplishment excepted

7. Social context
   a. Performance level of others in work group
   b. Existence of group norms
   c. Rater's status in group
   d. Ratee's status in group

Table 1. (Concluded)

8. Non-work variables
   a. Marital status
   b. Pre-school children at work
   c. Dual-career family
   d. Participation in company activities off the job
   e. Stressful life events in recent past

9. Performance constraints
   a. Poor information
   b. Equipment efficiency
   c. Supplies deficiency
   d. Time limitations
   e. Poor work environment

10. Organizational/unit norms
   a. Upper management's expectation of certain level of performance
   b. Immediate supervisor's expectation regarding level of performance
   c. Presence of a union
   d. Pay/rewards tied to performance levels by contract
   e. Pay/rewards tied to performance levels by informal norms

11. Public relations/administrative procedures
   a. Required or not
   b. Mode of presentation
   c. Content of procedure

12. Rater training
   a. Content of training
   b. Format of training
   c. Length of training

13. Measurement purpose
   a. Validation research only
   b. Employee growth and development
   c. Administrative purposes such as rewards
   d. Meeting legal guidelines

14. Performance feedback
   a. Required or not
   b. Sources of feedback
   c. Participative
   d. Clarity of feedback
   e. Frequency of feedback

15. Pay-performance relationship
   a. Are they related in the system?
   b. Equity of the relationship

Performance measurement systems have four major purposes: (a) administrative decisions, (b) employee growth and development, (c) validation/evaluation research, and (d) meeting legal guidelines. Since our purpose was to obtain performance measures for validation R&D, we eliminated those model components that were not related to that outcome. System characteristics

related to performance feedback and pay-for-performance relationships were eliminated. The resulting model is shown in Figure 1.

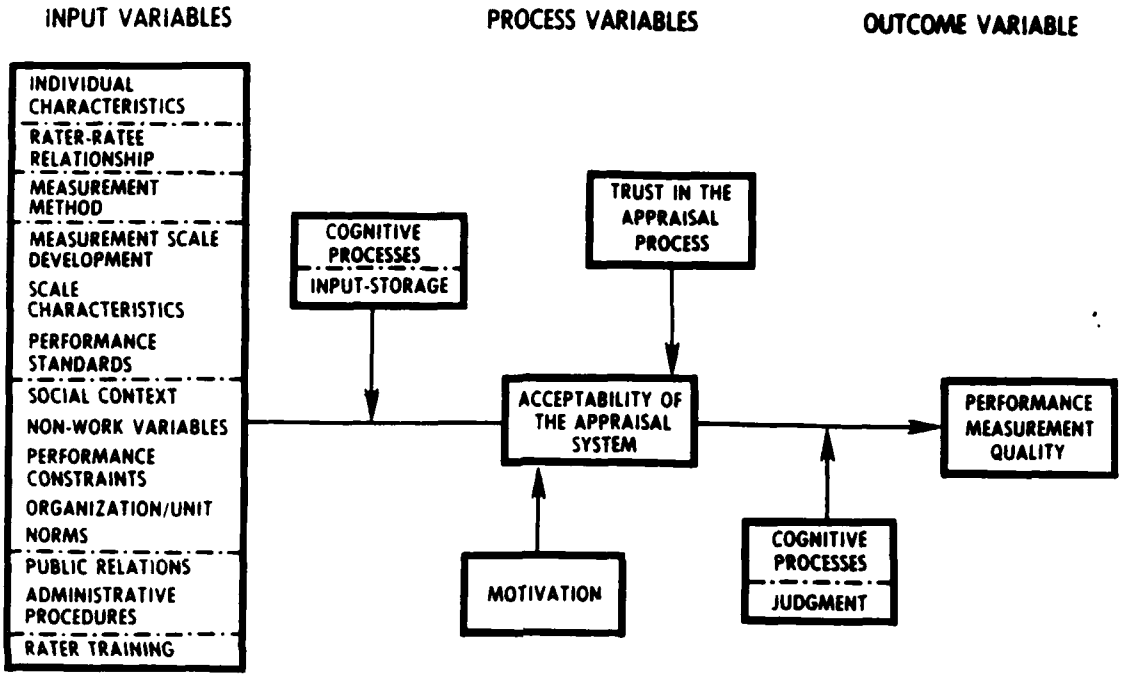


Figure 1. A Job Performance Measurement Classification Scheme for Validation Research.

The model categorizes the measurement system characteristics on the left; the outcome, performance measurement quality, on the right; and potential intervening variables in the center.

Regarding Figure 1, an analogy would be to consider the system characteristics as independent variables, the intervening variables as moderators, and measurement quality as the dependent variable in a multiple regression equation. One would expect the beta weights to change for the various terms in the equation as the type of measurement method or type of job being measured changes. These categories of independent and intervening variables provided a classification scheme by which the empirical literature was organized and the R&D issues identified. In establishing a connection between a system characteristic and the outcome (accuracy), the relevant literature gives specific implications for the measurement system. Accuracy and construct validity are the primary criteria by which the dependent measure, quality of performance measurement, must be assessed. The reasons for including intervening variables in the center of the model should become more apparent following a discussion of the nature of the measurement systems to be used in this effort.

Domain to be Measured Versus Measurement Method

Let us now consider the conclusions drawn between the relationship of measurement method and the validity of the resulting measures and how they were translated into a proposed approach.

Typically, performance measures have suffered varying degrees of criterion deficiency; that is, the measures did not adequately sample the domain of tasks to be performed in specific jobs. Either the measures were not based on an adequate job analysis, or the representative tasks could not be economically measured by the method employed.

Definition of job tasks is not a problem in the Air Force because of its active Occupational Survey Program, which provides current task-level data on all major enlisted specialties. More than 200 of the some 250 enlisted specialties are re-surveyed on an average of every 4 years. The surveys provide task-level measures of time spent performing, time required to learn to perform, and relative aptitude requirements, as well as how the tasks are organized into homogeneous clusters. The clusters of tasks which are performed in concert are called job types.

Identification of critical tasks in a specialty is not difficult, but selection of tasks for the measurement system immediately poses the question of just what "job" means in job performance. For example, Figure 2 shows a typical situation where there are relatively unique as well as overlapping job types within an Air Force specialty. Here, two job types share many common tasks, most of which are critical tasks--but a third job type shares very few tasks common to the other two job types. Thus, although performance can be measured at the job level, how do you aggregate across job types to make a collective statement about performance of individuals at the specialty level?

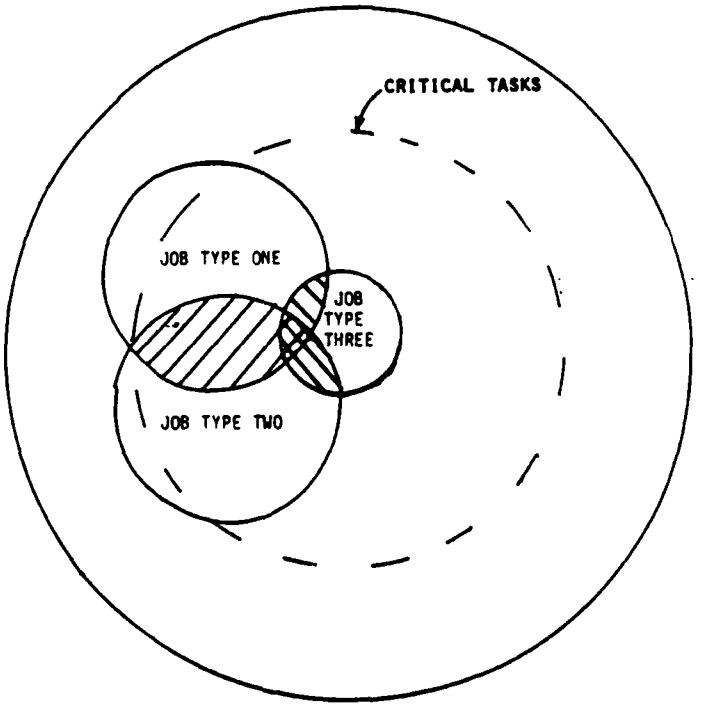


Figure 2. Job Performance Domain.

The solution to this problem for the Air Force is to measure both a sample of core specialty-wide critical tasks and a sample of job-type-specific tasks. Task-level experience measures will also be collected to assess the impact of differences in opportunities to perform, such that individuals can be compared on the specialty-wide set of core tasks, taking experience differences into account. However, in order to compare performance across job types as a means of drawing specialty-wide conclusions, a benchmarking/equating strategy must be developed.

For the Air Force, hands-on testing constitutes a particular problem because of the nature of Air Force equipment. In many specialties, the number of tasks that can be measured by hands-on testing is low, in that the tasks tend to take too long to complete, require replacement of expensive parts, or present possible damage to system components. Thus, a walk-through testing routine was devised to overcome the stated problems. WTPT is a task-level job performance measurement system which combines task performance and interview procedures to provide a high-fidelity measure of individual technical job competence. The test takes place in the work setting and is a type of show-and-tell procedure.

Development of the Performance Measures and the Measurement Process

The performance measures to be developed range from the micro task-level walk-through test to very macro, global ratings. From the occupational survey data, specialty- and job-type-specific critical core tasks will be identified. Subject-matter experts (SMEs) will aid the developers in dichotomizing the tasks into those which are economically observable and those which must be measured by interview. The SMEs will then develop the procedures for conducting the observations and interviews, as well as specify the performance standards for scoring responses. For a given specialty, a formal testing manual similar to, say, that for an individually administered intelligence test will be developed in two sections; i.e., a specialty core task section and a job-type-specific section. There will be an overlap between tasks measured by interview and hands-on testing, to ensure that the two methods demonstrate similar measurement fidelity and can be used to fully sample the domain of technical competency skills.

Before testing, the job incumbent will complete the self-ratings of experience and proficiency. The examiner will then take the ratee to the work site and administer the walk-through test, having the ratee point out the procedures, explain component functioning, or actually perform tasks. Target test administration time is 8 hours.

The detailed rating forms will have a one-to-one task correspondence with the tasks in the walk-through testing manual. In addition, the other forms will contain behavioral dimensions such as ability to perform troubleshooting or administrative tasks, global performance ratings such as technical competence and interpersonal/social skills, and ratings that apply Air Force-wide such as the Leadership and Technical Knowledge factors found in efficiency reports.

Other Research Issues

Considering the nature of the walk-through testing procedure, the reason for the concentration on the intervening variables in the center of Figure 1 should now be more apparent. Once the measures are developed, the value of the data obtained is going to be largely a function of the training, motivation, and ability of the rater/ratee. For this reason, much of our specific research must be aimed at these areas. In terms of system characteristics, the rater-ratee relationship, organization/unit norms, and public relations issues will initially be minimized by the use of professional examiners. However, later in the project, when non-professional examiners will be considered as possible test administrators, these issues will have to be addressed.

The measurement method issues have been discussed in an earlier part of this paper. Scale characteristics and measurement development issues have been largely resolved in the literature. Concerning specification of performance standards, we are confident (based on past experience) that, with our guidance, SMEs will be able to develop valid standards. Concerning administration procedures and rater training, there are many unanswered questions; therefore, we will concentrate much of the early work in these areas. Although the studies will not be detailed in

this series of papers, the basic mechanism we plan to use may be of interest. Most similar studies are plagued with using estimates of true performance as criterion measures. We plan to write scripts and videotape airmen performing actual tasks and then use the videotapes as the medium for presenting the rating situations during our research. This way we will know the true performance score, and the tapes will have both face and content validity for the experiments to be performed.

THE METHODOLOGY OF WALK-THROUGH PERFORMANCE TESTING

Jerry W. Hedge

Air Force Human Resources Laboratory

The purpose of this paper is to describe the development of a new assessment methodology known as WTPT. In discussing this approach, four main objectives of the Air Force Human Resources Laboratory's (AFHRL) job performance measurement R&D effort will be emphasized. These four objectives are: (a) to develop a measurement methodology that allows accurate evaluation of job proficiency, (b) to develop a measurement technique that expands coverage of the job content domain, (c) to evaluate the comparability of information gathered through the two components of WTPT (hands-on and interview testing), and (d) to adapt the WTPT methodology to a variety of Air Force specialties. The rationale and approach associated with each of these objectives will be detailed in the following sections of this paper.

Background

In assessing on-the-job performance, a variety of measures are available from which to choose. They range from subjective to objective, and from general to specific. When faced with the choice of which criterion to select, the researcher or practitioner typically relies on several informal decision rules:

1. Cost, in terms of time, money, safety, or mission effects.
2. Convenience in developing or obtaining measures.
3. Fidelity or accuracy of replicating behaviors relevant to the job.

The development of a criterion measure is frequently seen as a secondary concern to the main research focus (i.e., a training program, a selection system); as a result, Decision Rule 2--convenience--is frequently applied. The outcome is thus a generic, packaged-to-please rating form that in all probability will also satisfy Decision Rule 1, but not necessarily Decision Rule 3.

When the chief concern of researchers and practitioners shifts to Decision Rule 3, and fidelity becomes the overriding concern, the work sample orientation presents a viable alternative to the convenient but subjective rating form. As noted by Wilson (1962), over the years the primary use of the work sample has been for personnel selection; however, this orientation can also be a valuable aid in the measurement of job proficiency. Typically, work sample tests involve an individual's performing a task or set of tasks that are relevant to that person's job and are selected from the range of tasks performed by that person.

The value of the work sample methodology lies in the fidelity with which the selected set of tasks allows measurement of an incumbent's job proficiency. Unfortunately, the fact that tasks must be "selected" reflects the technique's chief weakness. Work sample procedures normally identify critical tasks, discard those not practically measurable, and let the remainder become the "selected set" of measured tasks. The Air Force's approach to work sample testing represents an attempt to overcome this criterion deficiency problem.

Walk-Through Performance Testing

For the Air Force, hands-on testing presents a particular problem because of the complexity and expense involved in performing many tasks. For example, many critical tasks cannot be

measured by hands-on testing because these tasks tend to take too long to complete, require replacement of expensive parts, and risk possible damage to components. AFHRL has developed a new methodology to deal with these problems. This new approach, WTPT, has as its foundation the work sample philosophy, but attempts to expand the measurement of critical tasks to include those tasks not measured by hands-on testing through the use of an interview testing component (Gould & Hedge, 1983).

WTPT is a task-level job performance measurement system that expands the range of job tasks on which an individual is measured, by combining hands-on task performance and interview procedures to provide a high-fidelity measure of individual technical job competence. The interview testing component has been added as a means of assessing those critical tasks previously eliminated from the content domain because of measurement constraints.

Overview of the Developmental Process

Details of the "Task Selection Plan" used to define the job content domain are presented in a separate paper by Lipscomb, which follows. At this time, however, two aspects of this selection process should be noted.

First, using this sampling strategy, tasks are classified into three major phases, based on measurement specificity. This allows work samples to be developed for tasks that are common to all jobs in the specialty (Phase I) or unique to a particular duty area or job type (Phases II & III). Secondly, no differentiation is yet made between tasks to be measured by hands-on versus interview testing. However, as part of this process, information is being collected concerning length of time required to perform each task and whether the task is measurable by hands-on testing. Once the job content universe has been reduced, and field interviews with SMEs initiated, those decisions will be made. Information about this process can be found in the Ballentine and Lipscomb paper contained in this volume.

WTPT Components

As noted previously, WTPT consists of two main components, hands-on testing and interview testing. The hands-on component resembles a traditional hands-on work sample test designed to measure proficiency on a critical task that has survived the imposition of time and/or measurement constraints. For example, the hands-on task outlined in Table 2 requires the incumbent to install a starter on the jet engine. On the first page of the task item, information is provided to the test administrator concerning testing time; required tools, technical orders, and job guides; pertinent background information and necessary engine configuration; and administrator's testing instructions. While the starter is being installed, the test administrator uses the checklist to indicate whether steps (e.g., lubricate the spline, index position of the starter, and install the locking device) are performed correctly or not. Finally, a 5-point rating scale is provided so an overall rating of proficiency on that task can also be recorded.

Interview testing allows the administrator to measure proficiency on tasks precluded from hands-on measurement (i.e., tasks that are either too time-consuming, too costly, or too dangerous for hands-on measurement). Interview testing requires the administrator to assess an incumbent's proficiency on a task by asking questions designed to uncover proficiency-based strengths and weaknesses related to the performance of that task. For example, the interview test item in Table 3 evaluates the jet engine mechanic's ability to determine the source of high oil consumption. Using categories similar to those for hands-on items, the test administrator asks the incumbent to show/explain procedures and provide answers to a variety of questions. In addition, a 5-point overall proficiency scale is completed by the administrator.

Table 2. Hands-On Task Item

Phase I J-79, J-57, TF-33 Hands-On Task 347
Shop and Flightline

Objective: To evaluate the incumbent's ability to install starters.

Estimated Time: 25 M Start: Finish: Time Req:

Time Limit: 35M #Times Performed: Last Performed:

Tools and Equipment: Consolidated Tool Kit, 0- to 150-inch-pound Torque Wrench, 10- to 300-inch-pound Torque Wrench, Lubricant.

	Appropriate T.O.:
J-79 (Fighter):	1F-4E-10
J-57 (Tanker):	1C-135(K)A-2-4JG-6
TF-33 (P7) (Cargo):	1C-141A-2-4JG-5 or 1C-141B-10
General Torquing	2-1-111 or 1-1A-8 or specific engine torquing T.O.:
	J-79: 2J-J79-86-7WP00100
	J-57: 1C-135(K)A-2-4JG-1
	TF-33: 1C-141B-10

Background Information: There are some common steps for all three engines, but each engine has some unique steps. The evaluation will be made on the common steps except when indicated. Differences include:
1. J-57 has two cannon plugs.
   J-79 and TF-33 (P7) have one cannon plug.
2. J-57 and TF-33 have one nut on the V-clamp.
   J-79 has two nuts on the V-clamp
Two-person task when actually putting the starter in place. This is the only task for which the incumbent will be required to actually get the technical order from the shelf.

Engine Configuration: The starter adapter pad must be on the engine. The starter is off the engine.

Instructions:

Administer in the shop.
The incumbent MUST use the T.O.
Compare the incumbent's response to the correct answer for the appropriate engine.

Table 2. (Continued)

Phase I J-79, J-57, TF-33 Hands-On Task 347
Shop and Flightline

SAY TO THE INCUMBENT

GET THE T.O. USED TO INSTALL A STARTER AND THE T.O. FOR GENERAL TORQUING PROCEDURES, THEN INSTALL THE STARTER USING THE APPROPRIATE PROCEDURES FROM BOTH T.O.s. FOLLOW GENERAL MAINTENANCE PROCEDURES AT ALL TIMES. TELL ME IF YOU PLAN TO DEVIATE FROM THE T.O. YOU MAY NOT ASK ANYONE TO HELP YOU FIND THE CORRECT T.O.

Performed or Answered Correctly	Yes	No
Did the incumbent:		
1. Obtain the appropriate T.O. for the starter installation and the torquing procedures within 10 minutes?	___	___
2. Hang the clamp per the specific T.O.?	___	___
3. Lubricate the spline?	___	___
4. Ensure that the starter was not left in an unsupported position (hung by the shaft) at any time?	___	___
5. Index (position) the starter per the appropriate T.O.? J-79: Breech at 8 o'clock position J-57: Breech at 3 o'clock position TF-33: Drain plug at 6 o'clock position	___	___
6. Properly seat the V-Band Clamp?	___	___
7. Torque the V-Band Clamp per the appropriate T.O.? J-79 Airsearch: 110 to 130 inch-pounds J-79 Sunstrand: 65 inch-pounds J-57: 65 to 70 inch-pounds TF-33: 60 to 70 inch-pounds	___	___
8. Install the locking device on the V-Band Clamp per the appropriate T.O.?	___	___
9. Connect the applicable electrical connector (cannon plug)? (Must not connect the tachometer generator plug on the J-57)	___	___

Table 2. (Concluded)

Phase I J-79, J-57, TF-33 Shop and Flightline — Hands-On Task 347

10. Use the correct tools and materials? ___ ___

STOP TIME: _____

OVERALL PERFORMANCE

5 Far exceeded the acceptable level of proficiency

4 Somewhat exceeded the acceptable level of proficiency

3 Met the acceptable level of proficiency

2 Somewhat below the acceptable level of proficiency

1 Far below the acceptable level of proficiency

Table 3. Interview Task

Phase III TF-33 P7 Interview Task 325
Flightline

Objective: To evaluate the incumbent's knowledge concerning the determination of high oil consumption on TF-33 engines.

Estimated Time: 20M Start: Finish: Time Req:

Time Limit 25M #Times Performed: Last Performed

Tools and Equipment: None. T.O. 1C-141B-2-4TS-1, page 6-7.

Background Information: N/A

Engine Configuration: N/A

Instructions:

Administer in the shop in a quiet place.
The incumbent may use the T.O. except when indicating the oil flow path.

SAY TO THE INCUMBENT

I AM GOING TO ASK YOU SOME QUESTIONS ABOUT TF-33 ENGINE OIL CONSUMPTION. YOU MAY USE THE T.O. AS A REFERENCE WHEN ANSWERING THESE QUESTIONS EXCEPT FOR THE FIRST QUESTION WHICH DEALS WITH THE OIL FLOW PATH.

Performed or Answered Correctly	Yes	No
1. Beginning and ending at the oil tank, tell me the path that the oil flows through the following components: oil tank, oil bypass valve, oil filter, scavenge pumps, oil jets for bearing cavities and sumps, oil pressure relief valve, air oil cooler, fuel oil cooler, oil pump. Remember, you may NOT use the T.O. while answering this question.		
ANSWER: Incumbent's order 1-10	___	___
a. Oil Tank ___		
b. Oil Pump ___		
c. Oil Pressure Relief Valve ___		
d. Oil Filter ___		
e. Oil Bypass Valve ___		
f. Oil Jet for Bearing cavities and sumps ___		

Table 3. (Continued)

Phase III TF-33 P7 Interview Task 325
Flightline

Performed or Answered Correctly		Yes	No
g. Scavenge Pumps	___		
h. Air Oil Cooler	___		
i. Fuel Oil Cooler	___		
j. Oil Tank	___		
SAY TO THE INCUMBENT			
NOW YOU MAY USE THE T.O. IF YOU WISH			
2. Name four areas other than the oil cooler and the engine cowling that you might check for external oil leaks or internal oil consumption. ANSWER: (The incumbent must mention at least four of the following for credit).		___	___
a. Oil tank	___		
b. Gear box	___		
c. Garloc seal leaks	___		
d. Oil pump accessory housing	___		
e. Pressure lines	___		
f. Scavenge lines	___		
g. Engine inlet	___		
h. Engine exhaust	___		
i. Combustion case split line	___		
3. Why is the engine cowling normally the first area to be inspected when determining the source of high oil consumption? ANSWER: Oil in the cowling would indicate an external leak.		___	___
4. What is the purpose of performing a breather isolation check? ANSWER: To determine the location of the <u>internal</u> oil leak.		___	___

Table 3. (Continued)

Phase III TF-33 P7 Interview Task 325
Flightline

Performed or Answered Correctly	Yes	No
5. Other than checking the servicing level yourself or asking the crew chief, what other source is available for determining when the oil system was last serviced? ANSWER: Aircraft Forms	___	___
6. What readings would indicate a restriction in the scavenge system? ANSWER: A normal oil breather pressure reading and a high oil scavenge pressure reading.	___	___
7. What two basic pieces of information would you need to determine whether or not you had an excessive oil consumption condition? ANSWER: (Incumbent must answer both for credit)	___	___
a. The number of flying hours ___ b. The number of quarts of oil serviced ___		
8. What creates the oil flow from the supply tank to the engine pressure pump? ANSWER: Gravity	___	___
9. What component regulates the oil pressure after it leaves the oil pump? ANSWER: The pressure relief valve.	___	___

STOP TIME: _____

NOTE: TURN PAGE FOR RATING SCALE

Table 3. (Concluded)

Phase III TF-33 P7 Interview Task 325 P
Flightline

OVERALL PERFORMANCE

5 Far exceeded the acceptable level of proficiency

4 Somewhat exceeded the acceptable level of proficiency

3 Met the acceptable level of proficiency

2 Somewhat below the acceptable level of proficiency

1 Far below the acceptable level of proficiency

The interview testing is conducted at the worksite in a "show-and-tell" fashion, such that the person being evaluated can "visually and verbally" describe how a step is to be accomplished (e.g., "that bolt is to be turned five revolutions," or "that component is to be lubricated prior to being assembled"). Thus, additional information, not otherwise collected, can be assembled along with hands-on information to provide a more thorough coverage of the content domain, and hopefully, a more accurate picture of an individual's job proficiency.

WTPT Design and Procedural Approach

At the beginning of this paper, it was stated that two of our objectives were to expand coverage of the job content domain and to determine if measurement comparability could be established between hands-on and interview testing. A short description of the procedure and design for the WTPT testing will clarify how these objectives are met. First (for the Jet Engine Mechanic), in approximately 7 hours of testing time, an individual's job proficiency is measured on 10 hands-on and 10 interview testing work samples. Five of the interview items are designed to cover unique aspects of the job not already measured through hands-on testing; the other five interview items measure proficiency on tasks already covered by existing hands-on items. These 20 work samples, then, make up the Jet Engine Mechanic WTPT. For testing purposes, the 20 items are intermixed and presented to job incumbents at the worksite. As noted previously, these tasks were selected according to a three-phased strategy. The breakdown of these 20 tasks across phases for the Jet Engine Mechanic is shown in Table 4.

Table 4. Breakdown of Work Sample Tests by Phases for the Jet Engine Mechanic Specialty

	Unique items		Overlap items	Total	
	Hands-on	Interview	Interview	Hands-on	Interview
Phase I	5	0	3	5	3
Phase II	3	2	2	3	4
Phase III	2	3	0	2	3
				10	10

Summary and Conclusions

Benefits and Drawbacks of WTPT

As mentioned in the introduction, AFHRL has initiated work in this area with four objectives in mind. We believe the hands-on work sample orientation provides a high-fidelity measurement methodology that can be enhanced by applying the WTPT methodology. Also, through the use of interview testing, measurement of the job content domain is expanded. In addition, if comparability of hands-on and interview testing is established (at least in some instances), a savings in terms of testing time can be realized (or if desired, further coverage of the content domain can be included). One major drawback to this approach is the cost associated with development and testing. In many instances, the benefits gained in fidelity may be offset by the high costs. However, although developmental costs do not differ for hands-on and interview testing, testing time can be significantly reduced if comparability can be demonstrated.

Generalizability of the WTPT Methodology

As noted earlier, the WTPT methodology is targeted for a variety of Air Force specialties in the next 5 to 7 years. Because this measurement methodology is designed to reflect the work performed in a specialty, its orientation should reflect the job content domain. For instance, in our first specialty, Jet Engine Mechanic, the job is (to a large extent) knowledge-based; this orientation is mirrored by the WTPT.

As the WTPT approach is applied to other specialties, the work sample content must necessarily change as well. For example, as selected specialties become more abstract (e.g., Administrative Specialist) or managerial in nature, the WTPT may begin to resemble an assessment center approach. In any event, as data are collected and evaluated, decisions will be made about the costs and benefits of adopting such a strategy for measuring job proficiency.

A TASK-LEVEL DOMAIN SAMPLING STRATEGY: A CONTENT VALID APPROACH

M. Suzanne Lipscomb

Air Force Human Resources Laboratory

In developing task-based job performance measures, it is impractical to assess performance on the universe of tasks within most Air Force specialties (AFSs). No individual performs all of the tasks in any specialty; and in most specialties, no individual performs an "average" job. Rather, the tasks of a specialty are distributed by management action to individuals in consistent ways so as to cluster into a variety of types of jobs, based on the co-performance of tasks and the variations in mission, equipment, or management in any given locale. This variance of jobs within AFSs is an exceedingly important phenomenon since it impacts on how the specialty is organized in the personnel system, the aptitudes required, the training provided, and the way individuals can be utilized in the workplace (Mitchell & Driskill, 1979).

This variance in Air Force jobs is of concern to Air Force managers and is one of the major issues of study in the occupational analysis program (Air Force Regulation 35-2, Occupational Analysis Program). Data on most AFSs indicate that the classification structure of the Air Force is highly dynamic, with frequent reallocation of tasks among specialties. In addition, although there may be some common tasks performed by a majority of individuals within a specialty, most of the tasks are performed only by members of the various job types within the specialty.

It is necessary therefore to rely on samples of performance that are both useful for differentiating between good and poor performers and representative of the performance domain. Differentiating good and poor performers can be accomplished by assessing job incumbent performance on tasks with a range of difficulty. In addition to identifying tasks with a range of difficulty, selecting tasks that adequately represent the total specialty domain is necessary to make inferences about performance from the sample of specific task measures used. If the specialty domain is adequately represented by the tasks selected, the task-based measurement system can be considered content valid.

Unlike other types of validity, the content validity of a measurement procedure is not a correlational process but an evaluation of adequacy and representativeness using rational judgments. Lennon (1956) stated that three assumptions underlie the use of content validity: (a) the area of concern to the user can be conceived as a meaningful, definable universe of responses; (b) a sample can be drawn from the universe in some purposeful, meaningful fashion; and (c) the sample and the sampling process can be defined with sufficient precision to enable the user to judge how adequately the sample of performance typifies the universe of performance.

Given the information available in the Air Force Occupational Research Data Base, these three assumptions can be met; thus, the issue of content validity can be addressed. The universe of responses can be defined as the universe of tasks for an AFS, as detailed by the occupational survey report task list. The sample can be drawn in a meaningful fashion based on the task-level occupational survey data available. Finally, the sampling process can be defined with precision using a task sampling plan which will allow a judgment to be made as to the adequacy of the sample.

A task sampling plan consisting of a procedural set of guidelines must be developed: (a) to specify the job and task domains of interest, (b) to establish the level of measurement specificity, and (c) to determine the proportional weighting (importance) of the work activities identified. Such guidelines will assure objectivity, replicability, and comparability of efforts to develop measures which detect meaningful differences in performance. These guidelines are presented as they apply to enlisted AFSs. Also provided is an illustration of their use for

selecting tasks within the first AFS to be investigated for the Joint-Service Job Performance Measurement Project, Jet Engine Mechanic (AFS 426X2).

Task Selection Procedural Guidelines

Defining the Job Domain

For most AFSs, the Air Force has a wealth of information sources, which give a comprehensive picture of the work domain. These sources give the AFS entrance requirements and a general specialty description (AFR 39-1, Airman Classification Regulation); AFS training requirements (AFR 50-5, USAF Formal Schools and Specialty Training Standards); and occupational survey data.

Occupational survey data include the percentage of incumbents performing specific tasks and the relative time they spend on each, as well as SMEs' judgments of the relative time required to learn to perform tasks (i.e., task difficulty) and the relative importance of training for each task (i.e., recommended training emphasis). Occupational survey and training information cover the full scope of tasks performed by incumbents in an AFS, and therefore can be applied to the development of the task-based performance measures.

Because occupational survey data provide the most detailed and comprehensive source, they will be used to define the work domain. The other sources will provide complementary information.

The goals of the Air Force job performance measurement program are to assess specific job competencies required within a specialty and general competencies applicable across AFSs. These two types of measures require four levels of measurement specificity: Air Force-wide, specialty-wide, duty-core, and incumbent-unique measures. Because the focus of this paper is on selecting tasks required to measure individuals' competence within an AFS, the latter three levels of measurement specificity will be highlighted.

To include an adequate representation for each of these three levels of measurement specificity, tasks within an AFS must be categorized accordingly. That is, tasks can be categorized into those performed throughout the specialty (i.e., specialty-wide), those specific to certain duties within an AFS (i.e., duty-core), and those uniquely performed by incumbents in certain job types (i.e., incumbent-unique).

The occupational survey task inventory will be used to define the work domain and categorize tasks. As task performance is often specific to equipment or work centers, tasks associated with equipment or work centers will be used to identify the duty-core domain. Finally, tasks associated with specific job types defined by the occupational analysis will delineate the incumbent-unique domain. Because it would be impractical to cover adequately all duty areas and job types within heterogeneous AFSs, those most representative of the work performed will be selected. That is, duty areas and job types which involve the largest percentage of personnel will be chosen.

Selecting Tasks Representative of the Job Domain

The procedures for sampling tasks representative of the three task domains are outlined in the following paragraphs, along with the rationale for these procedures. For each task domain, the number of tasks selected should be based on a judgment of the number of performance measures required to give an adequate sample, while conforming to a total testing time of no more than 8

hours for all measures. This time limit is considered the maximum time feasible to keep an airman away from his/her unit. Within this timeframe, individuals will be assessed on specialty-wide, duty-core, and incumbent-unique tasks.

Phase I. Selection of Specialty-Wide Tasks

Step 1. Select all tasks that are included in the Plan of Instruction (POI) for initial AFS training or, if not in the POI, are performed by at least 30% of the first-term incumbents with 1 through 48 months of total active Federal military service).

This will reduce the task pool to those tasks deemed important enough to train or those which are performed by a substantial number of first-term airmen across the AFS. (The 30% cutoff value may be varied by specialty according to the number of tasks performed by first-termers in that specialty.)

Step 2. Cluster tasks selected in Step 1 based on one of the following: (a) factor analysis based on tasks performed together, (b) Specialty Knowledge Test outline, (c) Specialty Training Standard outline, or (d) occupational survey inventory duty outline. Each of these is a means of organizing the pool of tasks into performance/knowledge areas based on occupational information. All will produce similar results; thus, the selection of the grouping strategy should be based on a judgment as to which is cost effective and best suited to the development of performance measures for a specific AFS.

Step 3. Weight each task cluster to reflect its relative importance to the overall performance of first-term airmen within the specialty. Possible indices for weighting clusters include the following: (a) Specialty Knowledge Test outline weights, (b) Specialty Training Standard proficiency-level requirements, (c) SME judgments of relative importance, or (d) weights derived from existing task factor data. Relevant task factor information includes recommended training emphasis ratings (i.e., SME judgments of the extent to which training is required for tasks) and percent time spent values (i.e., incumbent ratings of the relative time spent performing tasks). These task factor data can be used to derive weights by generating the product of the mean recommended training emphasis rating and the cumulative percent time spent performing tasks in a cluster.

Step 4. Determine the number of tasks to be selected from each cluster to reflect the assigned weights as follows. Total the cluster weights. Divide each cluster weight by the total to get a percentage. Multiply each cluster percentage by the total possible tasks to find the number of tasks to be selected for each cluster.

Step 5. Within each cluster, select the number of tasks determined in Step 4 to reflect a range of learning/task difficulty by: (a) ranking the tasks from low to high task difficulty, (b) dividing the ranked list into quartiles, (c) selecting 40% of the tasks from the fourth quartile, (d) selecting 30% from the third quartile, (e) selecting 20% from the second quartile, (f) selecting 10% from the first quartile, (g) repeating for each cluster. (It is important to sample tasks with a range of difficulty so incumbent performance assessment will reflect the rank-ordering of people of varying levels of job competence. The sampling is more heavily weighted on the more difficult tasks because they determine the aptitude requirements of the specialty and are also where most performance variation should occur.)

Step 6. Review the tasks identified in Step 5 to determine if they can be measured by either the hands-on or interview component of WTPT. Reject any task found to be unsuitable for WTPT, and document the reason it was judged unsuitable. If possible, select a replacement task from

the same task difficulty quartile. (The ability to assess performance through observation/interview procedures [WTPT] is a prerequisite to final task selection because performance measures obtained via these high-fidelity techniques will be the benchmarks against which surrogate measures are compared.)

Phase II. Selection of Duty-Core Tasks

Because the performance domain for a duty area (e.g., a specific engine type or work center) is less broad than for the entire specialty, fewer tasks are needed for an adequate sample. Also, because tasks selected for one duty area may be performed in another, tasks can be selected for more than one duty area. However, since tasks selected in Phase I for specialty-wide measures will be used to assess all incumbents, they should not be used to develop duty-core measures. The following steps apply for each duty area.

Step 1. Select from among those task not utilized in Phase I all tasks performed by at least 40% (as noted earlier, this cutoff may vary according to the number of tasks performed by first-termers) of the first-term airmen identified as performing the duty in question. (Within each duty area, a higher proportion of incumbents performing tasks can be used as the basis for identifying tasks to be assessed because the performance domain is more narrowly defined than across the entire specialty.)

Step 2. From the tasks identified in Step 1, select tasks which reflect a range of learning/task difficulty by repeating Phase I, Step 5.

Step 3. Repeat Phase I, Step 6.

Phase III. Selection of Incumbent-Unique Tasks

Because the performance domain for each job type is much less broad than for the entire specialty, fewer tasks are needed to provide an adequate sample. Also, the tasks selected for a job type may be applicable to more than one job type; however, tasks selected in Phases I or II should not be used to develop incumbent-unique measures. The following steps apply for each job type.

Step 1. Select all tasks performed by 50% or more of the incumbents in the incumbent-unique group and not utilized in Phases I or II. (Again, as the job domain becomes more specific, it is possible to select tasks performed by a higher proportion of incumbents. In addition, the cutoff may vary by number of tasks performed by first-termers.)

Step 2. From the tasks identified in Step 1, select tasks which reflect a range of learning/task difficulty by repeating Phase I, Step 5.

Step 3. Repeat Phase I, Step 6.

Review and Approval of Task Sample

Upon application of these task sampling procedures, the specialty-wide tasks selected for each AFS were reviewed by appropriate AFS functional managers and technical training representatives, who provided feedback concerning the adequacy of the tasks selected. Reviewers examined the task sample to ensure that work performed by first-term airmen and critical wartime requirements were well represented. Approval of the task sample by these policy-makers should increase the acceptance and utilization of the resulting job performance data.

Application to the Jet Engine Mechanic Specialty (AFS 426X2)

Before the sampling plan could be applied to the Jet Engine Mechanic Specialty, duty areas and incumbent-unique job types were identified using the following procedures.

Defining the Job Domain

Duty Areas Selected. Duty areas were selected based on the type of engine maintained. An inspection of the occupational survey data revealed that 20%, 18%, and 17% of AFS 426X2 first-term airmen performed maintenance tasks on J-57, J-79, and TF-33 engines, respectively. Because these percentages were the highest among the nine engine types maintained by AFS 426X2 personnel, these three engines were selected as being representative of equipment maintained by first-term jet engine mechanics.

Job Types Selected. The occupational survey data also revealed that the vast majority of first-term jet engine mechanics performed similar jobs (i.e., most airmen maintained similar engine accessory systems). The largest percentage of first-term incumbents in each major command (MAJCOM) spent the majority of their time performing general engine maintenance tasks in shop or on the flightline. As a result, these two functional areas were identified as being representative of AFS 426X2.

Phase I. Selecting Specialty-Wide Tasks

Task Clustering. Tasks were clustered by occupational survey duty area because this grouping adequately reflected the work done in the specialty and was cost effective. Weights were computed based on the product of the mean recommended training emphasis rating and the cumulative percent time spent performing tasks in a cluster. The following six task clusters received the weights indicated below.

Cluster	Weight
Preparing and Maintaining Forms, Records and Reports	10
Performing Quality Control Functions	5
Performing Flightline Engine Maintenance Functions	10
Performing In-Shop Engine Maintenance Functions	20
Performing Test Cell Functions	5
Performing General Engine Maintenance Functions	50

Task Selection and Review. The remaining Phase I steps were followed, and 18 tasks were selected to reflect the weights outlined above. Ten tasks each were selected for each engine type in Phase II and for each job type in Phase III. Selected tasks were reviewed by SMEs, and unsuitable tasks were deleted.

New tasks were selected and reviewed, and the task list was finalized, giving a representative set of tasks on which to develop performance measures. The main justifications for the task exclusions were:

1. Task not common to all engines (Phase I).
2. Task performed differently on different aircraft (Phase I).
3. Task not common to all functional areas (Phases I and II).

4. Task not representative of functional area (Phase III).

5. Task unclear, too broad, complex, or trivial.

6. Task overlapping or similar.

7. Task performed differently depending on how engine is shipped (air, rail, or truck) and its destination (depot, deployment).

8. Task performed differently depending on organizational unit (Examples: Some supervisors do not allow test cell personnel to transport engines. SAC flightline personnel do not make entries on oil analysis request forms (DD Form 2026), but MAC flightline personnel do make such entries).

9. Task involves equipment being changed within the year.

In summary, a strategy for task selection was developed to sample tasks representative of the job content. This strategy was applied to the Jet Engine Mechanic Specialty (426X2), and the selected tasks were used to develop the performance measures and standards described in the following paper.

DEVELOPING PERFORMANCE MEASURES AND STANDARDS FOR ACCURATE ASSESSMENT

Rodger D. Ballentine
and
M. Suzanne Lipscomb

Air Force Human Resources Laboratory

Once representative task statements from the occupational inventory have been identified by the Task Selection Plan, and refined through SME input, the developmental focus shifts toward task analysis and item writing. This paper will highlight the time period from task identification through construction of work sample tests to measure performance on these tasks. Discussion will center on the task analysis process, item writing, pilot testing, and final selection of test items.

Task Analysis Process

The objective of the task analysis process is to gather information essential to WTPT item development. Once the task statements have been identified using the Task Selection Plan discussed by Lipscomb in the previous paper, these statements are taken to operational units and discussed with SMEs (senior noncommissioned officers). Because occupational inventory task statements vary from rather general to quite specific, detailed discussions are required to clarify task boundaries and the nature of the work performed. Several interview workshops with SMEs from field units will provide the majority of the information needed to construct items.

In these working sessions, technical orders and job guides are used to identify the steps involved in task performance, the correct procedures to be used by job incumbents, and the sequencing of steps if a rigid ordering is required. An extensive listing of information gathered during the task analysis process can be found in Table 5, but several issues need to be highlighted here.

Table 5. Information Gathered During the Task Analysis Process

1.	Task/Step Information • Consequences of incorrect performance • Identification of critical steps (criticality criteria) --Safety of personnel/equipment --Required for proper task completion --Required to maintain proper maintenance procedures • Estimated frequency of incorrect performance • Sequencing of steps
2.	Amount of time required to perform task
3.	Required technical orders (job guides), tools, and equipment
4.	Number of people required to perform task
5.	Do first-termers typically perform task?
6.	Are general maintenance procedures used--or are there unique base, MAJCOM, or engine procedures?
7.	If interview testing is used, are there specific questions to ask that will identify ability to perform?

Much of the additional information gathered during task analysis can be categorized as logistical in nature. When decisions are being made about final testing items, and how testing is to be accomplished, logistical information is critical to successful data collection. For example, the amount of time required to perform a task, whether more than one person is required for task completion, and consequences of incorrect performance (i.e., personal injury or equipment damage) are all factors which must be considered in designing test items and selecting the most appropriate work sample testing modality--hands-on or interview procedures.

To gather this task-centered and logistical information for the Jet Engine Mechanic WTPT, test developers visited 12 bases. Workshops involving a total of 75 SMEs were conducted for task analysis. This extensive task analysis process was required because tasks were selected specific to three engine types in the specialty. Table 6 shows the types and numbers of job experts interviewed during the Jet Engine Mechanic task analysis.

Table 6. Bases Visted and Number of SMEs Interviewed During Task Analysis

Air Force base	Engine	Number of SMEs
Bergstrom AFB	J-79	10
Carswell AFB	J-57/TF-33	11
Altus AFB	J-57/TF-33	6
Seymour Johnson AFB	J-79	7
Shaw AFB	J-79	6
Barksdale AFB	J-57	5
Dyess AFB	J-57	5
Blytheville AFB	J-57	3
Travis AFB	TF-33	5
Norton AFB	TF-33	5
Kelly AFB	J-57/TF-33/J-79	5
Randolph AFB	J-57/TF-33/J-79	7

An important added benefit of these task analysis base visits was the opportunity afforded the test developers for equipment and work site familiarization. Periodically, during the course of interviews with SMEs, confusion arose concerning the placement and functioning of a particular step. By visting the work site, test developers could learn "first-hand" what was being discussed.

In addition to these base visits, a centralized workshop was held to gather task analysis information for Phase I tasks. Representatives for all engine types and personnel from the Technical Training School were assembled to generate all required information to construct tests for these tasks.

These SMEs also assisted test developers in evaluating tasks from all phases, specifically in terms of whether certain tasks could be clustered into modules (e.g., three documentation tasks combined into one test item).

In summary, the objective of the task analysis process is to gather information essential to WTPT item development. Relevant information includes beginning and ending points for each specific task, critical steps for task accomplishment, logistical requirements for task completion, required configuration of equipment, time-critical and safety steps, effects of local operating procedures on task performance, and representativeness of the task. This information

is gathered by referencing applicable regulations, technical orders, and local operating instructions, as well as through discussions with SMEs in the field. The desired result of the analysis is a comprehensive list of steps required for successful task completion. These steps should be generalizable to any situation in which the task might be observed and should be used to objectively evaluate an individual's performance on the task.

WTPT Item Writing

Once the task analysis process has been completed, test developers have the information necessary to write test items. Remember, data concerning logistics, feasibility for hands-on or interview testing, number and criticality of steps have been gathered. Now, the test developer utilizes this information to organize and compose test items reflecting steps required to perform each task.

When writing items, the test developer must decide (based on SME input) which steps represent the behavioral elements necessary to adequately characterize that task. This may involve eliminating unnecessary steps that add little to successful task performance (and little to the discriminating power of the item). Similarly, certain lengthy tasks may consist of several subtasks, one or more of which is representative of the behavioral requirements of the entire task. In such cases, an item can be constructed utilizing one subtask with similar properties and task difficulty levels. Specific criteria used in selecting subtasks are listed in Table 7.

Table 7. Criteria to Select Measurable Subtasks

Representative of the general task?
Discriminator between a good and bad performer?
Measurable using job performance methodology?
Capable of being performed in a testing situation?
Capable of being completed within the testing time limitations?
Frequently performed by first-term airmen?

Another important component of this process is an evaluation of whether hands-on or interview items should be written to represent a task. As Hedge noted earlier, the WTPT methodology calls for using interview testing when time, cost, or safety considerations suggest hands-on testing is impractical. In the present effort, because interview testing was to be evaluated as a potential surrogate for hands-on testing, both hands-on and interview items were developed for some tasks. As shown in Table 2 of Hedge's paper, hands-on and interview items were constructed such that six tests (3 engine types x 2 functional areas) were compiled, with each test consisting of 10 hands-on and 10 interview items. In all, 59 unique WTPT items were constructed for the Jet Engine Mechanic specialty (i.e., 8 for Phase I, 21 for Phase II, and 30 for Phase III). Each item was then reviewed and refined by three to four groups of SMEs.

Pilot Testing

Once items had been written, reviewed, and gathered into test booklets, pilot tests were set up at three Air Force bases (one per engine type). Pilot testing served multiple purposes. Items were evaluated to ensure that all critical steps were included and were sequentially

correct, that items were applicable to first-term airmen, etc. Booklets were also assessed in terms of clarifying instructions, test booklet layout, and inclusion of applicable technical orders or job guides. In addition, task setup and equipment requirements were checked, as well as base support required (e.g., personnel/equipment availability). Equipment use and safety issues were also identified. Finally, time to complete all tasks in the WTPT was examined to be sure the total test could be completed within the time allotted. Thus, information gained from the pilot test allowed appropriate revisions to instruments and procedures in preparation for data collection.

Lessons Learned

Having worked through the development of performance measures for the Jet Engine Mechanic, valuable insights were gained about the adequacy of the developmental process.

Field experience indicated that SME judgments during the task selection workshop, as described by Lipscomb, were correct and invaluable. Consequently, it is recommended that the task selection plan review and modification be based on SME analysis of tasks (rather than total reliance on occupational survey task data). Such input can provide important information about how specific or how general task statements are, as well as how much procedures vary because of equipment or environment. This helps to identify tasks having the same level of specificity, or modules of work which first-term airmen typically perform.

SME input is also essential during the task analysis process. It is improbable that test developers will possess the knowledge and skills necessary to understand the intricacies of the specialty for which items are being written. With the use of SMEs, in fact, test developer naivete becomes a benefit, in that SMEs are forced to explain task procedures in detail. In addition, SME input during task analysis and item development will assist test developers in identifying modules of work performed. This grouping of tasks will facilitate the development and testing process.

In summary, the task analysis and item writing process is a lengthy, yet necessary component in constructing work sample tests. The procedures used in task analysis and item writing were described, and the information gathered at each point was detailed. Throughout the process, the use of SMEs was shown to be essential to developing an accurate, well-constructed work sample test.

SOME COMMENTS ON WALK-THROUGH PERFORMANCE TESTING

Terry L. Dickinson

Old Dominion University

My remarks will primarily be directed to support the Air Force in its attempts to develop a program for job performance measurement. I am encouraged by the effort to implement state-of-the-art technologies in performance measurement and, also, by the concern to approach this implementation with an attitude of research-mindedness. I believe that applications in psychology often lead to the identification of new research issues as well as the resolution of old ones. It is with these thoughts in mind that I will discuss each of the papers in this symposium.

Dr. Gould and Dr. Hedge have indicated that the Air Force is developing its program of WTPT to address the "criterion problem." Performance measures constructed with this technology will serve as high-fidelity benchmarks against which other measures such as peer and supervisory ratings will be compared for substitutability. For many jobs and their tasks, WTPT would appear to provide these benchmarks. In particular, tasks that result in products (Osburn, 1973) are most appropriate for comparing ratings and WTPT measures. Assuming that the tasks have been sampled to represent critical requirements of job performance, the fidelity of WTPT measures cannot be questioned. Such measures will truly be work samples, and performance on these samples can be referenced to job content.

However, for jobs and their tasks that do not have products directly linked to performance, it is questionable whether WTPT can provide measures that can logically be thought of as high-fidelity benchmarks. For many jobs, the process of performance is as important as the final product of performance, and for some jobs, the product is the process of performance. Work samples could be constructed for these jobs. However, performance on the samples would necessarily involve observation by judges and the rating of that performance. Since there may be several acceptable ways to perform job tasks, the work samples must be constructed very carefully to allow for the full range of acceptable performance. Of course, it may be impossible to construct work samples for some jobs that will totally simulate the range of acceptable performance. Part of the Air Force's research efforts should be directed toward exploring the product versus process description of tasks in terms of identifying jobs for which WTPT can provide high-fidelity benchmarks.

Another goal of WTPT is to operationalize the test development and performance measurement procedures such that they can be conducted by technicians. Clearly, training programs need to be developed so that technicians can (a) construct task sampling plans and execute these plans, (b) construct the tests and their administration procedures, and (c) administer the tests at a variety of sites. The implementation of these programs suggests an interface between personnel psychology and organizational development activities. I believe that lessons can be learned from the technology of organizational development that can aid in the efficient accomplishment of this goal. Finally, these training programs will need to be evaluated rigorously to ensure that the WTPT measures that are produced by technicians meet acceptable professional standards.

I like the fact that a conceptual model of performance measurement is being used to guide WTPT. This model clearly suggests the appropriateness of different methods and sources for measuring different job content domains. For example, self-ratings are assumed to be best for measuring technical skills; peer ratings are better for interpersonal/social skills; and supervisory ratings are good for both types of skills. Continued iteration and refinement of the model should be an important goal in the Air Force's R&D program. Refinement of the model will ensure that the program is conceptually driven. For example, the current application of WTPT to

jet engine mechanics is assessing the comparability of the hands-on component and the interview component for describing performance on a subset of tasks. The selection of these tasks has been driven by practical considerations such as time to perform, safety, and potential damage to equipment. These are important considerations. However, I suggest that the Air Force attempt to identify the task attributes that specify the degree of interchangeability of hands-on and interview components. A key for identifying these task attributes may be the assumption that job knowledge (i.e., what to do) implies job proficiency (i.e., can do). Research should identify the task attributes that weaken or strengthen the relationship between job knowledge and job proficiency.

Dr. Hedge described the methodology of WTPT. He noted that a task sampling plan is used to define the job content domain from which tasks are sampled to develop work sample tests. Here, I think it is important to make the distinction between a job content domain and a job content universe. As Guion (1979) noted, the job content universe includes all the nontrivial tasks, responsibilities, and organizational relationships inherent in a job. The job content domain is a subset of the universe and consists of that portion of the universe that is identified for the purposes of testing.

Dr. Hedge noted that as the Air Force moves into more abstract specialties (e.g., supervisory or administrative), the job content universes may be less amenable to WTPT. The Air Force should consider developing for jobs indexes that describe the percentage of tasks that can be measured by WTPT technology. These indexes could be useful for constructing a job classification system that predicts the amenability of jobs to various performance measurement technologies.

Ms. Lipscomb described the task sampling plan that the Air Force used to develop WTPT measures for jet engine mechanics. In this plan, a wealth of information was available for sampling the job content domain; e.g., general job description, job training requirements, and occupational analysis information. Several parameters were defined for sampling the tasks from that domain, and they included being part of the plan of instruction at the training school, performed by at least 30% of first-term incumbents, and part of a duty cluster. This task sampling plan is impressive in terms of its specificity and detail. The plan certainly exemplifies the content-oriented approach to test construction.

Several research questions come to mind about the task sampling plan. First, is the extreme detail in the plan necessary? Furthermore, are the particular levels of the parameters important? Would essentially the same tasks have been selected with less detailed plans or different parameter levels? I propose that a sensitivity analysis (Fischoff, 1980) be conducted on the task sampling plan.

We know too little about the plans and parameters that are used to sample tasks for content-oriented test construction. If less costly and time-consuming plans can be used to identify tasks, they should be used.

Furthermore, sensitivity analyses that are done across several jobs may indicate that the parameters and complexity of the plans vary across specialties and job types. Such knowledge would be quite useful for expanding the theory and technique of content-oriented test construction.

Lt Col Ballentine described the lessons that were learned from the procedures used for the construction of WTPT measures for jet engine mechanics. Although extensive task information was available, it was not sufficient for test construction. SMEs were interviewed to check the appropriateness of the tasks for test construction. Some tasks were too complex for testing and needed to be subdivided, whereas other tasks were too trivial and were ignored. This result

underscores the generic nature of job analytic information and indicates the importance of adapting the information to meet its intended use.

Additional reviews by SMEs at different Air Force bases identified procedural differences among the bases in how tasks were accomplished. Tasks were not considered for test construction if the differences had to be included as steps in the work samples. Procedural differences bear on the generalizability of the test scores and should serve as a red flag. Ideally, WTPT measures should be assessed for location differences in task procedures at all Air Force bases. Each step in the measures must be applicable to every job incumbent. Otherwise, situational effects will account for some unknown proportion of variance in test performance. Clearly, assessing location differences at all bases is economically unfeasible. However, Generalizability Theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) could be used to ascertain the number of bases necessary to reasonably estimate the proportion of variance in test performance that is accounted for by procedural differences.

Finally, a concern that I have for WTPT is the quality of test administration. Test administrators must be trained in how to collect data with considerable accuracy. Unfortunately, we know too little about accuracy training. Furthermore, the limited body of research suggests that rating accuracy is not the same as observation accuracy (see Murphy, Garcia, Kerkar, Martin, & Balzer, 1982). Clearly, both rating accuracy and observational accuracy will be needed in WTPT. Depending on the job, WTPT measures can require the administrators to rate and observe test performance. The Air Force should give strong consideration to devoting some of its efforts to R&D in accuracy training.

REFERENCES

AFR 35-2 (1982, July). Occupational analysis. Washington, DC: Department of the Air Force.

AFR 39-1 (1982, January). Airman classification. Washington, DC: Department of the Air Force.

AFR 50-5 (1987, March). USAF formal schools (policy, responsibilities, general procedures, and course announcements). Washington, DC: Department of the Air Force.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Fischoff, B. (1980). Clinical decision analysis. Operations Research, 28, 28-43.

Gould, R.B., & Hedge, J.W. (1983, August). Air Force job performance criterion development. Paper presented at the annual meeting of the American Psychological Association, Anaheim, CA.

Guion, R.M. (1979). Principles of work sample testing: III. Construction and evaluation of work sample tests (TR-79-A10). U.S. Army Institute for the Behavioral and Social Sciences, Alexandria, VA.

Kavanagh, M.J., Borman, W.C., Hedge, J.W., & Gould, R.B. (1987), Job performance measurement in the military: A classification scheme, literature review, and directions for research (AFHRL-TR-87-15). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

Lennon, R.T. (1956). Assumptions underlying the use of content validity. Education Measurement, 16, 294-304.

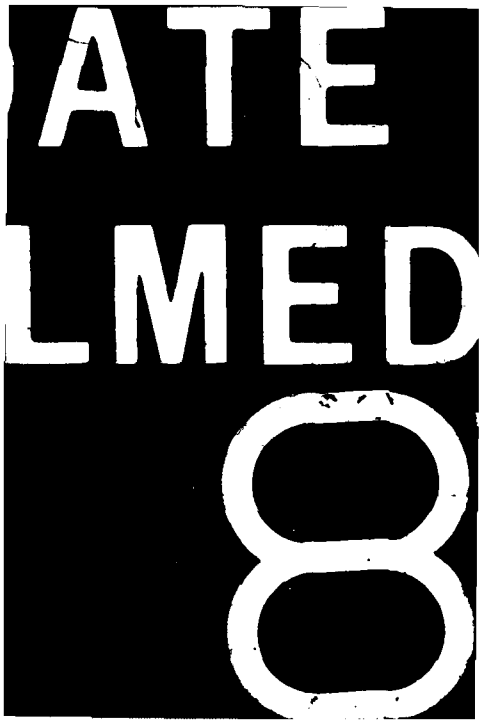
Mitchell, J.L., & Driskill, W.E. (1979, October). Variance within occupational fields: Job analysis versus occupational analysis. Proceedings of the 21st Annual Conference of the Military Testing Association (pp. 259-268). San Diego, CA: Navy Personnel Research and Development Center.

Murphy, K.R., Garcia, M., Kerkar, S., Martin, C., & Balzer, W.K. (1982). Relationship between observational accuracy and accuracy in evaluating performance. Journal of Applied Psychology, 67, 320-325.

Osburn, H.G. (1973, October). Process versus product measures in performance testing. Paper presented at the annual conference of the Military Testing Association, San Antonio, TX.

Uniform Guidelines for Employee Selection Procedures (1978, August). Federal Register, 43(166).

Wilson, C.L. (1962). On-the-job and operational criteria. In R. Glaser (Ed.), Training research and education. Pittsburgh: University of Pittsburgh Press.


ATE
LMED
8