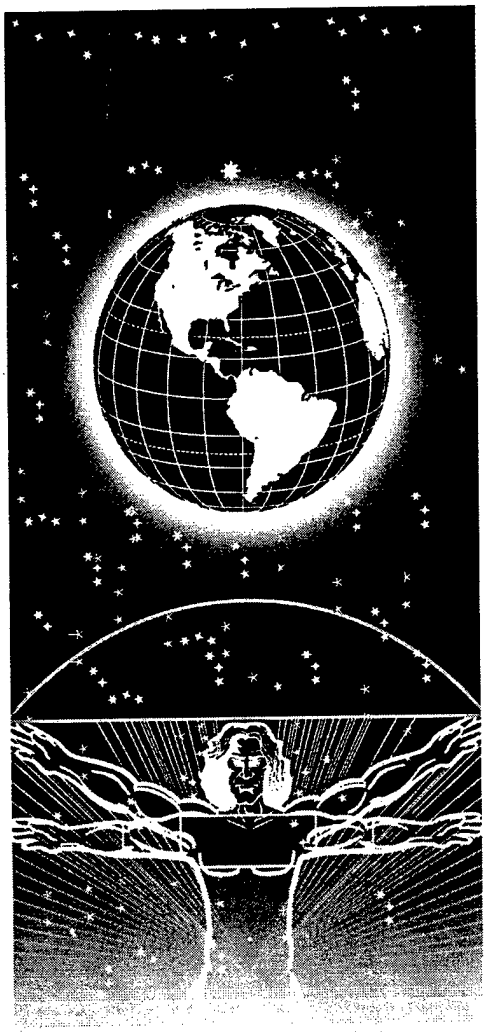
AFRL-HE-BR-TR-1998-0143



UNITED STATES AIR FORCE RESEARCH LABORATORY

GOOD NEWS: WORK SAMPLES ARE (ABOUT) AS VALID AS WE'VE SUSPECTED

Charles E. Lance
C. Douglas Johnson
Shane S. Douthitt

Department of Psychology
The University of Georgia
Athens, GA 30602-3013

Winston Bennett, Jr.

HUMAN EFFECTIVENESS DIRECTORATE
MISSION CRITICAL SKILLS DIVISION
7909 LIndbergh Drive
Brooks AFB, TX 78235-5352

May 1998

AIR FORCE MATERIEL COMMAND
AIR FORCE RESEARCH LABORATORY
HUMAN EFFECTIVENESS DIRECTORATE
7909 Lindbergh Drive
Brooks Air Force Base, TX 78235-5352

NOTICES

This report is published in the interest of scientific and technical information exchange and does not constitute approval or disapproval of its ideas or findings.

Using Government drawings, specifications, or other data included in this document for any purpose other than Government-related procurement does not in any way obligate the US Government. The fact that the Government formulated or supplied the drawings, specifications, or other data, does not license the holder or any other person or corporation, or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

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

This report has been reviewed and is approved for publication.

WINSTON BENNETT, JR
Project Scientist

R. BRUCE GOULD
Acting Chief
Mission Critical Skills Division

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	May 1998	Interim Report

4. TITLE AND SUBTITLE
Good News: Work Samples are (about) as Valid as We've Suspected

5. FUNDING NUMBERS
C – F41624-93-C-5011
PE – 62202F
PR – 1123
TA – A2
WU – 20

6. AUTHOR(S)
Charles E. Lance
C. Douglas Johnson
Shane S. Douthitt
Winston Bennett, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Department of Psychology
The University of Georgia
Athens, GA 30602-3013

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory
Human Effectiveness Directorate
Mission Critical Skills Division
7909 Lindbergh Drive
Brooks AFB, TX 78235-5352

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-HE-BR-TR-1998-0143

11. SUPPLEMENTARY NOTES
Air Force Research Laboratory Technical Monitor: Dr. Winston Bennett, Jr. (210) 536-1981

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release: distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

Data obtained on over 1,500 first-term U.S. Air Force enlisted personnel indicated that work sample administrators' global ratings of work sample performance substantially reflect actual ratee behavior in the work sample, and not potentially biasing factors (e.g., race, gender, amount of recent experience), supporting the "folk wisdom" that work samples are high-fidelity and valid measures of performance.

14. SUBJECT TERMS
Job Performance
Validation
Work Samples

15. NUMBER OF PAGES
28

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Table of Contents

PREFACE

An earlier version of this paper was presented at the 1998 annual meeting of the Society for Industrial and Organizational Psychology in Dallas, TX.

The views and opinions expressed in this paper are those of the authors and do not reflect the official policies or opinions of their respective organizations.

The authors wish to thank the United States Air Force personnel that participated in this project, for these scientific advances would not be possible without their support and cooperation. The authors thank Lillian Eby for her critical and constructive comments on an earlier version of this article. The authors would also like to thank Kathleen Sheehan for preparing the final format of this report.

SUMMARY

Data obtained on over 1,500 first-term U.S. Air Force enlisted personnel indicated that work sample administrators' global ratings of work sample performance substantially reflect actual ratee behavior in the work sample, and not potentially biasing factors (e.g., race, gender, amount of recent experience), supporting the "folk wisdom" that work samples are high-fidelity and valid measures of performance.

GOOD NEWS: WORK SAMPLES ARE (ABOUT) AS VALID AS WE'VE SUSPECTED

INTRODUCTION

A work sample may be defined as "...a measure of performance on a structured task that is directly reflective of the type of behaviors required in the job situation" (Smith, 1991, p. 28). As such, work sample measures may be distinguished from other related performance measurement approaches such as (a) trainability tests, which include a specified time to learn the task to be performed (Robertson & Downs, 1989), (b) situational judgment tests, which typically require the examinee to respond to a hypothetical situation, and which may be administered either in an oral or written mode, and (c) job knowledge tests, which assess declarative (as opposed to procedural) aspects of performance, and which usually are administered in written form. For many years, the work sample has been touted as an effective approach to the measurement of work-related behaviors for the purposes of predicting subsequent on-the-job performance, assessing training effectiveness, and measuring current job proficiency (Smith, 1991). Results of Terpstra's (1996) recent survey document the long-held and widespread belief in the effectiveness of work samples among human resource executives.

The effectiveness of the work sample as a predictor of job success has been documented in several reviews. For example, Asher and Sciarrino (1974) classified work samples as either motor ("...if the task was a physical manipulation of things..." p. 519) or verbal ("...if there was a problem situation that was primarily language-oriented or people oriented." p. 519), and found motor work samples to be second only to biodata in terms of their predictive efficiency; verbal work samples were somewhat less predictive of performance. Later reviews confirm these basic findings: work samples are among the most valid predictors of job performance, having mean validities in the .40s to .50s (Hunter & Hunter, 1984; Robertson & Kandola, 1982; Schmitt, Gooding, Noe, & Kirsch, 1984). M. Smith's (1994) theory of the validity of predictors of job performance suggests reasons why work samples are valid. First, work samples are very effective at assessing specific abilities and specialized skills that are required for the performance of particular jobs. Second, work samples are thought of as objective measures of samples of behavior that are highly representative of actual job duties. Thus, from M. Smith's (1994) theory, the high validity of work samples is seen to arise from the objective assessment of representative job duties that are required for successful job performance in specific jobs.

Work samples have also been held in high regard as criterion measures (Borman, White, & Dorsey, 1995; Kavanagh, Borman, Hedge, & Gould, 1987). For example, Borman and Hallum (1991) noted that "...some researchers have maintained that work samples...are the highest fidelity performance-measurement method available and that they provide the most valid indication of 'actual' performance" (p. 11). Some have even suggested that alternative measures of performance (e.g., performance ratings) might be "validated" in terms of their relationships with work sample measures (Wigdor & Green, 1991a).

Granted, work sample performance measures are high fidelity measures of "can-do" aspects of job performance (versus "will-do" aspects, see Borman et al., 1995; Borman, White, Pulakos, & Oppler, 1991; DuBois, Sackett, Zedeck, & Fogli, 1993; Sackett, Zedeck, & Fogli, 1988). However, they are not without their limitations. First, they can be time consuming, labor intensive, and expensive to develop and operate (Asher & Sciarrino, 1974; Hedge & Teachout, 1992; Hunter & Hunter, 1984; Smith, 1991). Second, although tasks included within work sample test batteries usually represent corresponding on-the-job elements with high fidelity, a relatively small range of job tasks is usually included in them due to the time and expense in developing and operating them. Thus except for some highly specialized jobs (e.g., life guard, toll taker, raisin washer), work samples may suffer more from criterion deficiency (Thorndike, 1949) as compared to alternative criterion measures. Finally, there is some question as to whether work sample criterion measures are really as "objective" as has been presumed. This is the subject of the present study.

F. D. Smith (1991) characterized three typical approaches to scoring performance in a work sample. In one (global rating approach), the work sample administrator observes examinee performance in the work sample and rates the examinee's performance on global, usually Likert-type, scales with anchors such as "Unsatisfactory" to "Exceeds performance standards." This approach is commonly used in scoring work sample task performance (Smith, 1991), and is also the approach that is most often used in the assignment of assessment center post-exercise dimension ratings (Klimoski & Brickner, 1987; Thornton, 1992). Ratings on a number of such scales (e.g., representing different performance dimensions) may be averaged to form an overall score for each work sample task, and ratings may be averaged across tasks to form an overall work sample test battery score. In a second approach, the work sample administrator is provided with behavioral recording forms which list specific examples of good and poor performance (developed by subject matter experts - SMEs) in the work sample that are intended to guide the administrator in observing examinee behaviors and in making summary, global ratings of examinee task performance. Thus in this approach, the work sample administrator still makes only a single global rating, but with the assistance of behavioral exemplars to guide the global performance judgment. Finally, work samples may be scored using a behavioral checklist. In this, probably the least often used approach, the work sample administrator indicates which of a number of prespecified task steps were completed either correctly or incorrectly by the examinee (see e.g., Brugnoli, Campion, & Basen, 1979; Campion, 1972). In this approach, work sample performance is usually scored as some form of percentage of task steps completed correctly for each task. Once again, overall work sample performance may be scored by computing an aggregate score across all tasks included in the work sample test.

Note that work sample administrator judgment is required in each of these scoring strategies. Consequently, it could be argued that none of these scoring schemes is entirely "objective." Despite the high fidelity of work sample tasks, work sample administrators still must make "clinical" global performance judgments of task performance based on their observations of examinee behaviors in the first scoring scheme (i.e., the global rating approach, though these may be aided with behavioral recording forms, as in the second scoring option, the behavior recording forms approach). Even the use of behavioral checklists (as in the third scoring scheme) may require judgment as to whether individual task steps are completed

correctly or incorrectly. For example, Borman and Hallam (1991) found there may not only be disagreement among work sample administrators as to whether particular executions of task performance steps are correct or incorrect, but there may even be more fundamental disagreement among SMEs as to what target behaviors should be scored as correct or incorrect.

Aside from Borman and Hallam's (1991) study, there has been almost no research on the "validity" of work samples as criterion measures. One exception was Hedge and Teachout's (1992) study of the convergence between work sample tasks as administered in a "hands-on" mode as compared to an "interview" mode. Work sample validity has been inferred from the fidelity with which actual job tasks are represented in the content of work sample test batteries. But replicating job functions in a high-fidelity simulated test environment does not insure that the examinee will be evaluated objectively in this environment. Human (i.e., work sample administrator) judgment, has been shown to be subject to a number of biases (Borman, 1991), and may be a significant factor in the measurement of work sample performance, regardless of how work sample performance is scored.

The purpose of this study was to provide an empirical assessment of whether work sample performance measures as they are usually scored (i.e., work sample administrator global ratings of examinee performance on work sample tasks) substantially reflect actual examinee performance in the work sample (as is generally presumed), or whether they may be biased by factors that are not directly related to examinee performance in the work sample. The particular work samples reported here afforded a unique opportunity to test these ideas, as data were collected on (a) work sample administrators' global task performance ratings (i.e., as in the global rating approach described by Smith, 1991), (b) whether discrete task steps that comprised the work sample tasks were completed correctly or incorrectly (i.e., as in the behavioral checklist approach described by Smith, 1991), and (c) a number of additional factors related to task performance (e.g., time to complete work sample tasks, previous experience performing the tasks, examinee demographic characteristics, etc. These are described in greater detail later.). We predicted that if work sample administrators' global task performance ratings are as valid as they have been presumed, then they should substantially reflect actual examinee behavior in the work sample and not the influences of other factors that are less directly related to task performance. To our knowledge, only one previously published study has attempted to address this issue (Brugnoli et al., 1979), in which it was found that global ratings, but not behavioral checklist measures of work sample task performance was subject to racial bias. However, this was a small-sample ($N = 46$) laboratory study in which work sample performance was depicted only in brief videotaped segments that showed only the examinees' arms and hands. So, in addition to the main focus of our study, we also were interested in the extent to which Brugnoli et al's. (1979) results could be replicated in a much larger sample and more ecologically valid measurement context.

Summary and Specific Predictions

Work samples have long been touted as "objective" performance measures, yet very little research has investigated their ostensible objectivity. As often scored (i.e., in terms of global ratings of task-level performance; Smith, 1991), work sample ratings may actually be subject to

many of the same biases as other ratings (e.g., supervisory job performance ratings), despite the high fidelity of the simulated performance situation. That is, even in high-fidelity measurement situations work sample global task ratings, like supervisory performance ratings, may reflect non-performance-based information as well as performance-based information available to the work sample administrator. We predicted that if global work sample ratings are as objective and valid as they have been presumed, then they should substantially (and perhaps exclusively) reflect actual examinee behavior in the work sample. In the present context, this would correspond to the situation in which examinee step-level behavior (i.e., percentage of task steps completed correctly) accounted for most or all of the predictable variance in global work sample ratings. As we explain in greater detail later, a number of steps were taken in the development of the work samples reported here to insure that task step-level measures were as objective as may be possible in operational work sample tests. If, on the other hand, global work sample ratings are as subject to effects of rating biases as other types of performance ratings (Borman, 1991), then significant portions of variance in global work sample ratings should be accounted for by other variables that are not directly related to performance in the work sample. Thus the present research was designed as a policy-capturing study (Cooksey, 1996; Hoffman, 1960) of the extent to which information which is available to the rater during work sample administration (information which relates both to performance-related and performance-irrelevant factors), combines to affect raters' overall judgments of work sample task performance. The question addressed in this study is the extent to which work sample administrators' overall judgments of examinee task performance substantially reflect performance-relevant information (i.e., actual examinee behavior in the work sample) as has been assumed, or whether information relating to other, potentially peripheral factors might also have nonminor effects on work sample overall task performance ratings. To date, there has been almost no research on this question.

METHOD

Study Context

Data were collected as part of a large-scale Joint-Service Job Performance Measurement (JPM)/Enlistment Standards project conducted by the U.S. military in the late 1980s and early 1990s (Wigdor & Green, 1991a, 1991b). The major purposes of this project were to link enlistment standards to on-the-job performance and to explore alternative technologies for measuring job performance. The conceptualization, design, and execution of the JPM Project has been discussed extensively elsewhere (Hedge & Teachout, 1986, 1992; Kavanagh et al., 1987; Lance, Teachout, & Donnelly, 1992; Laue, Hedge, Wall, Pederson, & Bentley, 1992; Ree, Earles, & Teachout, 1994; Teachout & Pellum, 1991). Thus only the particular aspects of the JPM Project that are relevant to the present study are highlighted here.

Samples

Samples were obtained from eight U.S. Air Force (USAF) specialties (AFSs) selected for inclusion in the JPM Project. These included Aircrew Life Support Specialist, $n = 229$; Air Traffic Control Operator, $n = 190$; Precision Measurement Equipment Laboratory Specialist, $n = 140$; Avionic Communication Specialist, $n = 98$; Aerospace Ground Equipment (AGE)

Mechanic, n = 269; Jet Engine Mechanic, n = 255; Information Systems Radio Operator, n = 155; and Personnel Specialist, n = 200. These AFSs were selected to be representatative of (a) the relatively more populous jobs in the enlisted occupational classification structure in existence at the time of data collection, (b) varying levels of occupational learning difficulty (see Burtch, Lipscomb, & Wissman, 1982; Mumford, Weeks, Harding, & Fleishman, 1987; Weeks, 1984), and (c) existing accession and classification policies based on mechanical, administrative, general, and electronic (MAGE) aptitude requirements (Department of Defense, 1984; i.e., two AFSs were chosen to represent each of the four MAGE aptitude areas).

Work Sample Task Selection

A number of criteria were used to select tasks for inclusion in each JPM AFS's work sample test battery. First, occupational survey (i.e., job analysis) data were analyzed to identify those tasks that were most widely performed by first-term incumbents in the respective AFSs. Second, tasks were selected from among the most frequently performed tasks to insure that most examinees would have some experience performing tasks included in the work sample test battery previously on the job. Third, tasks were also selected so as to reflect a range of task learning difficulties. Specifically, 40% of the work sample test battery tasks were sampled from the fourth quartile of task learning difficulty (i.e., the most difficult), 30% from the third quartile, 20% from the second quartile, and 10% from the first (least difficult) quartile, in order to reduce the likelihood of ceiling effects in work sample performance. Fourth, candidate tasks were reviewed by SMEs in "task validation workshops" (see Laue et al., 1992) to insure that constituent task steps were observable and that they could be scored unambiguously as being completed correctly or incorrectly. The purposes of this selection criterion were to insure that (a) discrete task steps were directly identifiable and observable by work sample administrators, and (b) performance on each task step could be scored as correct or incorrect according to specified criteria, thus minimizing potential ambiguities in scoring work sample test items that Borman and Hallam (1991) had identified earlier. Table 1 shows one example of a work sample task included here (Installation of engine pressure ratio probes for the Jet Engine Mechanic AFS) and constituent task steps, each of which was scored on a correct/incorrect or "go/no-go" basis. Candidate tasks whose steps were either not easily observable or scorable on to a "go/no-go" basis were replaced with alternate tasks. Combined with the extensive work sample administrator training that was given prior to administration of the work sample tasks (described below), these task selection criteria insured that evaluation of work sample performance at the step-level was as objective as may be possible in an operational work sample test battery.

Altogether, between 20 and 46 tasks per AFS were selected for the work sample task batteries. In most AFSs, some tasks were widely performed by all incumbents (referred to as "Phase I" tasks), while others ("Phase II" tasks) were performed only by incumbents in particular functional areas. For example, Jet Engine Mechanic Phase I tasks were commonly performed by all Jet Engine Mechanics, but Phase II tasks varied as a function of the particular type of jet engine that the incumbent serviced. In order to maximize sample sizes, only Phase I tasks were included in this study, resulting in the retention of between 8 and 31 tasks per AFS for analysis (see Table 3, below). Retention of Phase I tasks only insured that the work domains represented

here were content valid for all members of respective AFSs and not merely a more specialized subset.

TABLE 1: Example Work Sample Task and Constituent Task Steps

Task: Installation of engine pressure ratio probes (Task # 359)
Task Steps:
1. Insert the pressure sensing probes into the turbine exhaust case.
2. Install the bolts and nuts into the turbine exhaust case bosses.
3. Connect the tube and manifold assemblies into the sensing probes.
4. Torque the probe nuts.
5. Torque the manifold and probe connection B nuts.
6. Install the safety device on the B nut.
7. Install the brackets and clips to the rear turbine exhaust case.

Work Sample Administrator Training

Work sample tests were administered by active-duty or recently retired noncommissioned officers from the respective AFSs. Administrators received 1 - 2 weeks of intensive training in the observation and scoring of the work sample tests. Training included procedures for work sample administration, observation of examinee performance, and work sample scoring procedures. Training methods included lecture and discussion, role playing, and viewing and discussing videotaped target task performances. Videotaped task performances were scripted to reflect both correct and incorrect step-level performances, and to establish a common frame of evaluative reference among the test administrators. Inter-administrator reliability was estimated at .81 to .98 (see Hedge, Dickinson, & Bierstedt, 1988; Hedge & Teachout, 1992, for additional details).

Procedure

Upon arriving at the test station, examinees were briefed as to the general purpose of the work sample test and were administered an appropriate work sample test battery. Examinees were instructed and encouraged to do their best on each work sample task. Testing required 4 - 8 hours per examinee. For each work sample task, the work sample administrator recorded (a) incumbent-estimated number of times s/he had performed the task previously on the job ("Number of Times Performed"), (b) how long it had been (in weeks) since s/he had last performed the task ("Last Time Performed"), and (c) time of day at the beginning of the task administration. Next, the administrator administered the work sample task to the examinee, observed examinee task performance, and recorded whether each task step was completed correctly.[1] Third, the administrator recorded the time at the completion of the task and the total

time required to complete the task ("Time Required"). Finally, the administrator completed a global rating of task performance ("Overall Performance:" "1 = Far below the acceptable level of proficiency," to "5 = Far exceeded the acceptable level of profiency"). These four steps were repeated for each task in the work sample test battery. However, the second step (i.e., task administration) occurred in two different modes: hands-on and interview. In the hands-on mode, examinees were instructed to perform the task as they would on the job, and were allowed access to technical manuals and other written materials as they would ordinarily on the job. In the interview mode, examinees were asked to describe the steps necessary for task completion in a "show and tell" manner, but without the aid of technical manuals or other information (see Hedge & Teachout, 1992). Some work sample tasks were administered in the hands-on mode only, some in the interview mode only, and some in both (referred to by Hedge & Teachout, 1992 as "overlap tasks"). For overlap tasks, the interview mode of administration always preceded the hands-on administration of the work sample task. We included both hands-on and interview work sample tasks for analysis.

Measures

Overall Performance (OAP) was the work-sample administrator's global 5-point rating of work sample task performance, and was the primary criterion variable in this study. Note that this measure is typical of many work sample task-level performance measures, and exemplifies the overall work sample task ratings obtained in the "global rating" and "behavioral recording forms" approaches to scoring work samples described by F. D. Smith (1991).

Percent Steps Correct (%Correct) was measured as an unweighted percentage of task steps completed correctly as recorded by the work sample administrator. Note that this measure is typical of the "behavioral checklist" approach to scoring work sample task performance as described by F. D. Smith (1991). As such, it provides perhaps the closest possible link, particularly with the task selection and administrator training safeguards implemented in the work sample test batteries reported here, between measured task performance and actual examinee behavior in the work sample situation. The high interscorer (i.e., "shadow score") reliabilities reported earlier also are testimony to the objectivity of these measures. We predicted that %Correct would be positively related to OAP, and if OAP-type ratings are as objective and valid as has been presumed, that %Correct would account for substantially all of the predictable variance in OAP. Otherwise, we expected that OAPs might also reflect substantial influences of one or more of the following variables which relate more peripherally to actual performance in the work sample.

Number of Task Steps (#STEPS). As mentioned earlier, each work sample task consisted of a number of discrete task steps which were identified from the respective AFSs' technical and training manuals ("technical orders"). The number of constituent task steps ranged between 2 and 47. #STEPS can be considered as an indicator of task complexity. We expected that significant OAP -- #STEPS relationships would be negative, that is, that performance would generally be rated lower on more (versus less) complex tasks, as more complex tasks would be generally perceived as being more difficult.

Time to Complete Task (TIME), measured in minutes, was the difference between the work sample task finish time and start time. For cases in which the examinee did not finish the task within the pre-established time limit, TIME was set equal to the time limit. We expected that, all other things equal, OAPs would be higher for quicker (and perhaps more expertly executed, versus slower) task performances.

Last Time Performed (LTP), was computed as the number of weeks since the task had last been performed as part of the examinee's regular job duties. Thus LTP indicated the length of the interval in between the time the task was last performed and the time it was tested in the work sample (Lance, Parisi, Bennett, Teachout, Harville, & Welles, 1998). All other things equal, we expected higher OAPs for cases in which the task had been performed on the job more recently, as more recent experience might be expected to facilitate task performance.

Number of Times Performed (NTP) were incumbents' reports of the number of times they had previously performed the task on the job as part of their regular job duties. Previous research (e.g., Lance et al., 1989; Lance et al., in press) has found that NTP is markedly positively skewed and multimodal. Thus we transformed it (as in previous studies) as 1 = Never performed, 2 = 1 to 10 times performed previously, 3 = 11 to 20 previous performances, 4 = 21 to 50, 5 = 51 to 100, 6 = 101 to 800, and 7 = 801 to 999 previous performances ("999" indicated that the examinee had performed the task so often that they could not estimate the number of previous performances). We expected positive OAP -- NTP relationships, that is, higher OAPs for cases in which the task had been performed often previously, as more experienced examinees might be expected to perform more effectively than less experienced ones.

Examinee Motivation (MOT) to perform effectively in the work sample test was measured as a composite of six items anchored by 5-point Likert-type scales. These items were included on a questionnaire that was completed by the work sample examinee immediately after completing the work sample test battery. Example items included "Are you satisfied that you performed as well as you could on the (work sample) test?" and "How motivated were you to perform to the best of your ability on the (work sample) test?" Standardized coefficients alpha ranged between .81 and .86 across AFSs[2]. We predicted positive OAP -- MOT relationships on the basis that more motivated performance may serve as a cue to performance effectiveness (Martell, Guzzo, & Willis, 1995).

Demogaphic Variables. Sex was scored as Male = 1 and Female = 0. Personnel records included three racial codes for "White," "Black," and "Other." We recoded race as two binary variables: White (= 1, 0 = Nonwhite), and Black (= 1, 0 = Nonblack). We included these factors because gender and racial biases in performance measures have been found previously (e.g., Brugnoli et al., 1979; Ford, Kraiger, & Schechtman, 1986; Hamner, Kim, Baird, & Bigoness, 1974; Tosi & Einbender, 1985), although their effects are often minimal or nonexistent under performance measurement conditions such as in the present study (Pulakos, White, Oppler, & Borman, 1989; Sackett & DuBois, 1991; Tosi & Einbender, 1985).

Data Analyses

We performed two complementary sets of analyses. Both were aimed at determining what information that is available during work sample administration impacts administrators' OAP ratings. That is, both analytic approaches were directed toward capturing work sample administrators' OAP rating policies (Cooksey, 1996). In the first, we used ordinary least squares (OLS) multiple regression to regress the global task performance rating (OAP) for each task on %Correct in the first step, and in the second step, also on TIME, LTP, NTP, MOT, Sex, White, and Black. We entered TIME, LTP, NTP, MOT, Sex, White, and Black after entering %Correct, because some of these variables could be considered as performance determinants (e.g., task experience (indexed by NTP), and examinee motivation (MOT) should, theoretically, enhance task performance). Thus the effects of these variables on OAP should be considered as peripheral only to the extent that their effects on actual work sample task performance have already been controlled. Thus we controlled for these effects by entering %Correct into the policy-capturing equation first, followed by the remaining variables in step 2. We evaluated the change in R^2 (i.e., ΔR^2) from the first to the second step to investigate the statistical and practical significance of the variables included in the second step.

Altogether, we performed 134 such hierarchical regressions corresponding to the total number of Phase 1 tasks included in all eight AFSs. Sample sizes for each regression equation varied across AFSs (as were reported earlier). Support for the "validity" of the OAPs would be obtained if %Correct accounted for a substantial proportion of variance in OAP, and if the remaining variables accounted for very little variance in OAP beyond that which was accounted for by %Correct. Bias in OAPs would be indicated to the extent that one or more of the additional variables accounted for a substantial proportion of variance in OAP beyond that accounted for by %Correct.

The second analytic strategy combined data for all 134 tasks into a single "stacked" multi-level data set. This data set was multi-level in the sense that variables were operationalized at varying levels of specificity. For example, the study's dependent variable (OAP) indexed the ith examinee's (i --> N_k as reported earlier for each of the k --> K = 8 samples) performance on the jth work sample task (j --> J_k, J_k ranged between 8 and 31). Thus, the effective sample size was $\Sigma(N_k * J_k)$ = 14,965 after the deletion of missing data. %Correct also varied both across examinees and tasks, as did TIME, LTP, and NTP. Thus, OAP, %Correct, TIME, LTP, and NTP were task x examinee-level variables. On the other hand, #STEPS varied across the j tasks, but was constant for all N_k performers of jth task. Thus #STEPS was a task-level variable. Finally, examinee motivation (MOT), Black, White, and Sex were three examinee-level variables, as they varied appropriately across the N_k examinees, but were constant for the ith examinee across his/her performance of the J_k tasks attempted in the work sample test battery.

Again, the primary question investigated here was whether variance in work sample administrators' global task performance ratings (i.e., OAPs) is substantially predicted by examinees' actual behavior in the work samples (i.e., %Correct), or whether additional variance in OAPs is also explained on the basis of additional, potentially peripheral factors. We tested

this in the "stacked" multi-level data set using OLS to regress work sample administrator OAPs on %Correct in the first step and, in a second step, on #STEPS, TIME, NTP, LTP, White, Black, Sex, and MOT.

We also explored possible interactions between %Correct and an additional binary variable indicating whether the task was administered in the interview (=0) or hands-on (=1) mode ("H/I"), and the additional predictors, as Hedge and Teachout (1992) indicated that mode of administration may impact factors related to task performance. To do this we first centered %Correct, H/I and the remaining predictors (i.e., to a mean of zero), and then formed cross-products between %Correct and H/I and the additional predictors (e.g., %Correctx#STEPS, H/IxLTP, etc.). Finally, we entered these cross-product terms into the OAP regression equation in a third step. However, since we had no a priori predictions regarding interaction effects, we entered the cross-product terms using forward selection with an $\alpha < .05$ entry criterion. Finally, results reported later suggested that the form of the %CorrectxTIME interaction might vary between hands-on and interview tasks. We tested this by entering the 3-way H/Ix%CorrectxTIME interaction in a fourth step in the regression model.

RESULTS

Table 2 shows study variables' descriptive statistics and intercorrelations for all AFSs combined.[3] Mean OAP and %Correct values indicated the absence of ceiling effects and their SDs indicated that range restriction was not a problem. The mean NTP indicated that, on the average, examinees were experienced performing the tasks on which they were examined in the work sample test battery, but the mean LTP indicated that, on the average, it had been about 3 1/2 months since they had last performed the tasks included in the work sample on the job. MOT scores generally indicated that examinees were in fact motivated to perform well in the work sample. Finally, data in Table 2 show that the total sample was 80% White, 13% Black, and 85% Male.

Table 2 also shows that, as predicted, OAP was positively correlated with %Correct. However, NTP, TIME, LTP, MOT, and #STEPS also were significantly correlated with OAP, and in the hypothesized directions. Notably, correlations among most predictor variables were quite low (but statistically significant, due to the extremely high power afforded by the combined samples' size), and exceptions are easily understood. For example, (a) r(TIME,#STEPS) = .52 indicates that, on the average, it takes longer to perform tasks that have more constituent task steps, (b) r(NTP,TIME) = -.18 indicated some tendency for more experienced examinees to perform task more quickly, (c) r(NTP,LTP) = -.37 indicated that more experienced examinees also tended to have more recent experience on tasks in the work sample test battery, and (d) r(TIME.H/I) = .38 indicated that it took examinees somewhat longer (on the average) to actually perform hands-on tasks than it did for them to explain how they would perform tasks as administered in the interview mode. Also notable is the fact that correlations between demographic and more substantive variables are near zero, and many are statistically nonsignificant. This reinforces previous research indicating that when racial and gender biases are found, their effects are often quite small (Pulakos et al., 1989; Tosi & Einbender, 1985;

Sackett & DuBois, 1991). Finally, correlations with H/I indicated that there was some tendency for examinees to obtain higher performance scores on hands-on tasks as compared to tasks administered in the interview mode. Tables 3 through 5 address the study's main questions more directly.

TABLE 2: Study Variables' Descriptive Statistics and Intercorrelations

Variable	Mean	SD	Variable									
			1	2	3	4	5	6	7	8	9	10
1. OAP	2.48	1.19	1.00									
2. % Correct	.67	.29	.77	1.00								
3. NTP	3.50	2.10	.33	.38	1.00							
4. TIME	6.65	6.97	-.18	-.18	-.18	1.00						
5. LTP	14.82	24.98	-.14	-.12	-.37	.11	1.00					
6. MOT	3.80	.66	.07	.06	.04	-.01*	-.03	1.00				
7. # STEPS	12.01	7.06	-.14	-.10	-.11	.52	.10	-.02*	1.00			
8. White	.80	.40	-.01*	-.02	-.01*	.05	-.01*	-.05	.03	1.00		
9. Black	.13	.34	.02	.02	.02	-.04	.01*	.03	-.03	-.80	1.00	
10. Sex	.85	.36	-.05	-.04	-.01*	.08	.04	.04	.06	.12	-.14	1.00
11. H/I	.55	.50	.09	.14	.03	.38	.01*	.00*	.12	.00*	.00*	.00*

Note. OAP = Overall Performance Rating; %Correct = Percentage of Task Steps Completed Correctly; NTP = Number of Times Performed; TIME = Time to Complete work sample task; LTP = Last Time Performed; MOT = Examinee Motivation; Black and White (= 1, versus Other racial groups = 0); Sex (1 = Male, 0 = Female). *$p > .01$.

Table 3 shows the percentages of regression equations from the first set of analyses in which each variable was a statistically significant (i.e., $p < .05$) predictor of OAP. Numbers outside (inside) parentheses indicate the percentages of times that each predictor was statistically significant and was signed in the predicted (opposite) direction. For example, the first row of Table 3 shows that of the 19 regression equations for the Avionics Communication sample (i.e., one equation for each work sample task), %Correct was a statistically significant (and properly signed) predictor of OAP in 100% (i.e., all 19) of the equations; NTP was a statistically significant (and properly signed) predictor in 15.8% of the equations; TIME was a statistically significant (and properly signed) predictor in 36.8% of the equations but a statistically significant (and oppositely signed) predictor in 10.5% of the equations, and so forth. The last row summarizes the mean percentages across all samples. The first column of Table 3 shows that %Correct was a significant predictor of OAP in nearly every regression equation, and in no case was the effect of %Correct on OAP estimated to be statistically significant and negative. The last entry in the second column indicates that NTP was a statistically significant (and properly signed) predictor of OAP in 9% of the estimated equations, but in 1.5% of the equations the coefficient was statistically significant but negative (contrary to predictions). Table 3 also shows

that overall, LTP, MOT, Black, White, and Sex were "significant" predictors of OAP at, or well, below chance levels. Interestingly however, TIME was a significant predictor of OAP in a total of 29.1% of the regression equations, but in many cases (17.9% of the equations) its coefficient was negative (as was predicted) and in others (11.2% of the equations), the coefficient's sign was positive.

TABLE 3: Percentages of Statistically Significant Regression Weights for Predictors of OAP Ratings

Job	No. of Tasks	%Correct	NTP	TIME	LTP	MOT	Black	White	Sex
Avionics Communication	19	100.0 (0.0)	15.8 (0.0)	36.8 (10.5)	0.0 (0.0)	0.0 (0.0)	5.3 (0.0)	0.0 (5.3)	0.0 (0.0)
Air Traffic Control	14	100.0 (0.0)	0.0 (7.1)	14.3 (0.0)	0.0 (7.1)	N/A[a]	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
AGE Mechanic	31	100.0 (0.0)	6.5 (0.0)	0.0 (32.3)	3.2 (6.5)	0.0 (3.2)	0.0 (0.0)	0.0 (3.2)	3.2 (3.2)
Personnel	11	100.0 (0.0)	9.1 (0.0)	27.3 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (9.1)
Precision Measurement	21	100.0 (0.0)	4.8 (0.0)	52.4 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
Aircrew Life Support	14	78.6 (0.0)	28.6 (0.0)	7.1 (7.1)	21.4 (0.0)	28.6 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
Jet Engine Mechanic	8	100.0 (0.0)	0.0 (0.0)	0.0 (12.5)	0.0 (0.0)	N/A	0.0 (0.0)	0.0 (12.5)	0.0 (0.0)
Information Systems Radio Operator	16	100.0 (0.0)	6.3 (6.3)	0.0 (6.3)	6.3 (12.5)	0.0 (0.0)	6.3 (0.0)	0.0 (6.3)	0.0 (0.0)
Mean	134	97.8 (0.0)	9.0 (1.5)	17.9 (11.2)	3.7 (3.7)	3.0 (0.7)	1.5 (0.0)	0.0 (3.0)	0.7 (1.5)

Note. OAP = Overall Performance Rating; %Correct = Percentage of Task Steps Completed Correctly; NTP = Number of Times Performed; TIME = Time to Complete work sample task; LTP = Last Time Performed; MOT = Examinee Motivation; Black and White (= 1, versus Other racial groups = 0); Sex (1 = Male, 0 = Female). Numbers outside (inside) parentheses represent the percentages of regression coefficients that were statistically significant and in the hypothesized (opposite) direction.

[a]Examinee motivation was not assessed in the Air Traffic Control of Jet Engine Mechanic samples; see Footnote 2.

To try to pinpoint the reason for why TIME's coefficient was sometimes negative and sometimes positive, we summarized regression equations separately for hands-on and interview work sample tasks. For the 25 equations in which TIME (in addition to %Correct) was a statistically significant predictor of hands-on task OAPs, its coefficient was negative (as was predicted) in 20 (80%) of them. However, for the 14 equations in which TIME was a statistically significant predictor of interview task OAPs, its coefficient was positive (opposite to that predicted) in 10 (71%) of them. This difference in patterns of relationships between OAPs and TIME between hands-on and interview tasks was itself statistically significant: $\chi^2(1) = 10.06$, $p < .01$. That is, controlling for %Correct, administrators gave somewhat higher OAP ratings for quicker performances in hands-on tasks, and higher OAP ratings for slower performance in interview tasks. We interpret this as indicating that administrators were giving "extra credit" for quickly and smoothly-executed hands-on performances, and for more detailed and thorough (though slower) "show-and-tell" explanations of task performance in interview tasks.

On the whole, however, %Correct overshadowed every other predictor in accounting for variance in OAP ratings. This conclusion is further reinforced in Table 4 which shows mean R^2 and β (i.e., standardized regression coefficient) values ($\pm$ 1 SD) calculated across the 134 regression equations (values were converted to zs, averaged, and backtransformed to R^2s and βs). The mean R^2 (.54) approaches the reliability of global performance ratings as cited by Viswesvaran, Ones, and Schmidt (1996). That is, %Correct accounts for nearly all of the variance in OAP that could potentially be accounted for, given Viswesvaran et al.'s (1996) estimates of the reliability of performance ratings. Second, %Correct accounts for 88% of the variance in OAP that, on the average, is accounted for in the full regression equations (i.e., $\beta^2/R^2 = .69^2/.54 = .88$). Thus OAPs substantially reflect the influence of examinee behavior in the work sample (%Correct) and not the effects of additional factors that are more peripherally related to performance in the work sample.

TABLE 4: Mean R^2s and Standardized Regression Weights (βs)

	R^2	% Correct	NTP	TIME	LTP	MOT	Black	White	Sex
Mean	.54	.69	.04	-.03	.01	.02	-.01	-.01	.00
-1SD	.35	.48	-.04	-.17	-.07	-.04	-.09	-.10	-.06
+1SD	.68	.82	.12	.11	.09	.08	.07	.08	.06

Note. OAP = Overall Performance Rating; %Correct = Percentage of Task Steps Completed Correctly; NTP = Number of Times Performed; TIME = Time to Complete work sample task; LTP = Last Time Performed; MOT = Examinee Motivation; Black and White (= 1 versus Other racial groups = 0); Sex (1 = Male, 0 = Female).

Results from the second set of analyses complement and extend these findings. The overall β for %Correct shown in Table 5 ($\beta = .759$) is on the same order as the mean β for %Correct reported in Table 4 (.69). And although the variables added in Step 2 of the regression model explained a statistically significant proportion of variance in OAP above and beyond that which was predicted by %Correct ($\Delta R^2 = .006$, $F = 24.92$, $p < .001$), %Correct alone accounted for 99% of the variance explained on the basis of the Step 2 regression equation (i.e., .577/.583 = .9897). Nevertheless, effects of the additional variables, although small, were in the predicted directions. All other things equal, OAPs were somewhat higher for (a) examinees who reported as having been more motivated to perform well in the work sample (effect of MOT), (b) tasks with fewer steps (#STEPS, i.e., simpler tasks), (c) examinees who had performed the task on the job more recently (LTP), (d) examinees who had performed the task more often (NTP), and (e) examinees who performed tasks more quickly (TIME). There also were small effects favoring Blacks and Whites (versus "Other" groups) and against Males. However, all of these additional effects (i.e., beyond the effect of %Correct on OAP) must be interpreted in the contexts that (a) collectively, they account for only about 1% of the variance explained on the basis of the Step 2 regression model, and (b) these effects would likely remain undetected except for the extremely high statistical power afforded here by the large effective sample size.

A number of statistically significant 2-way interaction effects also were detected which, collectively, accounted for an additional 1.1% ($F = 57.18$, $p < .001$) of the variance in OAP. Again, most of these effects were small, and were detectable only by virtue of the extremely high power afforded by the large effective sample size in the second set of analyses. The %Correct x TIME interaction indicated that OAPs were low for low values of %Correct regardless of the amount of time taken to perform the task, but for higher values of %Correct, OAPs were higher for task performances that were executed more quickly than for slower task executions - raters "gave extra credit" to effective task performances that were also executed quickly. The remaining interactions with %Correct followed the same general pattern: raters "gave extra credit" for effective task performances (i.e., high values of %Correct) (a) that occurred on more complex (more task steps) versus simpler tasks (fewer task steps, the PCx#STEPS interaction), (b) for individuals who had performed the task more often (vs. less often) previously (the PCxNTP interaction), and (c) for individuals who had performed the task relatively recently (i.e., the PCxLTP interaction). Lastly, the 2-way H/IxTIME interaction indicated a positive relationship between TIME and OAPs (longer performance times were associated with higher ratings) for tasks administered in the interview mode, while this relationship was nil for tasks administered in the hands-on mode.[4]

TABLE 5:

Variable	β	t-ratio	R^2	F	ΔR^2	F
Step 1:						
% Correct (PC)	.759	142.86***	.577	20,410.05***	---	---
Step 2:						
MOT	.021	3.89***				
#STEPS	-.025	-4.01***				
LTP	-.034	-6.08***				
NTP	.032	5.41***				
TIME	-.022	-3.16**				
H/I	.006	n.s.				
Black	.018	2.01*				
White	.027	2.99**				
Sex	-.015	-2.80**	.583	2,092.72***	.006	24.92***
Step 3:						
PCxTIME	-.091	-13.96***				
PCx#STEPS	.093	14.55***				
PCxNTP	.038	6.08***				
PCxLTP	-.024	-4.16***				
H/IxTIME	-.061	-5.54***	.594	1,286.95***	.011	57.18***
Step 4:						
PCxH/IxTIME	-.070	-7.94***	.596	1,223.99***	.002	62.86***

Note. OAP = Overall Performance Rating; %Correct = Percentage of Task Steps Completed Correctly; MOT = Examinee Motivation; #STEPS = number of constituent task steps; LTP = Last Time Performed; NTP = Number of Times Performed; TIME = Time to Complete work sample task; H/I = hands-on (=1) vs. interview (=0) administration mode; Black and White (= 1, versus Other racial groups = 0); Sex (1 = Male, 0 = Female). * $p < .05$, ** $p < .01$, *** $p < .001$.

Finally, Table 5 shows that a 3-way interaction was supported between H/I, %Correct, and TIME which accounted for an additional .2% of the variance in OAPs. Consistent with findings from the first set of analyses, this 3-way interaction indicated that (a) for tasks administered in the interview mode, raters gave somewhat higher ratings for effective task performances (high %Correct) when time to perform the task was longer (versus shorter), but (b) for tasks administered in the hands-on mode, raters gave somewhat higher ratings for effective task performances when time to perform the task was shorter (versus longer). These results complement earlier findings indicating that raters gave "extra credit" for more detailed and thorough (though slower) "show-and-tell" explanations of task performance in interview tasks, and for quickly and smoothly-executed performances in hands-on tasks.

DISCUSSION

Combined, results in Tables 2 through 5 indicated that work sample administrator OAP ratings (a) substantially reflect the influence of examinee behaviors exhibited in the work sample, as indexed by %Correct, (b) do not reflect racial or gender biases of any practical consequence, (c) are largely independent of potentially biasing effects of administrator prior knowledge of previous experience (indexed by NTP), recent experience (indexed by LTP) and possible performance-cue effects of ratee motivation (MOT), but (d) may reflect subtle stylistic aspects of performance (automaticity of task execution or thoroughness of explanation) that are not captured in a simpler count of the number of task steps that were completed correctly (differential effects of TIME on OAPs for hands-on vs. interview tasks). Thus in one sense, the OAP ratings might be considered more valid than simple %Correct measures (or at least as more encompasing), since they tend not to be biased by peripheral information, and they tend to reflect qualitative aspects of performance that are not tapped by a %Correct measure. That is, results suggest that work samples, as they are often scored, are (about) as valid as has been presumed. However, we urge caution in generalizing the current findings to all work samples too readily.

First, we know of only three other studies to bear on the issue of work sample validity (Borman & Hallam, 1991; Brugnoli et al., 1979; Hedge & Teachout, 1992), so although empirical evidence is encouraging, it is still very limited. Second, the present results stem from work sample test batteries that were developed using "state of the technology" precision. Every step in the work sample test battery development and administrator training followed from scientifically established principles in the job analytic, psychometric, and performance appraisal literatures. In this sense, the present research context may be as good as it gets, and our findings should not be generalized to other settings in which work sample development follows more ad hoc procedures.

Third, the work sample measurement process in the current study was actually a combination of the scoring schemes described earlier by F. D. Smith (1991), and most closely resembled the behavioral recording forms approach in which the recording of task step-level performance information assists accurate OAP ratings. Consequently, our findings should not be readily generalized to situations in which only OAP ratings are obtained. Nevertheless, the present study's findings are the first to suggest that these ratings really are as valid as has been presumed.

Finally, our findings should not be generalized to other performance measurement situations that bear some similarities to the work samples studied here. For example, many assessment center (AC) exercises bear resemblances to work samples, and post-exercise dimensional ratings (PEDRs) often closely resemble the OAPs reported in the present study. PEDRs typically are made using the "global rating" approach discussed by F. D. Smith (1991) in which summary judgments of (dimensional) performance are made following the completion of task (i.e., exercise) performance. However, AC exercises are usually much less structured (e.g., in terms of the specification of intermediate performance steps) than the work sample items investigated here, and we know of no research that has linked PEDRs to actual assessee

behaviors as the OAPs were in the present study. We see this as a need for future, related, research.

Nevertheless, our findings seem to lend some reassurance to one of our "folk assumptions" regarding work samples. Work samples, at least in the form that we investigated here, appear to be (about) as valid as has been assumed. Good news!

REFERENCES

Asher, J. T., & Sciarrino, J. A. (1974). Realistic work sample tests: A review. Personnel Psychology, 27, 519-533.

Borman, W. C. (1991). Job behavior, performance, and effectiveness. In M. D. Dunnette & L. M. Hough (Eds.), Handbook of industrial and organizational psychology (2nd ed.) (Vol. 2) (pp. 271-326). Palo Alto, CA: Consulting Psychologists Press.

Borman, W. C., & Hallam, G. L. (1991). Observation accuracy for assessors of work-sample performance: Consistency across task and individual-difference correlates. Journal of Applied Psychology, 76, 11-18.

Borman, W. C., White, L. A., & Dorsey, D. W. (1995). Effects of task performance and interpersonal factors on supervisor and peer performance ratings. Journal of Applied Psychology, 80, 168-177.

Borman, W. C., White, L. A., Pulakos, E. D., & Oppler, S. H. (1991). Models of supervisory job performance ratings. Journal of Applied Psychology, 76, 863-872.

Brugnoli, G. A., Campion, J. E., & Basen, J. A. (1979). Racial bias in the use of work samples for personnel selection. Journal of Applied Psychology, 64, 119-123.

Burtch, L. D., Lipscomb, M. S., & Wissman, D. J. (1982). Aptitude requirements based on task difficulty: Methodology for evaluation. (AFHRL-TR-81-34) Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Campion, J. E. (1972). Work sampling for personnel selection. Journal of Applied Psychology, 56, 40-44.

Cooksey, R. W. (1996). Judgment analysis: Theory, methods, and applications. San Diego, CA: Academic Press.

Department of Defense (1984). Test manual for the armed services vocational aptitude battery. North Chicago, IL: United States Military Entrance Processing Command.

DuBois, C. L., Sackett, P. R., Zedeck, S., & Fogli, L. (1993). Further exploration of typical and maximum performance criteria: Definitional issues, prediction, and white-black differences. Journal of Applied Psychology, 78, 205-211.

Ford, J. K., Kraiger, K., & Schechtman, S. L. (1986). Study of race effects in objective indices and subjective evaluations of performance: A meta-analysis of performance criteria. Psychological Bulletin, 99, 330-337.

Hamner, C. W., Kim, J. S., Baid, L., & Bigoness, W. J. (1974). Race and sex as determinants

ratings by potential employers in a simulated work sample task. Journal of Applied Psychology, 59, 705-711.

Hedge, J. W., Dickinson, T. L., & Bierstedt, S. A. (1988). The use of videotape technology to train administrators of walk-through performance testing. (AFHRL-TP-87-71). Brooks AFB, TX: Air Force Human Resources Laboratory, Training Systems Division.

Hedge, J. W., & Teachout, M. S. (1986). Job performance measurement: A systematic program of research and development. (AFHRL-TP-86-37). Brooks AFB, TX: Air Force Human Resources Laboratory, Training Systems Division.

Hedge, J. W., & Teachout, M. S. (1992). An interview approach to work sample criterion measurement. Journal of Applied Psychology, 77, 453-461.

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. Psychological Bulletin, 57, 116-131.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. Psychological Bulletin, 96, 72-98.

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: Air Force Human Resources Laboratory, Training Systems Division.

Klimoski, R. , & Brickner, M. (1987). Why do assessment centers work? The puzzle of assessment center validity. Personnel Psychology, 40, 243-260.

Lance, C. E., Hedge, J. W., & Alley, W. E. (1989). Joint relations of task proficiency with ability, experience, and task difficulty: A cross-level, interactional study. Human Performance, 2, 249-272.

Lance, C. E., Parisi, A. G., Bennett, W. Jr., Teachout, M. S., Harville, D. L., & Welles, M. L. (1998). Moderators of skill retention interval/performance decrement relationships in eight U.S. Air Force enlisted specialties. Human Performance, 11, 103-123.

Lance, C. E., Teachout, M. S., & Donnelly, T. M. (1992). Specification of the criterion construct space: An application of hierarchical confirmatory factor analysis. Journal of Applied Psychology, 77, 437-452.

Laue, F. J., Hedge, J. W., Wall, M. L., Pederson, L. A., & Bentley, B. A. (1992). Job performance measurement system development process. (AL-TR-1992-0120). Brooks AFB, TX: Armstrong Laboratory, Human Resources Directorate, Technical Training Research Division.

Martell, R. F., Guzzo, R. A., & Willis, C. E. (1995). A methodological and substantive note on the performance-cue effect in ratings of work-group behavior. Journal of Applied Psychology, 80, 191-195.

Mumford, M. D., Weeks, J. L., Harding, F. D., & Fleishman, E. A. (1987). Measuring occupational difficulty: A construct validation against training criteria. Journal of Applied Psychology, 72, 578-587.

Pulakos, E. D., White, L. A., Oppler, S. H., & Borman, W. C. (1989). Examination of race and sex effects on performance ratings. Journal of Applied Psychology, 74, 770-780.

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than g. Journal of Applied Psychology, 79, 518-524.

Robertson, I. T., & Downs, S. (1989). Work-sample tests of trainability: A meta-analysis. Journal of Applied Psychology, 74, 402-410.

Robertson, I. T., & Kandola, R. S. (1982). Work sample tests: Validity, adverse impact, and applicant reaction. Journal of Occupational Psychology, 55, 171-183.

Sackett, P. R., & DuBois, C. L. Z. (1991). Rater-ratee race effects on performance evaluation: Challenging meta-analytic conclusions. Journal of Applied Psychology, 76, 873-877.

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. Journal of Applied Psychology, 73, 482-486.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. Personnel Psychology, 37, 407-422.

Smith, F. D. (1991). Work samples as measures of performance. In A. K. Wigdor & B. F. Green Jr. (Eds.), Performance assessment for the workplace (Vol. 2) (pp. 27-52). Washington, DC: National Academy Press.

Smith, M. (1994). A theory of the validity of predictors in selection. Journal of Occupational and Organizational Psychology, 67, 13-31.

Teachout, M. S., & Pellum, M. W. (1991). Air Force research to link standards for enlistment to on-the-job performance. (AFHRL-TR-90-90). Brooks AFB, TX: Air Force Human Resources Laboratory, Training Systems Division.

Terpstra, D. E. (1996). The search for effective methods. HR Focus, 73(5), 16-17.

Thorndike, R. L. (1949). Personnel selection. New York: Wiley.

Thornton, G. C. (1992). Assessment centers in human resource management. Reading, MA: Addison Wesley.

Tosi, H. L., & Einbender, S. W. (1985). The effects of the type and amount of information in sex discrimination research: A meta-analysis. Academy of Management Journal, 28, 712-723.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. Journal of Applied Psychology, 81, 557-574.

Weeks, J. (1984). Occupational learning difficulty: A standard for determining the order of aptitude requirement minimums. (AFHRL-SR-84-26). Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Wigdor, A. K., & Green, B. F. Jr. (Eds.), (1991a). Performance assessment for the workplace (Vol. 1). Washington, DC: National Academy Press.

Wigdor, A. K., & Green, B. F. Jr. (Eds.), (1991b). Performance assessment for the workplace (Vol. 2). Washington, DC: National Academy Press.

FOOTNOTES

[1]Data collection for the JPM project occurred in three sequential "waves." Data collection began with the Jet Engine Mechanic and Air Traffic Control Operator AFSs in the first wave, with data collection following for the remaining AFSs in subsequent waves. In the latter two waves, approximately 15% of the examinees' performance was evaluated using "shadow scoring," in which two test administrators independently observed and scored the examinee's step-level performance. Median interscorer reliabilities were $\underline{r} = .97$ and $\underline{r} = .93$ (Hedge & Teachout, 1992) for hands-on and interview work sample tasks (this distinction is described shortly), supporting the accuracy and objectivity of these step-level performance measures.

[2]Items relating to examinee motivation were administered only in data collection waves two and three. Consequently, these data were unavailable for the Jet Engine Mechanic and Air Traffic Control samples.

[3]Descriptive statistics for each AFS separately are available from Charles E. Lance.

[4]Data regarding statistically significant interaction effects are available from Charles E. Lance.