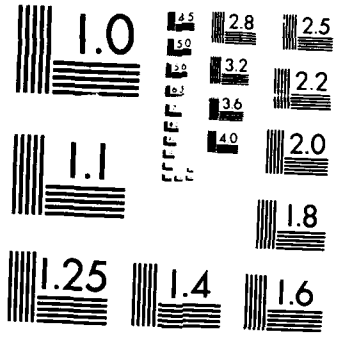
AD-A193 344 THE USE OF VIDEOTAPE TECHNOLOGY TO TRAIN ADMINISTRATORS OF WALK-THROUGH PERFORMANCE TESTING(U) AIR FORCE HUMAN RESOURCES LAB BROOKS AFB TX J W HEDGE ET AL JUL 88 1/1

UNCLASSIFIED AFHRL-TP-87-71 F/G 5/9 NL



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

AFHRL-TP-87-71

(2)

AIR FORCE HUMAN RESOURCES LABORATORY

THE USE OF VIDEOTAPE TECHNOLOGY TO TRAIN ADMINISTRATORS OF WALK-THROUGH PERFORMANCE TESTING

Jerry W. Hedge

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

Terry L. Dickinson
Old Dominion University
Department of Psychology
Norfolk, Virginia 23508-8559

Sheryl A. Bierstedt
Universal Energy Systems
8961 Tesoro Drive
San Antonio, Texas 78217-6225

July 1988

Interim Technical Paper for Period February 1985 - March 1987

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

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-87-71	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	PROJECT NO.	TASK NO	WORK UNIT ACCESSION NO
62703F	7734	08	22

11. TITLE (Include Security Classification)
The Use of Videotape Technology to Train Administrators of Walk-Through Performance Testing

12. PERSONAL AUTHOR(S)
Hedge, J.W.; Dickinson, T.L.; Bierstedt, S.A.

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Interim	FROM Feb 85 TO Mar 87	July 1988	14

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	
05	09		interrater reliability, rater accuracy, ~~work sample testing~~
05	10		Job Performance Measurement, rater training,
			performance measurement, Walk-Through Performance Testing (WTPT),

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

When assessing the performance of individuals with work sample tests, it is essential that test administrators observe and rate performance accurately. The present effort used videotapes of work sample test performance with known target scores both as a training device to improve observational and rating skills as well as an evaluation tool to determine level of rater accuracy and identify type and frequency of rating errors. Data were collected from nine test administrators at two workshops held 2 1/2 months apart and in a field test conducted in the interim between the two workshops. Rater accuracy and interrater agreement were measured for both workshops; only interrater agreement was measured during the field test. Results indicated high levels of accuracy and interrater agreement for the test administrators. Analyses of types and frequency of rating errors suggested that errors resulted most often from failure to identify correct performance when it occurred. Results are discussed in terms of the training and evaluation benefits of using videotape technology in work sample testing.

20 DISTRIBUTION/AVAILABILITY OF ABSTRACT	21 ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL	22b TELEPHONE (Include Area Code)	22c OFFICE SYMBOL
Nancy J. Allin, Chief, STINFO Office	(512) 536-3877	AFHRL/TSR

DD Form 1473, JUN 86 *Previous editions are obsolete.* SECURITY CLASSIFICATION OF THIS PAGE

Unclassified

AFHRL Technical Paper 87-71

July 1988

THE USE OF VIDEOTAPE TECHNOLOGY TO TRAIN ADMINISTRATORS OF WALK-THROUGH PERFORMANCE TESTING

Jerry W. Hedge

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

Terry L. Dickinson

Old Dominion University
Department of Psychology
Norfolk, Virginia 23508-8559

Sheryl A. Bierstedt

Universal Energy Systems
8961 Tesoro Drive
San Antonio, Texas 78217-6225

Reviewed and submitted for publication by

Jerry W. Hedge, Chief
Job Performance Measurement Section
Training Systems Division

This publication is primarily a working paper. It is published solely to document work performed.

SUMMARY

The Job Performance Measurement project has developed a work sample test known as Walk-Through Performance Testing (WTPT) to assess the performance of first-term enlisted personnel in the United States Air Force. In order to train test administrators, videotapes of task performances were developed. These videotapes were used both as a training device to improve observational rating skills, as well as an evaluation tool to determine the level of rater accuracy and to identify the type and frequency of rating errors. Results indicated high levels of rater accuracy and interrater agreement for all test administrators, and suggested the continued use of videotape technology to enhance and evaluate work sample test administrator assessments.

PREFACE

The Training Systems Division of the Air Force Human Resources Laboratory is engaged in a multi-year effort to develop performance criteria for use in validating the Air Force selection/classification system and evaluating training programs. The high-fidelity criterion developed for these purposes utilizes a work sample testing approach known as Walk-Through Performance Testing (WTPT). This paper describes the Laboratory's approach to training test administrators to be accurate assessors of work performance. An earlier version of this paper was presented in 1985 at the annual meeting of the American Psychological Association in Los Angeles, California.

TABLE OF CONTENTS

LIST OF TABLES

THE USE OF VIDEOTAPE TECHNOLOGY TO TRAIN ADMINISTRATORS OF WALK-THROUGH PERFORMANCE TESTING

I. INTRODUCTION

When performance is assessed with a work sample test, the skills of test administrators in observing and rating performance are critical to accurate measurement. Consequently, it is important that the researcher devise a strategy that focuses on developing these observational and rating skills. Ideally, this strategy would also provide information that could be used to evaluate the level of accuracy achieved in applying these skills.

The most popular approach for evaluating rating accuracy has employed videotape technology, whereby pre-defined scenarios of behavior are acted out by "ratees" and recorded on videotapes for later showing to raters. The raters assess the videotaped performance of the ratees, and their ratings are compared to expected or "expert-generated" target scores to derive measures of rating accuracy.

This videotape technology for evaluating rating accuracy was pioneered by Borman and his colleagues (Borman, Hough, & Dunnette, 1976). To our knowledge, the technology has been used exclusively in structured laboratory situations, where researchers have evaluated the effects of such variables as type of rater training (Pulakos, 1984), purpose of appraisal (McIntyre, Smith, & Hassett, 1984), and rating format (Nugent, Laabs, & Penell, 1982).

The primary purpose of the present effort was to evaluate administrators of work sample tests in a non-laboratory setting using videotape technology. It was anticipated that intensive training to develop observation and rating skills would result in high levels of interrater reliability and rating accuracy. A second purpose was to identify the types of errors committed by raters. This information could be useful for altering the focus of training or guiding the selection of additional training strategies.

II. METHOD

Background

Currently, the Armed Services are engaged in developing performance measures in order to validate the Armed Services Vocational Aptitude Battery (ASVAB) (Wigdor & Green, 1986). These performance measures are to be administered to current job incumbents at the work setting to obtain criterion data for validation.

Work Sample Tests

The Air Force Human Resources Laboratory has developed a methodology known as Walk-Through Performance Testing (WTPT) (Hedge, 1984) as part of the Air Force's contribution to validate the Armed Services Vocational Aptitude Battery. A hands-on component in WTPT resembles a traditional work sample test, and it is designed to measure hands-on performance on critical job tasks. A second component, interview testing, was developed to ensure adequate coverage of the job content domain, because practical constraints often prevent the development of a measure of hands-on performance for a critical task. The focus of the current investigation was the hands-on component.

The WTPT methodology was applied initially to the Jet Engine Mechanic Specialty (AFS 426X2), and performance measures were constructed for three jobs in that specialty. Each job was defined by the type of engine that a mechanic repaired (TF-33, J-57, and J-79).

An extensive task sampling plan was developed for each job using information obtained from the Air Force's Occupational Survey Program (Lipscomb, 1984). This Program maintains job content domains for over 200 of the 250 enlisted specialties in the Air Force. Surveys are administered approximately every 4 years to keep the job content domains current. Available information in each job content domain includes the tasks performed, the relative amount of time spent performing these tasks, and emphasis given to the tasks in training. This information was used to select tasks for developing WTPT components.

Visits were made to several Air Force bases in order to interview subject-matter experts (SMEs) about these tasks (Alba, Dickinson, & Lipscomb, 1985). They were asked to describe the procedural steps involved in performing each task, whether mechanics differed in their performance on a task, and whether the development of a hands-on test for a task was infeasible (e.g., because of time constraints, potential damage to expensive equipment, or examinee injury). This information was used to revise the list of tasks.

Hands-on and interview tests were written for each of the appropriate tasks. These tests were reviewed by SMEs, and based on their input, the tests were refined. Finally, the tests were field tested at several Air Force bases.

Test Administrator Training

Former jet engine mechanics with extensive work experience as mechanics were hired to collect test data. They administered the hands-on and interview tests and rated the performance of the job incumbents. Teams of three test administrators were assigned to each of the three engine types, such that each three-member team was to assess incumbents only on one engine type. Training occurred in three distinct phases: orientation training, a training workshop prior to field testing, and a retraining workshop between field testing and full-scale collection of data for ASVAB validation.

Orientation training occurred over a 3-month period and focused on familiarization with project goals, measurement instruments, and technical orders and job guides. In addition, test administrators observed the work performance of active-duty jet engine mechanics, performed the WTPT tasks themselves, and gained experience in administering the hands-on and interview tests to the mechanics.

The training workshop lasted 2 days and the retraining workshop 1 day. The training workshop consisted of five major activities: (a) project overview, (b) general requirements for Air Force base visits, (c) presentation and practice of briefing given to the unit commander explaining the purpose of research and testing requirements, (d) WTPT training, and (e) rater training and rating form administration. The retraining workshop focused on activities (d) and (e). Both workshops concentrated on learning and mastering skills necessary to administer WTPT. The WTPT training focused on three major areas: (a) a discussion of test administration requirements; (b) discussion, modeling, and practice of good interviewing techniques; and (c) practice using the hands-on tests of WTPT.

Since the test administrators were experienced jet engine mechanics who had become highly familiar with WTPT during orientation training, workshop training on the hands-on tasks stressed observation and rating skills. Videotapes of jet engine mechanics performing the tasks allowed the test administrators to practice observing and rating performance for the hands-on tests.

Videotapes

Videotapes were constructed for seven tasks for each engine type. Scenarios were generated by consulting SMEs as to where and how performance errors could be made within each task. This information was used to direct eight job incumbents (who were the actors for the videotapes) in performing task steps correctly or incorrectly, and if incorrectly, in telling them what errors were to be made. The scenarios were discussed with each incumbent prior to videotaping. Both correct and incorrect versions of task performance were videotaped and used for training and obtaining ratings. Additional details concerning the development of videotapes can be found in Bierstedt and Hedge (1987).

Data Collection

Data were collected from the three teams of raters at three separate points in time. During the training workshop, the correct or incorrect versions of task performance were shown to the raters. After the presentation of each task performance, the videotape was stopped and the performance of the mechanic was rated independently by each team member. Each step was rated by a member as "yes" or "no" (i.e., the performance was correct or incorrect on that step). Next, the team members compared their ratings to the target scores and discussed among themselves any discrepancies, with the aim of increasing agreement and accuracy of observation and rating within the team. Whenever an incorrect version of task performance was shown, the correct version followed. This viewing of videotapes was completed by a team for all seven tasks, and it required approximately 4 hours to complete. Two and one-half months after the first workshop, the retraining workshop was held, and the 4-hour session of viewing and rating was repeated.

In addition to the two workshops, data were collected in a field test held 2 weeks after the training workshop and 2 months preceding the retraining workshop. This pretest was conducted at three separate Air Force bases (one per engine type), with WTPTs being administered to a total of 14 incumbents per engine. The incumbents were all first-term (13-48 months of active military service) jet engine mechanics randomly selected from the population of first-term mechanics at each of the three bases. For each engine type, nine incumbents were assessed by single raters. The remaining incumbents were rated by the team, allowing an evaluation of interrater reliability. One team member administered the tests to each incumbent, while all three members rated the incumbent's performance separately. The team members alternated in serving the administrator role for the incumbents tested by the team.

III. RESULTS

The focus of the data analysis was threefold: (a) evaluation of interrater agreement at three points in time (training workshop, pretesting, retraining workshop); (b) evaluation of rater accuracy at both workshops; and (c) identification of the types of errors committed by the raters.

Interrater Agreement

A number of correlational indices have been suggested for describing interrater agreement, but certain precautions are warranted when evaluating dichotomous responses (as required with the WTPT procedure). The distributions of dichotomous responses are often skewed, and correlational indices of agreement may be sensitive to this skewing (Jones, Johnson, Butler, & Main, 1983). Consequently, both pairwise correlations and percent agreements were computed between the raters for each task and ratee. The arithmetic averages (across tasks, raters, and ratees) of these

indices are reported in Table 1. The indices suggest that a high level of interrater agreement was obtained for the three teams at all points in time. In addition, agreement tended to improve over time.

Table 1. Interrater Agreement by Workshop, Pretest, and Engine Type

Engine type	Workshop 1	Pretest	Workshop 2
TF-33	.786	.882	.918
	(74.4)	(78.7)	(83.6)
J-57	.813	.947	.886
	(84.9)	(90.0)	(90.0)
J-79	.712	.904	.764
	(76.4)	(85.8)	(84.1)

Note. Primary values reported are correlations; numbers in parentheses are percent agreement values.

Rater Accuracy

Correlational and percent agreement accuracy indices were calculated for each rater on the seven tasks in each workshop. These indices were computed between ratings and target scores. The averages of the indices for each team on the seven tasks are reported in Table 2. Accuracy was quite high for all teams at both workshops.

Table 2. Rater Accuracy by Workshop and Engine Type

Engine type	Workshop 1	Workshop 2
TF-33	.772	.922
	(73.9)	(96.1)
J-57	.794	.907
	(86.8)	(95.0)
J-79	.635	.770
	(67.2)	(78.5)

Rating Errors

Additional insight into rater accuracy can be gained by evaluating types and percentages of errors committed within and across the two training workshops. Borrowing from the concepts of signal detection theory, task-step ratings were analyzed for decision errors (cf. Baker & Schuck, 1975; Lord, 1985). In this theory, a "hit" is defined as a decision by the rater that a behavior occurred when it actually did occur (i.e., for a task step performed correctly, the rater marks "yes"). Conversely, a "miss" is a decision by the rater that a behavior occurred when it really did not (i.e., for a task step performed incorrectly, the rater marks "yes"). In addition, a "false alarm" is a decision by the rater that a behavior did not occur when, in fact, it did occur (i.e., for a step performed correctly, the rater marks "no"). Finally, a "correct rejection" indicates a decision by a rater that a behavior did not occur when, in fact, it did not occur. The analysis of rater accuracy should focus on two types of errors: failure to say

"yes" correctly and failure to say "no" correctly (Lord, 1985). The failure to say "yes" correctly was operationalized by means of the hit rate. Hit rates were computed for each rater for a task by dividing the frequency of hits for that task by the frequency of hits plus misses. Then, the hit rate was subtracted from 1.00 and the result multiplied by 100 to reflect the percentage of task steps for which the rater failed to say "yes" correctly. The failure to say "no" correctly was operationalized using the false alarm rate. These rates were computed by dividing the frequency of false alarms by the frequency of false alarms plus correct rejections, and then multiplying the result by 100 to reflect the percentage of steps for which the rater failed to say "no" correctly.

The arithmetic averages (across tasks, raters, and ratees) for the two measures of rating errors are shown in Table 3. The results indicate that all three teams erred much more in rating "no" than "yes." Moreover, this phenomenon was prevalent in both the initial and retraining workshops. Thus, raters failed to note correct performance (rated "no" incorrectly) much more frequently than incorrect performance (rated "yes" incorrectly).

Table 3. Percentages of 1.0 - Hit Rate (HR) and False Alarm Rate (FAR) by Workshop and Engine Type

	Workshop 1		Workshop 2	
Engine type	1.0 - HR	FAR	1.0 - HR	FAR
TF-33	6.62	38.46	1.32	14.89
J-57	6.67	54.29	2.24	18.92
J-79	17.01	20.00	6.08	41.18
Overall Mean	10.10	37.58	3.27	25.00

Note. Values reported are percentages.

IV. DISCUSSION

In this investigation, high levels of interrater agreement and rater accuracy were obtained in using work sample testing to assess the performance of job incumbents. The results were obtained by hiring former job incumbents to serve as test administrators and raters, and by providing them with intensive training. The use of videotape technology in workshops allowed the raters to practice observing incumbents perform hands-on components of WTPT and practice using test booklets to rate performance. The usefulness of the videotape approach was reflected not only in the high levels of rater accuracy obtained but also in the verbal comments of the raters. After viewing the videotapes, the raters would engage in detailed discussions as to the key behaviors that an incumbent should perform and avoid. The outcome of these discussions was apparently a common "frame of reference" for rating performance.

The use of videotape technology also provides the opportunity for evaluating rater accuracy. Interrater reliability indices can be collected in field testing to reflect the level of rater agreement; however, these indices do not address the more pressing issue of agreement between the performance displayed by incumbents and the rating of that performance. Although the results reported here were aggregated across tasks and raters, a more detailed analysis could provide data for each rater and task in each workshop to diagnose specific rating deficiencies. Furthermore, the use of signal detection theory can provide information on the types of errors that are being committed. Though the workshops in this investigation did not profit from this knowledge, rater attention could have been directed to particular task steps and the nature of the errors made in rating that performance.

An additional use for videotape technology is suggested for data collection designs in which incumbents are rated by a single test administrator. Over extended periods of data collection, test administrators may fluctuate in their rating accuracy. By periodically requiring raters to observe and rate videotapes of incumbent performance, researchers could evaluate interrater agreement and rater accuracy, and if needed, "recalibrate" the raters to reduce fluctuations in their skill.

The raters in this study erred more frequently in observing correct performance than incorrect performance. As former job incumbents, the test administrators knew that incorrect performance by a jet engine mechanic is highly costly to the Air Force, and apparently, this knowledge influenced the value they attached to choosing a "yes" versus "no" response for the task steps. This result is desirable and reasonable, and if replicated for other specialties (e.g., Air Traffic Control Operator), it suggests the desirability of employing former job incumbents as test administrators.

In conclusion, the results of the current investigation indicate that videotape technology can play an important role in achieving high levels of interrater agreement and rater accuracy in a non-laboratory setting. The use of videotape technology provides the researcher with a flexible training and evaluation tool that is highly recommended for work sample testing.

REFERENCES

Alba, P. A., Dickinson, T. L., & Lipscomb, M. S. (1985). Walk-Through Performance Testing documentation for jet engine mechanic (AFS 426X2). Unpublished manuscript.

Baker, E. M., & Schuck, J. R. (1975). Theoretical note: Use of signal detection theory to clarify problems of evaluating performance in industry. Organizational Behavior and Human Performance, 13, 307-317.

Bierstedt, S. A., & Hedge, J. W. (1987, September). Job performance measurement system (JPMS) trainer's manual (AFHRL-TP-86-34, AD-A115 294). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

Borman, W. C., Hough, L., & Dunnette, M. D. (1976). Performance ratings: An investigation of reliability, accuracy, and relationship between individual differences and rater error. Minneapolis: Personnel Decisions Research Institute.

Hedge, J. W. (1984, August). The methodology of Walk-Through Performance Testing. Paper presented at the annual meeting of the American Psychological Association, Toronto.

Jones, A. P., Johnson, L. A., Butler, M. C., & Main, D. S. (1983). Apples and oranges: An empirical comparison of commonly used indices of interrater agreement. Academy of Management Journal, 26, 507-519.

Lipscomb, M. S. (1984, August). A task-level domain sampling strategy: A content-valid approach. Paper presented at the annual meeting of the American Psychological Association, Toronto.

Lord, R. G. (1985). Accuracy in behavioral measurement: An alternative definition based on raters' cognitive schema and signal detection theory. Journal of Applied Psychology, 70, 66-71.

McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of ratings. Journal of Applied Psychology, 69, 147-156.

Nugent, W. A., Laabs, G. J., & Penell, R. C. (1982). Performance test objectivity: A comparison of rater accuracy and reliability using three observation forms (NPRDC-TR-82-30). San Diego, CA: Navy Personnel Research and Development Center.

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. Journal of Applied Psychology, 69, 581-588.

Wigdor, A. K., & Green, B. F., Jr. (Eds.). (1986). Assessing the performance of enlisted personnel. Washington, DC: National Academy Press.

END
DATE
9-88
DTIC